

A Survey on Self-Evolution of Large Language Models

Zhengwei Tao^{12*}, Ting-En Lin², Xiancai Chen¹, Hangyu Li², Yuchuan Wu²,
Yongbin Li^{2†}, Zhi Jin^{1†}, Fei Huang², Dacheng Tao³, Jingren Zhou²

¹ Key Lab of HCST (PKU), MOE; School of Computer Science, Peking University

² Alibaba Group ³Nanyang Technological University

{tttzw, xiancaich}@stu.pku.edu.cn, zhijin@pku.edu.cn

{ting-en.lte, shengxiu.wyc, shuide.lyb, jingren.zhou}@alibaba-inc.com
dacheng.tao@ntu.edu.sg

Abstract

Large language models (LLMs) have significantly advanced in various fields and intelligent agent applications. However, current LLMs that learn from human or external model supervision are costly and may face performance ceilings as task complexity and diversity increase. To address this issue, self-evolution approaches that enable LLM to autonomously acquire, refine, and learn from experiences generated by the model itself are rapidly growing. This new training paradigm inspired by the human experiential learning process offers the potential to scale LLMs towards superintelligence. In this work, we present a comprehensive survey of self-evolution approaches in LLMs. We first propose a conceptual framework for self-evolution and outline the evolving process as iterative cycles composed of four phases: experience acquisition, experience refinement, updating, and evaluation. Second, we categorize the evolution objectives of LLMs and LLM-based agents; then, we summarize the literature and provide taxonomy and insights for each module. Lastly, we pinpoint existing challenges and propose future directions to improve self-evolution frameworks, equipping researchers with critical insights to fast-track the development of self-evolving LLMs. Our corresponding GitHub repository is available at <https://github.com/AlibabaResearch/DAMO-ConvAI/tree/main/Awesome-Self-Evolution-of-LLM>.

1 Introduction

With the rapid development of artificial intelligence, large language models (LLMs) like GPT-3.5 (Ouyang et al., 2022), GPT-4 (Achiam et al., 2023), Gemini (Team et al., 2023), LLaMA (Touvron et al., 2023a,b), and Qwen (Bai et al., 2023)

mark a significant shift in language understanding and generation. These models undergo three stages of development as shown in Figure 1: pre-training on large and diverse corpora to gain a general understanding of language and world knowledge (Devlin et al., 2018; Brown et al., 2020), followed by supervised fine-tuning to elicit the abilities of downstream tasks (Raffel et al., 2020; Chung et al., 2022). Finally, the human preference alignment training enables the LLMs to respond as human behaviors (Ouyang et al., 2022). Such successive training paradigms achieve significant breakthroughs, enabling LLMs to perform a wide range of tasks with remarkable zero-shot and in-context capabilities, such as question answering (Tan et al., 2023), mathematical reasoning (Collins et al., 2023), code generation (Liu et al., 2024b), and task-solving that require interaction with environments (Liu et al., 2023b).

Despite these advancements, humans anticipate that the emerging generation of LLMs can be tasked with assignments of greater complexity, such as scientific discovery (Miret and Krishnan, 2024) and future events forecasting (Schoenegger et al., 2024). However, current LLMs encounter challenges in these sophisticated tasks due to the inherent difficulties in modeling, annotation, and the evaluation associated with existing training paradigms (Burns et al., 2023). Furthermore, the recently developed Llama-3 model has been trained on an extensive corpus comprising 15 trillion tokens¹. It's a monumental volume of data, suggesting that significantly scaling model performance by adding more real-world data could pose a limitation. This has attracted interest in self-evolving mechanisms for LLMs, akin to the natural evolution of human intelligence and illustrated by AI developments in gaming, such as the transition from

*Work done while interning at Alibaba Group.

†Corresponding authors.

¹<https://huggingface.co/meta-llama/Meta-Llama-3-70B-Instruct>

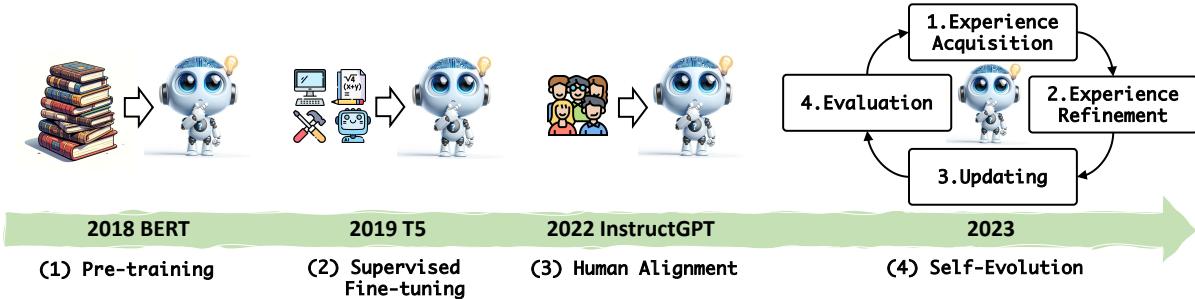


Figure 1: Training paradigms shift of LLMs.

AlphaGo (Silver et al., 2016) to AlphaZero (Silver et al., 2017). AlphaZero’s self-play method, requiring no labeled data, showcases a path forward for LLMs to surpass current limitations and achieve superhuman performance without intensive human supervision.

Drawing inspiration from the paradigm above, research on the self-evolution of LLMs has rapidly increased at different stages of model development, such as self-instruct (Wang et al., 2023b), self-play (Tu et al., 2024), self-improving (Huang et al., 2022), and self-training (Gulcehre et al., 2023). Notably, DeepMind’s AMIE system (Tu et al., 2024) outperforms primary care physicians in diagnostic accuracy, and Microsoft’s WizardLM-2² exceeds the performance of the initial version of GPT-4. Both models are developed using self-evolutionary frameworks with autonomous learning capabilities and represent a potential LLM training paradigm shift. However, the relationships between these methods remain unclear, lacking systematic organization and analysis.

Therefore, we first comprehensively investigate the self-evolution processes in LLMs and establish a conceptual framework for their development. This self-evolution is characterized by an iterative cycle involving experience acquisition, experience refinement, updating, and evaluation, as shown in Figure 2. During the cycle, an LLM initially gains experiences through evolving new tasks and generating corresponding solutions, subsequently refining these experiences to obtain better supervision signals. After updating the model in-weight or in-context, the LLM is evaluated to measure progress and set new objectives.

The concept of self-evolution in LLMs has sparked considerable excitement across various research communities, promising a new era of models that can adapt, learn, and improve au-

tonomously, akin to human evolution in response to changing environments and challenges. Self-evolving LLMs are not only able to transcend the limitations of current static, data-bound models but also mark a shift toward more dynamic, robust, and intelligent systems. This survey deepens understanding of the emerging field of self-evolving LLMs by providing a comprehensive overview through a structured conceptual framework. We trace the field’s evolution from the past to the latest cutting-edge methods and applications while examining existing challenges and outlining future research directions, paving the way for significant advances in developing self-evolution frameworks and next-generation models.

The survey is organized as follows: We first present the overview of self-evolution (§ 2), including background and conceptual framework. We summarize existing evolving abilities and domains of current methods (§ 3). Then, we provide in-depth analysis and discussion on the latest advancements in different phases of the self-evolution process, including experience acquisition (§ 4), experience refinement (§ 5), updating (§ 6), and evaluation (§ 7). Finally, we outline open problems and prospective future directions (§ 8).

2 Overview

In this section, we will first discuss the background of self-evolution and then introduce the proposed conceptual framework.

2.1 Background

Self-Evolution in Artificial Intelligence. Artificial Intelligence represents an advanced form of intelligent agent, equipped with cognitive faculties and behaviors mirroring those of humans. The aspiration of AI developers lies in enabling AI to harness self-evolutionary capabilities, paralleling the experiential learning processes characteristic of

²<https://wizardlm.github.io/WizardLM2/>

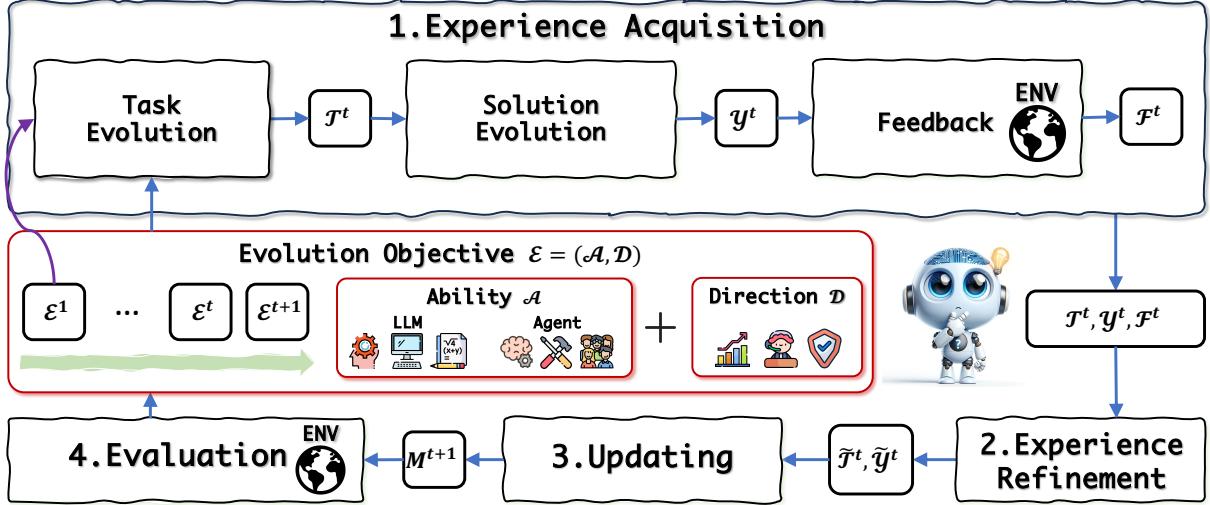


Figure 2: Conceptual framework of self-evolution. For the t^{th} iteration: \mathcal{E}^t is the evolution objective; \mathcal{T}^t and \mathcal{Y}^t denote the task and solution; \mathcal{F}^t represents feedback; M^t is the current model. Refined experiences are marked as $\tilde{\mathcal{T}}^t$ and $\tilde{\mathcal{Y}}^t$, leading to the evolved model \tilde{M} . ENV is the environment. The whole self-evolution starts at \mathcal{E}^1 .

human development. The concept of self-evolution in AI emerges from the broader fields of machine learning and evolutionary algorithms (Bäck and Schwefel, 1993). Initially influenced by the principles of natural evolution, such as selection, mutation, and reproduction, researchers have developed algorithms that simulate these processes to optimize solutions to complex problems. The landmark paper by Holland (1992), which introduced the genetic algorithm, marks a foundational moment in the history of AI’s capability for self-evolution. Subsequent developments in neural networks and deep learning have furthered this capability, allowing AI systems to modify their own architectures and improve performance without human intervention (Liu et al., 2021).

Can Artificial Entities Evolve Themselves? Philosophically, the question of whether artificial entities can self-evolve touches on issues of autonomy, consciousness, and agency. While some philosophers argue that true self-evolution in AI would require some form of consciousness or self-awareness, others maintain that mechanical self-improvement through algorithms does not constitute genuine evolution (Chalmers, 1997). This debate often references the works of thinkers like Dennett (1993), who explore the cognitive processes under human consciousness and contrast them with artificial systems. Ultimately, the philosophical inquiry into AI’s capacity for self-evolution remains deeply intertwined with interpretations of what it means to ‘evolve’ and whether such processes can

purely be algorithmic or must involve emergent consciousness (Searle, 1986).

2.2 Conceptual Framework

In the conceptual framework of self-evolution, we describe a dynamic, iterative process mirroring the human ability to acquire and refine skills and knowledge. This framework is encapsulated within Figure 2, emphasizing the cyclical nature of learning and improvement. Each iteration of the process focuses on a specific evolution goal, allowing the model to engage in relevant tasks, optimize its experiences, update its architecture, and evaluate its progress before moving to the next cycle.

Experience Acquisition At the t^{th} iteration, the model identifies an evolution objective \mathcal{E}^t . Guided by this objective, the model embarks on new tasks \mathcal{T}^t , generating solutions \mathcal{Y}^t and receiving feedback \mathcal{F}^t from the environment, ENV. This stage culminates in the acquisition of new experiences $(\mathcal{T}^t, \mathcal{Y}^t, \mathcal{F}^t)$.

Experience Refinement After experience acquisition, the model examines and refines these experiences. This involves discarding incorrect data and enhancing imperfect ones, resulting in refined outcomes $(\tilde{\mathcal{T}}^t, \tilde{\mathcal{Y}}^t)$.

Updating Leveraging the refined experiences, the model undergoes an update process, integrating $(\tilde{\mathcal{T}}^t, \tilde{\mathcal{Y}}^t)$ into its framework. This ensures the model remains current and optimized.

Evaluation The cycle concludes with an evaluation phase, where the model’s performance is assessed through an evaluation in external environment. The outcomes of this phase inform the objective \mathcal{E}^{t+1} , setting the stage for the subsequent iteration of self-evolution.

The conceptual framework outlines the self-evolution of LLMs, akin to human-like acquisition, refinement, and autonomous learning processes. We illustrate our taxonomy in Figure 3.

3 Evolution Objectives

Evolution objectives in self-evolving LLMs serve as predefined goals that autonomously guide their development and refinement. Much like humans set personal objectives based on needs and desires, these objectives are crucial as they determine how the model iteratively self-updates. They enable the LLM to autonomously learn from new data, optimize algorithms, and adapt to changing environments, effectively "feeling" its needs from feedback or self-assessment and setting its own goals to enhance functionality without human intervention.

We define an evolution objective as combining an evolving ability and an evolution direction. An evolving ability stands for an innate and detailed skill. The evolution direction is the aspect of the evolution objective aiming to improve. We formulate the evolution objective as follows:

$$\mathcal{E}^t = (\mathcal{A}^t, \mathcal{D}^t), \quad (1)$$

where \mathcal{E}^t is the evolution objective, composed by evolving abilities \mathcal{A}^t and evolution directions \mathcal{D}^t . Take "reasoning accuracy improving" as an example, "reasoning" is the evolving ability and "accuracy improving" is the evolution direction.

3.1 Evolving Abilities

In Table 1, we summarize and categorize the targeted evolving abilities in current self-evolution research into two groups: LLMs and LLM Agents.

3.1.1 LLMs

These are fundamental abilities underlying a broad spectrum of downstream tasks.

Instruction Following: The capability to follow instructions is essential for effectively applying language models. It allows these models to address specific user needs across different tasks and domains, aligning their responses within the given context (Xu et al., 2023a).

Reasoning: LLMs can self-evolve to recognize statistical patterns, making logical connections and deductions based on the information. They evolve to perform better reasoning involving methodically dissecting problems in a logical sequence (Cui and Wang, 2023).

Math: LLMs enhance the intricate ability to solve mathematical problems covering arithmetic, math word, geometry, and automated theorem proving (Ahn et al., 2024) towards self-evolution.

Coding: Methods improve the LLM coding abilities to generate more precise and robust programs (Singh et al., 2023; Zelikman et al., 2023). Furthermore, EvoCodeBench (Li et al., 2024a) provides an evolving benchmark that updates periodically to prevent data leakage.

Role-Play: It involves an agent understanding and acting out a particular role within a given context. This is crucial in scenarios where the model must fit into a social structure or follow a set of behaviors associated with a specific identity or function (Lu et al., 2024a).

Others: Apart from the above fundamental evolution objectives, self-evolution can also achieve and a wide range of NLP tasks (Stammer et al., 2023; Koa et al., 2024; Gulcehre et al., 2023; Zhang et al., 2024b,c).

3.1.2 LLM-based Agents

The abilities discussed here are characteristic of advanced artificial agents used for task-solving or simulations in digital or physical world. These capabilities mirror human cognitive functions, allowing these agents to perform complex tasks and interact effectively in dynamic environments.

Planning: It involves the ability to strategize and prepare for future actions or goals. An agent with this skill can analyze the current state, predict the outcomes of potential actions, and create a sequence of steps to achieve a specific objective (Qiao et al., 2024).

Tool Use: This is the capacity to employ objects or instruments in the environment to perform tasks, manipulate surroundings, or solve problems (Zhu et al., 2024).

Embodied Control: It refers to an agent’s ability to manage and coordinate its physical form within an environment. This encompasses locomotion, dexterity, and the manipulation of objects (Bousmalis et al., 2023).

Communication: It is the skill to convey information and understand messages from other agents

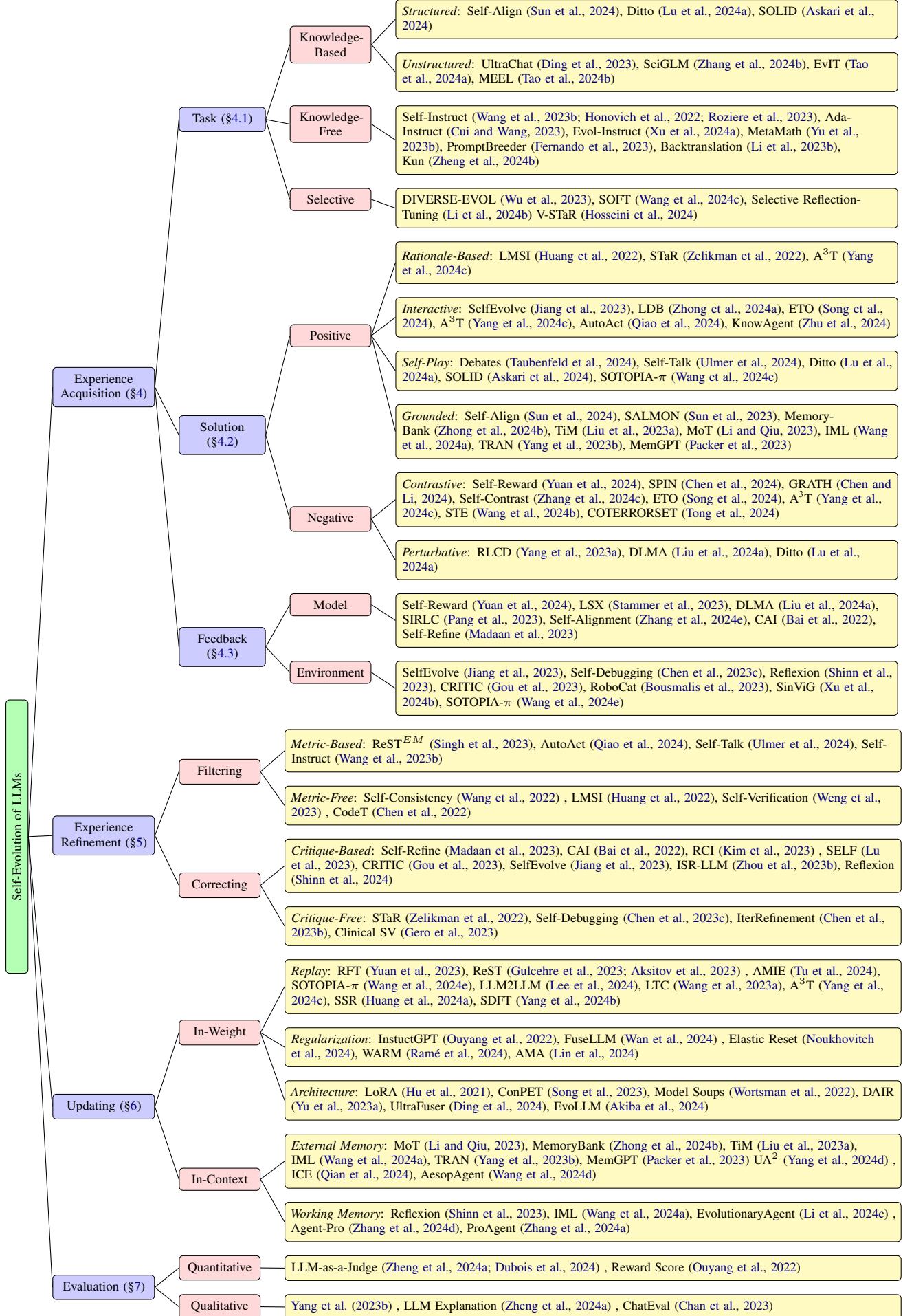


Figure 3: Taxonomy of self-evolving large language models.

or humans. Agents with advanced communication abilities can participate in dialogue, collaborate with others, and adjust their behaviour based on the communication received (Ulmer et al., 2024).

3.2 Evolution Directions

Examples of evolution directions include but are not limited to:

Improving Performance: The goal is to continuously enhance the model’s understanding and generation across various languages and abilities. For instance, a model initially trained for question answering and chitchat can autonomously extend its proficiency and develop abilities like diagnostic dialogue (Tu et al., 2024), social skills (Wang et al., 2024e), and role-playing (Lu et al., 2024a).

Adaptation to Feedback: This involves improving model responses based on feedback to better align with preferences or adapt to environments (Yang et al., 2023a; Sun et al., 2024).

Expansion of Knowledge Base: The aim is to continuously update the model’s knowledge base with the latest information and trends. For example, a model might automatically integrate new scientific research into its responses (Wu et al., 2024).

Safety, Ethic and Reducing Bias: The goal is to identify and mitigate biases in the model’s responses, ensuring fairness and safety. One effective strategy is to incorporate guidelines, such as constitutions or specific rules, to identify inappropriate or biased responses and corrected through model updates (Bai et al., 2022; Lu et al., 2024b).

4 Experience Acquisition

Exploration and exploitation (Gupta et al., 2006) are fundamental strategies for learning in humans and LLMs. Among that, exploration involves seeking new experiences to achieve objectives and is analogous to the initial phase of LLM self-evolution, known as experience acquisition. This process is crucial for self-evolution, enabling the model to autonomously tackle core challenges such as adapting to new tasks, overcoming knowledge limitations, and enhancing solution effectiveness. Furthermore, experience is a holistic construct, encompassing not only the tasks encountered (Dewey, 1938) but also the solutions developed to address these tasks (Schön, 2017), and the feedback (Boud et al., 2013) received as a result of task performance.

Inspired by that, we divide experience acquisi-

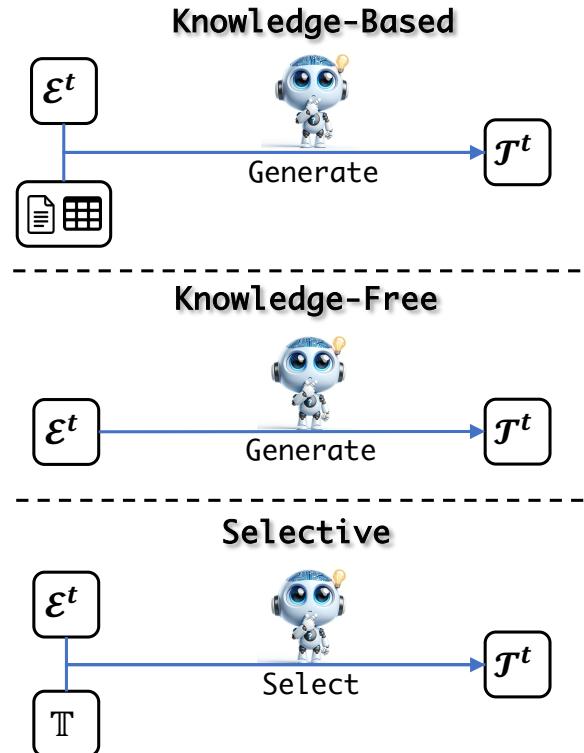


Figure 4: Task evolution. \mathcal{E}^t and \mathcal{T}^t are the evolving objective and task of t^{th} iteration. \mathbb{T} is the set of all tasks to be selected. The first two are generative methods that differ based on their respective use of knowledge. The third method, in contrast, employs a discriminative approach to select what to learn.

tion into three parts: task evolution, solution evolution, and obtaining feedback. In task evolution, LLMs curate and evolve new tasks aligning with evolution objectives. For solution evolution, LLMs develop and implement strategies to complete these tasks. Finally, LLMs may optionally collect feedback from interacting with the environment for further improvements.

4.1 Task Evolution

To gain new experience, the model first evolves new tasks according to the evolution objective \mathcal{E}^t in the current iteration. Task evolution is the crucial step in the engine that starts the entire evolution process. Formally, we denote the task evolution as:

$$\mathcal{T}^t = f^{\mathcal{T}}(\mathcal{E}^t, M^t), \quad (2)$$

where $f^{\mathcal{T}}$ is the task evolution function. \mathcal{E}^t , M^t , and \mathcal{T}^t denote the evolution objective, the model, and the evolved task at iteration t , respectively. We summarize and categorize existing studies on the task evolution method $f^{\mathcal{T}}$ into three groups: *Knowledge-Based*, *Knowledge-Free*, and *Selective*.

METHOD	ACQUISITION			REFINEMENT $f^{\mathcal{R}}$	UPDATING $f^{\mathcal{U}}$	OBJECTIVE \mathcal{E}
	TASK f^T	SOLUTION f^Y	FEEDBACK f^F			
LARGE LANGUAGE MODELS						
Self-Align (Sun et al., 2024)	Context-Based	Pos-G	-	Filtering	In-W	IF
SciGLM (Zhang et al., 2024b)	Context-Based	-	-	-	In-W	Other
EvIT (Tao et al., 2024a)	Context-Based	-	-	-	In-W	Reasoning
MEEL (Tao et al., 2024b)	Context-Based	-	-	-	In-W	Reasoning
UltraChat (Ding et al., 2023)	Context-Based	-	-	-	In-W	Role-Play
SOLID (Askanari et al., 2024)	Context-Based	Pos-S	-	Filtering	In-W	Role-Play
Ditto (Lu et al., 2024a)	Context-Based	Pos-S, Neg-P	-	-	In-W	Role-Play
MetaMath (Yu et al., 2023b)	Context-Free	Pos-R	-	-	In-W	Math
Self-Rewarding (Yuan et al., 2024)	Context-Free	-	Model	-	In-W	IF,Reasoning,Role-Play
Kun (Zheng et al., 2024b)	Context-Free	-	-	Filtering	In-W	IF,Reasoning
PromptBreeder (Fernando et al., 2023)	Context-Free	-	-	-	In-C	Math, Reasoning
Ada-Instruct (Cui and Wang, 2023)	Context-Free	-	-	-	In-W	Math, Reasoning, Code
Backtranslation (Li et al., 2023b)	Context-Free	-	-	-	In-W	IF
DiverseEvol (Wu et al., 2023)	Selective	Pos-I	-	-	In-W	Code
Grath (Chen and Li, 2024)	Selective	Neg-C	Model	-	In-W	Reasoning
REST ^{em} (Singh et al., 2023)	Selective	-	Model	Filtering	In-W	Math, Code
SOFT (Wang et al., 2024c)	Selective	-	-	-	In-W	IF
LSX (Stammer et al., 2023)	-	Pos-R	Model	Correcting	In-W	Other
LMSI (Huang et al., 2022)	-	Pos-R	-	Filtering	In-W	Math
TRAN (Yang et al., 2023b)	-	Pos-G	-	-	In-C	Reasoning
MOT (Li and Qiu, 2023)	-	Pos-R, Pos-G	-	Filtering	In-C	Math, Reasoning
STaR (Zelikman et al., 2022)	-	Pos-R, Neg-C	Model	Correct	In-W	Reasoning
COTERRORSET (Tong et al., 2024)	-	Pos-R, Neg-C	-	-	In-W	Math, Reasoning
Self-Debugging (Chen et al., 2023c)	-	Pos-I	Env	-	In-C	Code
SelfEvolve (Jiang et al., 2023)	-	Pos-I	-	-	In-C	Code
Reflexion (Shinn et al., 2024)	-	Pos-I, Pos-G	-	-	In-C	Code, Reasoning
V-STaR (Hosseini et al., 2024)	-	Neg-C	Model	Filter	In-W	Math, Code
Self-Contrast (Zhang et al., 2024e)	-	Neg-C	Model	-	In-W	Reasoning
SALMON (Sun et al., 2023)	-	Neg-C	Model	-	In-W	IF,Reasoning,Role-Play
SPIN (Chen et al., 2024)	-	Neg-C	-	-	In-W	IF,Reasoning,Role-Play
RLCD (Yang et al., 2023a)	-	Neg-P	Model	-	In-W	IF
DLMA (Liu et al., 2024a)	-	Neg-P	Model	-	In-W	IF
SELF (Lu et al., 2023)	-	-	Model	Correct	In-W	IF, Math
LLM AGENTS						
AutoAct (Qiao et al., 2024)	Context-Based	Pos-I	Env	Filtering	In-W	Planning, Tool
KnowAgent (Zhu et al., 2024)	Context-Based	Pos-I, Pos-G	Env	Filtering	In-W	Embodied, Planning, Tool
RoboCat (Bousmalis et al., 2023)	Context-Free	Pos-I	Env	-	In-W	Embodied
STE (Wang et al., 2024b)	Context-Free	Pos-I, Neg-C	Env	Correct	In-W	Tool
IML (Wang et al., 2024a)	-	Pos-R, Pos-G	-	-	In-C	Reasoning
SinViG Xu et al. (2024b)	-	Pos-I	Env	Filtering	In-W	Embodied
ETO (Song et al., 2024)	-	Pos-I, Neg-C	Env	Correct	In-W	Tool
A ³ T (Yang et al., 2024c)	-	Pos-I, Neg-C	Env	Correct	In-W	Tool
Debates (Taubenfeld et al., 2024)	-	Pos-S	-	-	In-W	Communication
SOTOPIA-π (Wang et al., 2024e)	-	Pos-S, Pos-G	Env	-	In-W	Communication
Self-Talk (Ulmer et al., 2024)	-	Pos-S, Pos-G	Model	Filtering	In-W	Communication
MemGPT (Packer et al., 2023)	-	Pos-G	Env	Filtering	In-C	Communication
MemoryBank (Zhong et al., 2024b)	-	Pos-G	Env	Filtering	In-C	Communication
ProAgent (Zhang et al., 2024a)	-	Pos-G	Env	-	In-C	Embodied
Agent-Pro (Zhang et al., 2024d)	-	Pos-G	Env	-	In-C	Planning
AesopAgent (Wang et al., 2024d)	-	Pos-G	Env	-	In-C	Planning
ICE (Qian et al., 2024)	-	Pos-G	Env	-	In-C	Planning
TiM (Liu et al., 2023a)	-	Pos-G	-	-	In-C	Communication
Werewolf (Xu et al., 2023b)	-	Pos-G	-	-	In-C	Planning

Table 1: Overview of self-evolution methods, detailing approaches across evolutionary stages. Key: Pos (Positive), Neg (Negative), R (Rationale-based), I (Interactive), S (Self-play), G (Grounded), C (Contrastive), P (Perturbative), Env (Environment), In-W (In-Weight), In-C (In-Context), IF (Instruction-Following). For the evolution objectives, **Adaptation to Feedback** is in green, **Expansion of Knowledge Base** is in blue, and **Safety, Ethic and Reducing Bias** is in brown. Improving Performance is in the default color, black.

We detail each type in the following parts and show the concepts in Figure 4.

Knowledge-Based The objective \mathcal{E}^t may associate with external knowledge to evolve where the knowledge is not inherently comprised in the current LLMs. Explicitly sourcing from knowledge enriches the relevance between tasks and evolution objectives. It also ensures the validity of relevant facts in the tasks. We delve into the Knowledge-

Based methods seeking to evolve new tasks of the evolving objective assisted by external information.

The first kind of knowledge is structured. Structured knowledge is dense in information and well-organized. Self-Align (Sun et al., 2024) curates topic-guided tasks generated covering 20 scientific topics such as scientific and legal expertise. Apart from topic knowledge, DITTO (Lu et al., 2024a) includes character knowledge from Wikidata and

Wikipedia. The knowledge comprises attributes, profiles, and concise character details for role-play conversations. SOLID (Askari et al., 2024) generates structured entity knowledge as conversation starters.

The second group consists of tasks evolving from an unstructured context. Unstructured context is easy to obtain but is sparse in knowledge. UltraChat (Ding et al., 2023) gathers unstructured knowledge of 20 types of text materials based on 30 meta-concepts to construct conversation tasks. SciGLM (Zhang et al., 2024b) derives questions from the text of diversified science subjects, which covers rich scientific knowledge. EvIT (Tao et al., 2024a) derives event reasoning tasks based on large-scale unstructured events mined from the unsupervised corpus. Similarly, MEEL (Tao et al., 2024b) evolves multi-modal events in both image and text to construct the tasks for MM event reasoning.

Knowledge-Free Unlike previous methods that require extensive human effort to gather external knowledge, Knowledge-Free approaches operate independently using the evolving object \mathcal{E}^t and the model itself. These efficient methods can generate more diversified tasks without additional knowledge restrictions.

First, the LLMs can prompt themselves to generate new tasks according to \mathcal{E}^t . Self-Instruct (Wang et al., 2023b; Honovich et al., 2022; Roziere et al., 2023) is a typical methodology of Knowledge-Free task evolution. These methods self-generate a variety of new task instructions based on evolution objectives. Ada-Instruct (Cui and Wang, 2023) further proposes an adaptive task instruction generation strategy that fine-tunes open-source LLMs to generate lengthy and complex task instructions for code completion and mathematical reasoning.

Second, extending and boosting original tasks increases the quality of instructions. WizardLM (Xu et al., 2023a) proposes Evol-Instruct that evolves instruction following tasks with in-depth and in-breadth evolving and further expands it in code generation (Luo et al., 2024). MetaMath (Yu et al., 2023b) rewrites the question in multiple ways, including rephrasing, self-verification, and FOBAR. It evolves a new MetaMathQA dataset for fine-tuning LLMs to improve mathematical task-solving. Promptbreeder (Fernando et al., 2023) evolves seed tasks via mutation prompts. It further evolves mutation prompts via the hyper mutation

prompts to increase the task diversity.

Third, deriving tasks from plain text is another way. Backtranslation (Li et al., 2023b) extracts self-contained segments in unlabelled data and regards it as the answers to tasks. Similarly, Kun (Zheng et al., 2024b) presents a task self-evolving algorithm utilizing instruction harnessed from unlabelled data towards back-translation.

Selective Instead of task generation, we may start with large-scale existing tasks. At each iteration, LLMs can select tasks that exhibit the highest relevance to the current evolving objective \mathcal{E}^t without additional generation. This approach obviates the intricate curation of new tasks, streamlining the evolution process (Zhou et al., 2024; Li et al., 2023a; Chen et al., 2023a).

A simple task selecting method is to randomly sample tasks from the task pool like REST (Gulcehre et al., 2023), REST^{em} (Singh et al., 2023), and GRATH (Chen and Li, 2024) do. Rather than random selection, DIVERSE-EVOL (Wu et al., 2023) introduces a data sampling technique where the model selects new data points based on their distinctiveness in the embedding space, ensuring diversity enhancement in the chosen subset. SOFT (Wang et al., 2024c) then splits the initial training set. Each iteration selects one chunk of the split set as the evolving task.

Li et al. (2024b) propose Selective Reflection-Tuning and select a subset of tasks via a novel metric calculating to what extent the answer is related to the question. V-STaR (Hosseini et al., 2024) selects the correct solutions in the previous iteration and adds their task instructions to the task set of the next iteration.

4.2 Solution Evolution

After obtaining evolved tasks, LLMs solve the tasks to acquire the corresponding solution. The most common strategy is to generate the solution directly according to the task formulation (Zelikman et al., 2022; Gulcehre et al., 2023; Singh et al., 2023; Zheng et al., 2024b; Yuan et al., 2024). However, this straightforward approach might reach solutions irrelevant to the evolution objective, leading to sub-optimal evolution (Hare, 2019). Therefore, solution evolution uses different strategies to solve tasks and enhance LLM capabilities by ensuring that solutions are not just generated but are also relevant and informative. In this section, we comprehensively survey these strategies and illustrate them in

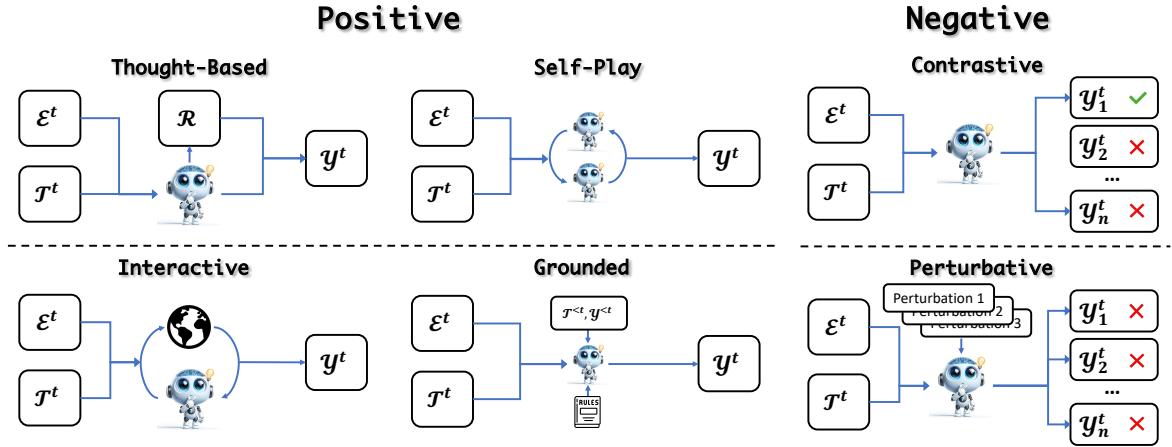


Figure 5: Solution evolution. \mathcal{E}^t , \mathcal{T}^t , and \mathcal{Y}^t are the evolving objective, task, and solution of t^{th} iteration. \mathcal{R} is the rational thought.

Figure 5. We first formulate the solution-evolution as follows:

$$\mathcal{Y}^t = f^{\mathcal{Y}}(\mathcal{T}^t, \mathcal{E}^t, M^t), \quad (3)$$

where $f^{\mathcal{Y}}$ is the model’s strategy to approach the evolution objective.

We then categorize these methods into positive and negative according to the correctness of the solutions. The positive methods introduce various approaches to acquire correct and desirable solutions. On the contrary, negative methods elicit and collect undesired solutions, including unfaithful or mis-align model behaviors, which are then used for preference alignment. We elaborate on the details of each type in the following sections.

4.2.1 Positive

Current studies explore diverse methods beyond vanilla inference for positive solutions to obtain correct solutions aligned with evolution objectives. We categorize the task-solving process into four types: Rationale-Based, Interactive, Self-Play, and Grounded.

Rationale-Based The model incorporates rationale explanations towards approaching the evolving objective when solving the tasks and can self-evolve by utilizing such rationales. These methods enable models to explicitly acknowledge the evolution objective and complete this task in that direction (Wei et al., 2022; Yao et al., 2024; Besta et al., 2024; Yao et al., 2022).

Huang et al. (2022) proposes a method where an LLM self-evolves using "high-confidence" rationale-augmented answers generated for unla-

beled questions. Similarly, STaR (Zelikman et al., 2022) generates rationale when solving the task. If the answer is wrong, it further corrects the rationale and the answer. Then, it uses the answer and rationale as experiences to fine-tune the model. Similarly, LSX (Stammer et al., 2023) proposes the novel paradigm to generate an explanation of the answer, incorporating an iterative loop between a learner module performing a base task and a critic module that assesses the quality of explanations given by the learner. Song et al. (2024); Yang et al. (2024c) obtain rationales in the ReAct (Yao et al., 2022) style when solving the tasks. The rationales are further engaged in training the agents in the following step.

Interactive Models can interact with the environment to enhance the evolution process. These methods can obtain environmental feedback that is valuable for guiding self-evolution directions.

SelfEvolve and LDB (Jiang et al., 2023; Zhong et al., 2024a) improve code generation ability via self-evolution. They allow the model to generate code and acquire feedback via running the code on the interpreter. As another environment, Song et al. (2024); Yang et al. (2024c) interact in embodied scenarios and acquire feedback. They learn to take proper actions based on their current state. For agent abilities, AutoAct (Qiao et al., 2024) introduces self-planning from scratch, focusing on an intrinsic self-learning process. In this process, agents enhance their abilities through recursive planning iterations with environment feedback. Following AutoAct, (Zhu et al., 2024) further enhances agent training by integrating self-evolution

and an external action knowledge base. This approach guides action generation and boosts planning ability through environment-driven corrective feedback loops.

Self-Play It’s the situation where a model learns to evolve by playing against copies of itself. Self-play is a powerful evolving method because it enables systems to communicate with themselves to get feedback in a closed loop. It’s especially effective in environments where the model can simulate various sides of the roles, like multi-player games (Silver et al., 2016, 2017). Compared with interactive methods, self-play is an effective strategy to obtain feedback without an environment.

Taubenfeld et al. (2024) investigate the systematic biases in simulations of debates by LLMs. On the contrary to debating, Ulmer et al. (2024) engage LLMs in conversations following generated principles. Another kind of conversation via role-playing. Lu et al. (2024a) proposes self-simulated role-play conversation. The process involves instructing the LLM with character profiles and aligning its responses to maintain consistency with the character’s knowledge and style. Similarly, Askari et al. (2024) propose SOLID to generate large-scale intent-aware role-play dialogues. This self-playing aspect harnesses the expansive knowledge of LLMs to construct information-rich exchanges that streamline the dialog generation process. Wang et al. (2024e) introduces a novel approach whereby each LLM follows a role and communicates with each other to achieve their goals.

Grounded To reach the evolving objective and reduce exploration space, models can be grounded on existing rules (Sun et al., 2024) and previous experiences for further explicit guidance when solving the tasks.

LLMs can generate desirable solutions more effectively by being grounded on pre-defined rules and principles. For instance, Self-Align (Sun et al., 2024) generated self-evolved questions with principle-driven constraints to guide the task-solving process. SALMON (Sun et al., 2023) design a set of combined principles that requires the model to follow when solving the task. Self-Talk (Ulmer et al., 2024) ensures the LLMs generate a workflow-aligned conversation based on preset agent characters. They generate the workflow in advance based on GPT-4.

Besides pre-defined rules, grounding on previous experiences can improve the solutions. Memory-

Bank (Zhong et al., 2024b) and TiM (Liu et al., 2023a) answer current questions by incorporating previous question-answer records. Rather than previous solution histories, MoT (Li and Qiu, 2023), IML (Wang et al., 2024a), and TRAN (Yang et al., 2023b) incorporate induced rules from the histories to answer new questions. MemGPT (Packer et al., 2023) combines these merits and retrieves previous questions, solutions, induced events, and user portrait knowledge.

4.2.2 Negative

In addition to acquiring positive solutions, recent research illustrates that LLMs can benefit from negative ones for self-improvement (Yang et al., 2023b). This strategy is analogous to trial and error in human behavior when learning skills. This section summarises typical methods of gaining negative solutions to assist in self-evolution.

Contrastive A widely used group of methods is to collect multiple solutions for a task and then contrast the positive and negative ones to get improvements.

Self-Reward, SPIN (Yuan et al., 2024; Chen et al., 2024) updates the model by comparing the answers of high and low scores. Similarly, GRATH (Chen and Li, 2024) generates both correct and incorrect answers. It then trains the model by comparing these two answers. Self-Contrast (Zhang et al., 2024c) contrasts the differences and summarizes these discrepancies into a checklist that could be used to re-examine and eliminate discrepancies. In ETO (Song et al., 2024), the model interacts with the embodied environment to complete tasks and optimizes from the failure solutions. A³T (Yang et al., 2024c) improves ETO by adding rationale after each action for solving tasks. STE (Wang et al., 2024b) implements trial and error where the model solves the tasks with unfamiliar tools. It learns by analyzing failed attempts to improve problem-solving strategies in future tasks. More recently, COTERRORSET (Tong et al., 2024) obtains incorrect solutions generated by PALM-2 and proposes mistake tuning, which requires the model to void making mistakes.

Perturbative Compared to *Contrastive*, *Perturbative* methods seek to add perturbations to obtain negative solutions intentionally. Models can later learn to avoid generating these negative answers. Adding perturbations to obtain negative solutions is more controllable than contrastive methods.

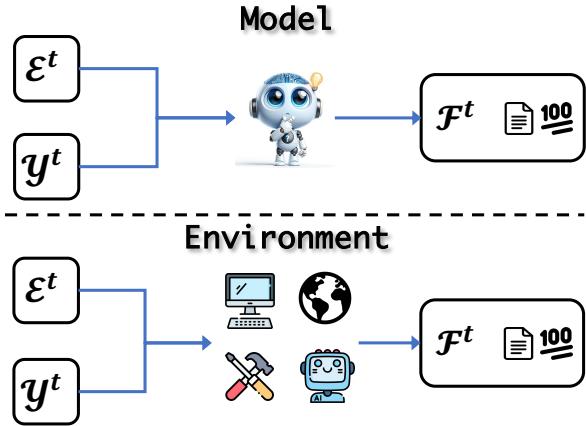


Figure 6: Types of Feedback. \mathcal{E}^t and \mathcal{Y}^t are the evolving objective and task solutions of t^{th} iteration.

Some methods add perturbation to generate harmful solutions (Yang et al., 2023a; Liu et al., 2024a). Given a task, RLCD (Yang et al., 2023a) curates both positive and negative instructions and generates positive and negative solutions. DLMA (Liu et al., 2024a) gathers both positive and negative instructional prompts and subsequently produces corresponding positive and negative solutions.

Rather than harmful perturbation, incorporating negative context is another way. Ditto (Lu et al., 2024a) adds negative persona characters to generate incorrect conversations. The model then learns from the negative conversations to evolve persona dialogue ability.

4.3 Feedback

As humans learn skills, feedback plays a critical role in demonstrating the correctness of the solutions. This key information enables humans to reflect and then update their skills. Akin to this process, LLMs should obtain feedback during or after the task solution in the cycle of self-evolution. We formalize the process as follows:

$$\mathcal{F}^t = f^{\mathcal{F}}(\mathcal{T}^t, \mathcal{Y}^t, \mathcal{E}^t, M^t, ENV), \quad (4)$$

where $f^{\mathcal{F}}$ is the method to acquire feedback.

In this part, we summarize two types of feedback. Model feedback refers to gathering the critique or score rated by the LLMs themselves. Besides, Environment denotes the feedback received directly from the external environment. We illustrate these concepts in Figure 6.

4.3.1 Model

Current studies demonstrate that LLMs can play well as a critic (Zheng et al., 2024a). In the cycle of self-evolution, the model judges itself to acquire the feedback of the solutions.

One type of feedback is a score that indicates correctness. Self-Reward (Yuan et al., 2024), LSX Stammer et al. (2023), and DLMA (Liu et al., 2024a) rate their own solutions and output the scores via LLM-as-a-Judge prompting. Similar to that, SIRLC (Pang et al., 2023) utilizes self-evaluation results of LLM as the reward for further reinforcement learning. Self-Alignment (Zhang et al., 2024e) leverages the self-evaluation capability of an LLM to generate confidence scores on the factual accuracy of its outputs.

Another type provides a textual description, offering multi-dimensional information. To alter the distribution of the responses via supervised learning, CAI (Bai et al., 2022) asks the model to critique its response according to a principle in the constitution. In contrast to supervised learning and reinforcement learning approaches, Self-Refine (Madaan et al., 2023) allows the model to generate natural language feedback on its own output in a few-shot manner.

4.3.2 Environment

Another form of feedback comes from the environment, common in tasks where solutions can be directly evaluated. This feedback is precise and elaborate and can provide sufficient information for model updating. They may be derived from code interpreter (Jiang et al., 2023; Chen et al., 2023c; Shinn et al., 2024), tool execution (Qiao et al., 2024; Gou et al., 2023), the embodied environment (Bousmalis et al., 2023; Xu et al., 2024b; Zhou et al., 2023b), and other LLMs or agents (Wang et al., 2024e; Taubenfeld et al., 2024; Ulmer et al., 2024).

For code generation, Self-Debugging Chen et al. (2023c) utilizes execution results on test cases as part of feedback while SelfEvolve (Jiang et al., 2023) receives the error message from the interpreter. Similarly, Reflexion (Shinn et al., 2023) also obtains the run-time feedback from the code interpreter. It then further reflects to generate thoughts. This run-time feedback contains the trace-back information that can point out the key information for improved code generation.

Recently, methods endow tool-using ability to LLMs and agents. Executing tools leading to feed-

back in return (Gou et al., 2023; Qiao et al., 2024; Song et al., 2024; Yang et al., 2024c; Wang et al., 2024b).

RoboCat (Bousmalis et al., 2023) and SinViG (Xu et al., 2024b) act in the robotic embodied environment. This type of feedback is precise and strong to guide self-evolution.

Communication feedback is common and effective in LLM-based multi-agent systems. Agents can correct and support each other, enabling co-evaluation (Wang et al., 2024e; Taubenfeld et al., 2024; Ulmer et al., 2024).

5 Experience Refinement

After experience acquisition and before updating in self-evolution, LLMs may improve the quality and reliability of their outputs through experience refinement. It helps LLMs adapt to new information and contexts without relying on external resources, leading to more reliable and effective assistance in dynamic environments. This process is formulated as follows:

$$\tilde{\mathcal{T}}^t, \tilde{\mathcal{Y}}^t = f^R(\mathcal{T}^t, \mathcal{Y}^t, \mathcal{F}^t, \mathcal{E}^t, M^t), \quad (5)$$

where f^R is the methods of experience refinement, $\tilde{\mathcal{T}}^t, \tilde{\mathcal{Y}}^t$ are the refined tasks and solutions. We classify the methods into two categories: filtering and correcting.

5.1 Filtering

Refinement in self-evolution involves two primary filtering strategies: *Metric-Based* and *Metric-Free*. The former uses external metrics to assess and filter outputs, while the latter does not rely on these metrics. This ensures that only the most reliable and high-quality data is utilized for further updating.

5.1.1 Metric-Based

By relying on feedback and pre-defined criteria, metric-based filtering improves the quality of the outputs (Singh et al., 2023; Qiao et al., 2024; Ulmer et al., 2024; Wang et al., 2023b), ensuring the progressive enhancement of LLM capabilities through each iteration of refinement.

For example, ReST^{EM} (Singh et al., 2023) incorporates a reward function to filter the dataset sampled from the current policy. The function provides binary rewards based on the correctness of the generated samples rather than a learned reward model trained on human preferences in ReST (Gulcehre et al., 2023). AutoAct (Qiao et al., 2024) leverages

F1-score and accuracy as rewards for synthetic trajectories and collects trajectories with exactly correct answers for further training. Self-Talk (Ulmer et al., 2024) measures the number of completed subgoals to filter the generated dialogues, ensuring that only high-quality data is used for training. To encourage diversity of the source instructions, Self-Instruct (Wang et al., 2023b) automatically filters low-quality or repeated instructions using ROUGE-L similarity and heuristics before adding them to the task pool.

The filtering criteria or metrics are crucial for maintaining the quality and reliability of the generated outputs, thereby ensuring the continuous improvement of the model's capability.

5.1.2 Metric-Free

Some methods seek filtering strategies beyond external metrics, making the process more flexible and adaptable. *Metric-Free* filtering typically involves sampling outputs and evaluating them based on internal consistency measures or other model-inherent criteria (Huang et al., 2022; Weng et al., 2023; Chen et al., 2022). The filtering in Self-Consistency (Wang et al., 2022) is based on the consistency of the final answer across multiple generated reasoning paths, with higher agreement indicating higher reliability. LMSI (Huang et al., 2022) utilizes CoT prompting plus self-consistency for generating high-confidence self-training data.

Designing internal consistency measures that accurately reflect output quality can be challenging. Self-Verification (Weng et al., 2023) allows the model to select the candidate answer with the highest interpretable verification score, calculated by assessing the consistency between the predicted and original condition values. For the code generation task, CodeT (Chen et al., 2022) considers both the consistency of the outputs against the generated test cases and the agreement of the outputs with other code samples.

These methods emphasize the language model's ability to self-assess and filter its outputs based on internal agreement, showcasing a significant step forward in self-evolution without the direct intervention of external metrics.

5.2 Correcting

Recent advancements in self-evolution have highlighted the significance of iterative self-correction, which enables models to refine their experiences. This section divides the methods employed into

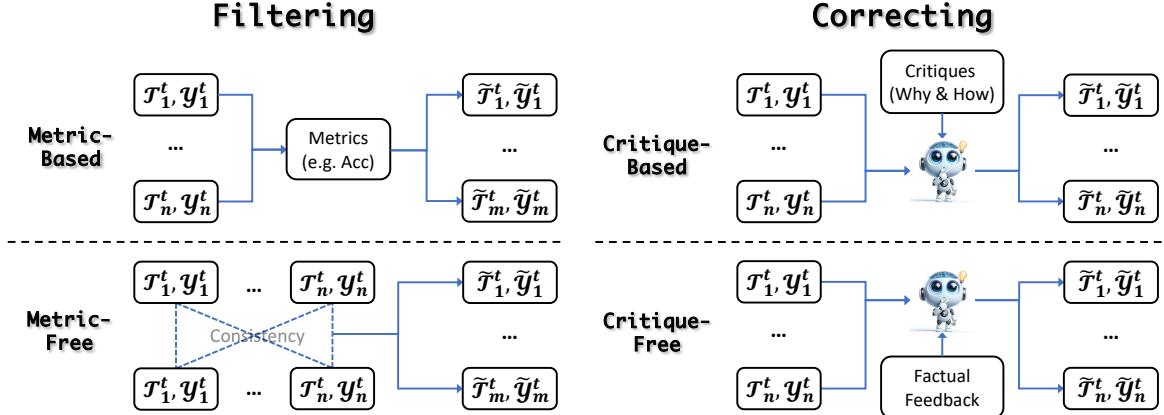


Figure 7: Experience refinement.

two categories: *Critique-Based* and *Critique-Free* correction. Critiques often serve as strong hints that include the rationale behind perceived errors or suboptimal outputs, guiding the model towards improved iterations.

5.2.1 Critique-Based

These methods rely on additional judging processes to draw the critiques of the experiences. Then, the experiences are refined based on the critiques. By leveraging either self-generated (Madaan et al., 2023; Bai et al., 2022; Shinn et al., 2023; Lu et al., 2023) or environment-interaction generated critiques (Gou et al., 2023; Jiang et al., 2023; Zhou et al., 2023b), the model benefits from detailed feedback for nuanced correction.

LLMs have demonstrated their ability to identify errors in their outputs. Self-Refine (Madaan et al., 2023) introduces an iterative process in which the model refines its initial outputs conditioned on actionable self-feedback without additional training. To evolve from the correction, CAI (Bai et al., 2022) generates critiques and revisions of its outputs in the supervised learning phase, significantly improving the initial model. Applied to an agent automating computer tasks, RCI (Kim et al., 2023) improves its previous outputs based on the critique finding errors in the outputs.

Since weaker models may struggle significantly with the self-critique process, several approaches enable models to correct the outputs using critiques provided by external tools. CRITIC (Gou et al., 2023) allows LLMs to revise the output based on the critiques obtained during interaction with tools in general domains. SelfEvolve (Jiang et al., 2023) prompts an LLM to refine the answer code based on the error information thrown by the interpreter.

ISR-LLM (Zhou et al., 2023b) helps the LLM planner find a revised action plan by using a validator in an iterative self-refinement process.

The primary advantage of this method lies in its ability to process and react to detailed feedback, potentially leading to more targeted and nuanced corrections.

5.2.2 Critique-Free

Contrary to critique-based, critique-free methods correct the experiences directly leveraging objective information (Zelikman et al., 2022; Chen et al., 2023c,b; Gero et al., 2023). These methods offer the advantage of independence from nuanced feedback that critiques provide, allowing for corrections that adhere strictly to factual accuracy or specific guidelines without the potential bias introduced by critiques.

One group of critique-free methods modifies the experiences on the signal of whether the task was correctly resolved. Self-Taught Reasoner (STaR) (Zelikman et al., 2022) proposes a technique that iteratively generates rationales to answer questions. If the answers are incorrect, the model is prompted again with the correct answer to generate a more informed rationale. Self-Debug (Chen et al., 2023c) enables the model to perform debugging steps by investigating execution results from unit tests and explaining the code on its own.

Different from depending on the task-solving signal, other information produced during the solving process can be leveraged. IterRefinement (Chen et al., 2023b) relies on a series of refined prompts that encourage the model to reconsider and improve upon its previous outputs without any direct critique. For information extraction tasks, Clinical SV (Gero et al., 2023) grounds each element

in evidence from the input and prunes inaccurate elements using supplied evidence.

These critique-free approaches simplify the correction mechanism, allowing for easier implementation and faster adjustments.

6 Updating

After experience refinement, we enter the crucial updating phase that leverages the refined experiences to improve model performance. We formulate updating as follows:

$$\mathbf{M}^{t+1} = f^{\mathcal{U}}(\tilde{\mathcal{T}}^t, \tilde{\mathcal{Y}}^t, \mathcal{E}^t, \mathbf{M}^t), \quad (6)$$

where $f^{\mathcal{U}}$ is the updating functions. These update methods keep the model effective by adapting to new experiences and continuously improving performance in changing environments and during iterative training.

We divide these approaches into in-weight learning, which involves updates to model weights, and in-context learning, which involves updates to external or working memory.

6.1 In-Weight

Classical training paradigms in updating LLMs in weight encompass continuous pretraining (Brown et al., 2020; Roziere et al., 2023), supervised fine-tuning (Longpre et al., 2023), and preference alignment (Ouyang et al., 2022; Touvron et al., 2023a). However, in the iterative training process of self-evolving, the core challenge lies in **achieving overall improvement** and **preventing catastrophic forgetting**, which entails refining or acquiring new capabilities while preserving original skills. Solutions to this challenge can be categorized into three main strategies: replay-based, regularization-based, and merging-based methods.

6.1.1 Replay-based

Replay-based methods reintroduce previous data to retain old knowledge. One is experience replay, which mixes the original and new training data to update LLMs (Roziere et al., 2023; Yang et al., 2024c; Zheng et al., 2023; Lee et al., 2024; Wang et al., 2023a). For example, Rejection sampling Fine-Tuning (RFT) (Yuan et al., 2023) and Reinforced Self-Training (ReST) (Gulcehre et al., 2023; Aksitov et al., 2023) method iteratively updates large language models by mixing seed training data with filtered new outputs generated by the model itself. AMIE (Tu et al., 2024) utilizes a self-play simulated learning environment for iterative

improvement and mixes generated dialogues with supervised fine-tuning data through inner and outer self-play loops. SOTONIA- π (Wang et al., 2024e) leverages behavior cloning from the expert model and self-generated social interaction trajectory to reinforce positive behaviors.

Another is generative replay, which adopts the self-generated synthesized data as knowledge to mitigate catastrophic forgetting. For instance, Self-Synthesized Rehearsal (SSR) (Huang et al., 2024a) generates synthetic training instances for rehearsal, enabling the model to preserve its ability without relying on real data from previous training stages. Self-Distillation Fine-Tuning (SDFT) (Yang et al., 2024b) generates a distilled dataset from the model itself to bridge the distribution gap between task datasets and the LLM’s original distribution to mitigate catastrophic forgetting.

6.1.2 Regularization-based

Regularization-based methods constrain the model’s updates to prevent significant deviations from original behaviors, exemplified by function-and weight-based regularization. Function-based regularization focuses on modifying the loss function that a model optimizes during training (Zhong et al., 2023; Peng et al., 2023). For example, InstuctGPT (Ouyang et al., 2022) employs a per-token KL-divergence penalty from the output probabilities of the initial policy model π_{SFT} on the updated policy model π_{RL} . FuseLLM (Wan et al., 2024) employs a technique akin to knowledge distillation (Hinton et al., 2015), leveraging the generated probability distributions from source LLMs to transfer the collective knowledge to the target LLM.

Weight-based regularization (Kirkpatrick et al., 2017) directly targets the model’s weights during training. Techniques such as Elastic Reset (Noukhovitch et al., 2024) counters alignment drift in RLHF by periodically resetting the online model to an exponentially moving average of its previous states. Furthermore, Ramé et al. (2024) introduced WARM, which combines multiple reward models through weight averaging to address reward hacking and misalignment. Moreover, AMA (Lin et al., 2024) adaptively average model weights to optimize the trade-off between reward maximization and forgetting mitigation.

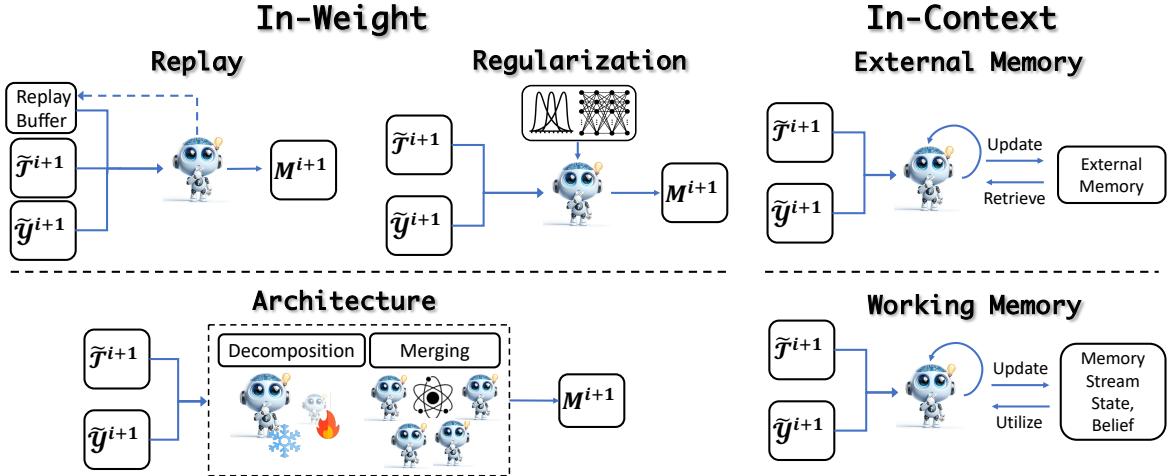


Figure 8: The illustration of *Updating* methods, including in-weight and in-context updating. The terms \tilde{T}^t and \tilde{y}^t represent refined experiences, each containing task and corresponding solutions, respectively. M^t denotes the updated model.

6.1.3 Architecture-based

Architecture-based methods explicitly utilize extra parameters or models for updating, including decomposition- and merging-based approaches. Decomposition-based methods separate large neural network parameters into general and task-specific components and only update the task-specific parameters to mitigate forgetting. LoRA (Hu et al., 2021; Dettmers et al., 2024) inject trainable low-rank matrices to significantly reduce the number of trainable parameters while maintaining or improving model performance across various tasks. This paradigm is later adopted by GPT4tools (Yang et al., 2024a), OpenAGI (Ge et al., 2024) and Dromedary (Sun et al., 2024). Dynamic Con-PET (Song et al., 2023) combines pre-selection and prediction with task-specific LoRA modules to prevent forgetting, ensuring scalable and effective adaptation of LLMs to new tasks.

Merging-based methods, on the other hand, involve combining multiple models or layers to achieve general improvements, including but not limited to merging multiple generic and specialized model weights into a single model (Wortsman et al., 2022; Ilharco et al., 2022; Yu et al., 2023a; Yadav et al., 2024), through mixture-of-expert approach (Ding et al., 2024) or even layer-wise merging and up-scaling such as EvoLLM (Akiba et al., 2024).

6.2 In-Context

In addition to directly updating model parameters, another approach is to leverage the in-context capabilities of LLMs to learn from experiences, thereby

enabling fast adaptive updates without expensive training costs. The methods could be divided into updating external and working memory.

External Memory This approach utilizes an external module to collect, update, and retrieve past experiences and knowledge, enabling models to access a rich pool of insights and achieve better results without updating model parameters. The external memory mechanism is common in AI Agent systems (Xu et al., 2023b; Qian et al., 2024; Wang et al., 2024d). This section provides a detailed overview of the latest methods for updating external memory, emphasizing the aspects of memory *Content* and *Updating Operations*, and summarized in Table 2.

Content: External memory mainly stores two type of content: past experiences and reflected rationale, each serving distinct purposes. For instance, past experience provide valuable historical context, serving as a guiding force toward achieving improved outcomes. MoT (Li and Qiu, 2023) archives filtered question-answer pairs to construct a beneficial memory repository. Additionally, the FIFO Queue mechanism in MemGPT (Packer et al., 2023) maintains a rolling history of messages, encapsulating interactions between agents and users, system notifications, and inputs and outputs of function calls.

On the other hand, reflected rationales offer condensed explanations, such as rules, that support decision-making. For instance, TRAN (Yang et al., 2023b) archives rules inferred from experiences alongside information on mistakes to miti-

METHOD	CONTENT	OPERATION
MoT (Li and Qiu, 2023)	Experience	Insert
TRAN (Yang et al., 2023b)	Rationale	Insert, Reflect
MemoryBank (Zhong et al., 2024b)	Experience, Rationale	Insert, Reflect, Forget
MemGPT (Packer et al., 2023)	Experience	Insert, Forget
TiM (Liu et al., 2023a)	Rationale	Insert
IML (Wang et al., 2024a)	Rationale	Insert, Reflect
ICE (Qian et al., 2024)	Rationale	Insert, Reflect
AesopAgent (Wang et al., 2024d)	Experience, Rationale	Insert, Reflect

Table 2: Content and operations of updating external memory.

gate future errors. Correspondingly, TiM (Liu et al., 2023a) preserves inductive reasoning, defined as text elucidating the relationships between entities. Moreover, IML (Wang et al., 2024a) and ICE (Qian et al., 2024) store comprehensive notes and rules derived from a series of trajectories, demonstrating the broad spectrum of content types that memory systems can accommodate.

MemoryBank (Zhong et al., 2024b) and AesopAgent (Wang et al., 2024d) establish both experience and reflection knowledge stores, which are the integration of both two memories.

Updating Operation: We categorize the operations to the memory into Insert, Reflect, and Forget. The most common operation is insert, methods insert text content into the memory for storage (Li and Qiu, 2023; Yang et al., 2023b; Zhong et al., 2024b; Packer et al., 2023; Liu et al., 2023a; Wang et al., 2024a). Another operation is reflection, which is to think and summarize previous experiences to conceptualize rules and knowledge for future use (Yang et al., 2023b; Zhong et al., 2024b; Wang et al., 2024a; Qian et al., 2024). Last, due to the limited storage of memory, forgetting content is crucial to keeping memory efficient and the content valid. MemGPT (Packer et al., 2023) adopts the FIFO queue to forget the contents. MemoryBank (Zhong et al., 2024b) establishes a forgetting curve on the insert time of each item.

Working Memory The methods use past experience to evolve the capabilities of agents by updating internal memory streams, states, or beliefs, known as working memory, often in the form of verbal cues. Reflexion (Shinn et al., 2023) introduces verbal reinforcement learning for decision-making improvement without conventional model updates. Similarly, IML (Wang et al., 2024a) enables LLM-based agents to autonomously learn and adapt to

their environment by summarizing, refining, and updating knowledge based on past experience directly in working memory.

EvolutionaryAgent (Li et al., 2024c) aligns agents with dynamically changing social norms through evolution and selection principles, leveraging environmental feedback for self-evolution. Agent-Pro (Zhang et al., 2024d) employs policy-level reflection and optimization, allowing agents to adapt their behavior and beliefs in interactive scenarios based on past outcomes. Lastly, ProAgent (Zhang et al., 2024a) enhances cooperation in multi-agent systems by dynamically interpreting teammates' intentions and adapting behavior.

These collective works demonstrate the importance of integrating past experiences and knowledge into the agents' memory stream to refine their state or beliefs for improved performance and adaptability across various tasks and environments.

7 Evaluation

Much like the human learning process, it is essential to ascertain whether the present level of ability is adequate and meets the application requirements through evaluation. Furthermore, it is from these evaluations that one can identify the direction for future learning. However, how to accurately assess the performance of an evolved model and provide directions for future improvements is a crucial yet underexplored research area. For a given evolved model M^t , we conceptualize the evaluation process as follows:

$$\mathcal{E}^{t+1}, \mathcal{S}^{t+1} = f^{\mathcal{E}}(M^t, \mathcal{E}^t, \text{ENV}), \quad (7)$$

where $f^{\mathcal{E}}$ represents the evaluation function that measures the performance score (\mathcal{S}^{t+1}) of the current model and provide evolving goal (\mathcal{E}^{t+1}) for the next iteration. Evaluation function $f^{\mathcal{E}}$ can be

categorized into quantitative and qualitative approaches, each providing valuable insights into model performance and areas for improvement.

7.1 Quantitative Evaluation

This method focuses on providing measurable metrics to reliably assess LLM performance, such as automatic (Papineni et al., 2002; Lin, 2004) and human evaluation. However, traditional automatic metrics struggle to accurately evaluate increasingly complex tasks, and human assessment is not an ideal option for autonomous self-evolution. Recent trends use LLMs as human proxy for automatic evaluators, offering cost-effective and scalable solutions for evaluations.

For example, reward model score has been widely used to measure model or task performances (Shinn et al., 2024) and select the best checkpoint (Ouyang et al., 2022). LLM-as-a-judge (Zheng et al., 2024a) using LLMs to evaluate LLMs, employing methods like pairwise comparison, single answer grading, and reference-guided grading. It shows that LLMs can closely match human judgment, enabling efficient large-scale evaluations.

7.2 Qualitative Evaluation

Qualitative evaluation involves case studies and analysis to derive insights, offering evolving guidance for subsequent iterations. Initiatives such as LLM-as-a-judge (Zheng et al., 2024a) provide the reasoning behind its assessments; ChatEval (Chan et al., 2023) explores the strengths and weaknesses of model outputs through debate mechanisms. Furthermore, TRAN (Yang et al., 2023b) leverages past errors to formulate rules that enhance future LLM performances. Nonetheless, compared with instance-level critic or reflection, qualitative evaluation at the task- or model-level still needs comprehensive investigation.

8 Open Problems

8.1 Objectives: Diversity and Hierarchy

Section 3 summarizes existing evolution objectives and their coverage. Nonetheless, these highlighted objectives can only satisfy a small fraction of the vast human needs. The extensive application of LLMs across various tasks and industries highlights unresolved challenges in establishing self-evolution frameworks for evolving objectives that can comprehensively address a broader spectrum of real-world tasks (Eloundou et al., 2023).

Furthermore, the concept of evolving objectives entails a potential hierarchical structure; for instance, UltraTool (Huang et al., 2024b) and T-Eval (Chen et al., 2023d) categorize the capability of tool usage into various sub-dimensions. Exploring evolutionary objectives into manageable sub-goals and pursuing them individually emerges as a viable strategy.

Overall, a clear and urgent need exists to develop self-evolution frameworks that effectively address diversified and hierarchical objectives.

8.2 Level of Autonomy: From Low to High

Self-evolution in large models is emerging, yet lacks clear definitions for its autonomous levels. We categorize self-evolution into three tiers: low, medium, and high-level autonomy

Low-level In this level, the user predefined the evolving object \mathcal{E} and it remains unchanged. The user needs to design the evolving pipeline, namely all modules f^\bullet , on its own. Then, the model completes the self-evolution process based on the designed framework. We denote this level of self-evolution in the following formula:

$$\tilde{M} = \text{Evol}^L(M, \mathcal{E}, f^\bullet, \text{ENV}), \quad (8)$$

where M denotes the model to be evolved. \tilde{M} is the evolving output. ENV is the environment. Most of the current works lie at this level.

Medium-level In this level, the user only sets the evolving object \mathcal{E} and keeps it unchanged. The user doesn't need to design the specific modules f^\bullet in the framework. The model can construct each module f^\bullet independently for self-evolution. This level denotes as follows:

$$\tilde{M} = \text{Evol}^M(M, \mathcal{E}, \text{ENV}), \quad (9)$$

High-level In the final level, the model diagnoses its deficiency and constructs the self-evolution methods to improve itself. This is the ultimate purpose of self-evolution. The user model sets its own evolving object \mathcal{E} according to the evaluation f^E output. The evolving objective would change during the iteration. Besides, the model designs the specific modules f^\bullet in the framework. We represent this level as:

$$\tilde{M} = \text{Evol}^H(M, \text{ENV}), \quad (10)$$

As discussed in previous open problem (§ 8.1), there are a large of unfulfilled objectives. However, most of the existing self-evolution frameworks are at the Low-level which requires specifically designed modules (Yuan et al., 2024; Lu et al., 2024a; Qiao et al., 2024). These frameworks are objective-dependent and rely on large human efforts to develop. Exhausting all objectives are not deployment-efficient which brings about the urgent need to develop medium and high levels self-evolution frameworks. At the medium level, it doesn't require expert efforts to design specific modules. LLMs can self-evolve according to targeted objectives. Then at the high level, LLMs can investigate their current deficiencies and evolve in a targeted manner. In all, developing highly autonomous self-evolution frameworks remains an open problem.

8.3 Experience Acquisition and Refinement: From Empirical to Theoretical

Suppose we have addressed the previous two challenges that we have developed promising self-evolution frameworks, the exploration of self-evolution LLMs lacks solid theoretical grounding. This idea posits that LLMs can self-improve or correct their outputs, with or without feedback from the environment. However, the mechanisms behind it remain unclear. Studies show mixed results: Huang et al. (2023) observed self-corrective behavior in models with over 22 billion parameters, while Ganguli et al. (2023) finds LLMs struggle to self-correct reasoning errors without external feedback.

A related challenge is the use of self-generated data for learning. Critics argue this approach could reduce linguistic diversity (Guo et al., 2023) and lead to "model collapse," where models fail to capture complex, long-tailed data distributions (Shumailov et al., 2023). Furthermore, Alemohammad et al. (2023) reveal that generative models trained on their synthetic outputs progressively lose output quality and diversity. Fu et al. (2024) extend this by theoretically analyzing the impact of self-consuming training loops on model performance, emphasizing the importance of balancing synthetic and real data to mitigate error accumulation.

Recent studies (Yang et al., 2024c; Singh et al., 2023) also show that current methods struggle to improve after more than three rounds of self-evolution. One hypothesized reason is that the self-critic of LLM has not co-evolved with the evolving

objective, but more experimental and theoretical support is still needed. These findings highlight a pressing need for more theoretical exploration in self-evolving LLMs. Addressing these concerns is crucial for advancing the field and ensuring that models can effectively learn and improve over time.

8.4 Updating: Stability-Plasticity Dilemma

The stability-plasticity dilemma represents a crucial yet unresolved challenge that is essential for iterative self-evolution. This dilemma reflects the difficulty of balancing the need to retain previously learned information (stability) while adapting to new data or tasks (plasticity). Existing LLMs either overlook this issue or adopt conventional methods that may be ineffective. While training models from scratch could mitigate the problem of catastrophic forgetting, it is highly inefficient, particularly as model parameters increase exponentially and autonomous learning capabilities advance. Finding a balance between acquiring new skills and preserving existing knowledge is crucial for achieving effective and efficient self-evolution, leading to overall improvement.

8.5 Evaluation: Systematic and Evolving

To effectively assess LLMs, a dynamic, comprehensive benchmark is crucial. This becomes even more pivotal as we progress towards Artificial General Intelligence (AGI). Traditional static benchmarks risk obsolescence due to LLMs' evolving nature and potential access to test data through interacting with environments, such as search engines, undermining their reliability. A dynamic benchmark, like Sotopia (Zhou et al., 2023a), proposes a solution by creating an LLM-based environment tailored for evaluating the social intelligence of LLMs, thereby avoiding the limitations posed by static benchmarks.

8.6 Safety and Superalignment

The advancement of LLMs opens the possibility for AI systems to achieve or even surpass expert-level capabilities in both supportive and autonomous decision-making. For safety, ensuring these LLMs align with human values and preferences is crucial, particularly to mitigate inherent biases that can impact areas such as political debates, as highlighted by Taubenfeld et al. (2024). OpenAI's initiative, Superalignment (Leike and Sutskever, 2023), aims to align a superintelligence by developing scalable training methods, validating models for alignment,

and stress-testing the alignment process through scalable oversight (Saunders et al., 2022), robustness (Perez et al., 2022), automated interpretability (Bills et al., 2023), and adversarial testing. Although challenges remain, Superalignment marks an initial attempt to develop a self-evolving LLM that closely aligns with human ethics and values in a scalable way.

9 Conclusion

The evolution of LLMs towards self-evolution paradigms represents a transformative shift in artificial intelligence akin to the human learning process. It is promising to overcome the limitations of current models that rely heavily on human annotation and teacher models. This survey presents a comprehensive framework for understanding and developing self-evolving LLMs, structured around iterative cycles of experience acquisition, refinement, updating, and evaluation. By detailing advancements and categorizing the evolution objectives within this framework, we offer a thorough overview of current methods and highlight the potential for LLMs to adapt, learn, and improve autonomously. We also identify existing challenges and propose directions for future research, aiming to accelerate the progress toward more dynamic, intelligent, and efficient models. This work deepens the understanding of self-evolving LLMs. It paves the way for significant advancements in AI, marking a step towards achieving superintelligent systems capable of surpassing human performance in complex real-world tasks.

Acknowledgements

This work was supported by Alibaba Group through Alibaba Research Intern Program.

References

- Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.
- Janice Ahn, Rishu Verma, Renze Lou, Di Liu, Rui Zhang, and Wenpeng Yin. 2024. Large language models for mathematical reasoning: Progresses and challenges. *arXiv preprint arXiv:2402.00157*.
- Takuya Akiba, Makoto Shing, Yujin Tang, Qi Sun, and David Ha. 2024. Evolutionary optimization of model merging recipes. *arXiv preprint arXiv:2403.13187*.
- Renat Aksitov, Sobhan Miryoosefi, Zonglin Li, Daliang Li, Sheila Babayan, Kavya Kopparapu, Zachary Fisher, Ruiqi Guo, Sushant Prakash, Pranesh Srinivasan, et al. 2023. Rest meets react: Self-improvement for multi-step reasoning llm agent. *arXiv preprint arXiv:2312.10003*.
- Sina Alemohammad, Josue Casco-Rodriguez, Lorenzo Luzi, Ahmed Imtiaz Humayun, Hossein Babaei, Daniel LeJeune, Ali Siahkoohi, and Richard Baraniuk. 2023. Self-consuming generative models go mad. In *The Twelfth International Conference on Learning Representations*.
- Arian Askari, Roxana Petcu, Chuan Meng, Mohammad Aliannejadi, Amin Abolghasemi, Evangelos Kanoulas, and Suzan Verberne. 2024. Self-seeding and multi-intent self-instructing llms for generating intent-aware information-seeking dialogs. *arXiv preprint arXiv:2402.11633*.
- Thomas Bäck and Hans-Paul Schwefel. 1993. An overview of evolutionary algorithms for parameter optimization. *Evolutionary computation*, 1(1):1–23.
- Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.
- Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. 2022. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*.
- Maciej Besta, Nils Blach, Ales Kubicek, Robert Gerstenberger, Michal Podstawski, Lukas Gianinazzi, Joanna Gajda, Tomasz Lehmann, Hubert Niewiadomski, Piotr Nyczyk, et al. 2024. Graph of thoughts: Solving elaborate problems with large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 17682–17690.
- Steven Bills, Nick Cammarata, Dan Mossing, Henk Tillman, Leo Gao, Gabriel Goh, Ilya Sutskever, Jan Leike, Jeff Wu, and William Saunders. 2023. Language models can explain neurons in language models. URL <https://openaipublic.blob.core.windows.net/neuron-explainer/paper/index.html>. (Date accessed: 14.05. 2023).
- David Boud, Rosemary Keogh, and David Walker. 2013. *Reflection: Turning experience into learning*. Routledge.
- Konstantinos Bousmalis, Giulia Vezzani, Dushyant Rao, Coline Manon Devin, Alex X Lee, Maria Bauza Villalonga, Todor Davchev, Yuxiang Zhou, Agrim Gupta, Akhil Raju, et al. 2023. Robocat: A self-improving generalist agent for robotic manipulation. *Transactions on Machine Learning Research*.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind

- Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.
- Collin Burns, Pavel Izmailov, Jan Hendrik Kirchner, Bowen Baker, Leo Gao, Leopold Aschenbrenner, Yining Chen, Adrien Ecoffet, Manas Joglekar, Jan Leike, et al. 2023. Weak-to-strong generalization: Eliciting strong capabilities with weak supervision. *arXiv preprint arXiv:2312.09390*.
- David J Chalmers. 1997. *The conscious mind: In search of a fundamental theory*. Oxford Paperbacks.
- Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. 2023. Chateval: Towards better llm-based evaluators through multi-agent debate. In *The Twelfth International Conference on Learning Representations*.
- Bei Chen, Fengji Zhang, Anh Nguyen, Daoguang Zan, Zeqi Lin, Jian-Guang Lou, and Weizhu Chen. 2022. Codet: Code generation with generated tests. *arXiv preprint arXiv:2207.10397*.
- Lichang Chen, Shiyang Li, Jun Yan, Hai Wang, Kalpa Gunaratna, Vikas Yadav, Zheng Tang, Vijay Srinivasan, Tianyi Zhou, Heng Huang, et al. 2023a. Alpagasus: Training a better alpaca with fewer data. *arXiv preprint arXiv:2307.08701*.
- Pinzhen Chen, Zhicheng Guo, Barry Haddow, and Kenneth Heafield. 2023b. Iterative translation refinement with large language models. *arXiv preprint arXiv:2306.03856*.
- Weixin Chen and Bo Li. 2024. Grath: Gradual self-truthifying for large language models. *arXiv preprint arXiv:2401.12292*.
- Xinyun Chen, Maxwell Lin, Nathanael Schaeferli, and Denny Zhou. 2023c. Teaching large language models to self-debug. In *The 61st Annual Meeting Of The Association For Computational Linguistics*.
- Zehui Chen, Weihua Du, Wenwei Zhang, Kuikun Liu, Jiangning Liu, Miao Zheng, Jingming Zhuo, Songyang Zhang, Dahua Lin, Kai Chen, et al. 2023d. T-eval: Evaluating the tool utilization capability step by step. *arXiv preprint arXiv:2312.14033*.
- Zixiang Chen, Yihe Deng, Huizhuo Yuan, Kaixuan Ji, and Quanquan Gu. 2024. Self-play fine-tuning converts weak language models to strong language models. *arXiv preprint arXiv:2401.01335*.
- Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.
- Katherine M Collins, Albert Q Jiang, Simon Frieder, Lionel Wong, Miri Zilka, Umang Bhatt, Thomas Lukasiewicz, Yuhuai Wu, Joshua B Tenenbaum, William Hart, et al. 2023. Evaluating language models for mathematics through interactions. *arXiv preprint arXiv:2306.01694*.
- Wanyun Cui and Qianle Wang. 2023. Ada-instruct: Adapting instruction generators for complex reasoning. *arXiv preprint arXiv:2310.04484*.
- Daniel C Dennett. 1993. *Consciousness explained*. Penguin uk.
- Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2024. Qlora: Efficient finetuning of quantized llms. *Advances in Neural Information Processing Systems*, 36.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
- John Dewey. 1938. Experience and education: Kappa delta pi. *International Honor Society in Education*.
- Ning Ding, Yulin Chen, Ganqu Cui, Xingtai Lv, Ruobing Xie, Bowen Zhou, Zhiyuan Liu, and Maosong Sun. 2024. Mastering text, code and math simultaneously via fusing highly specialized language models. *arXiv preprint arXiv:2403.08281*.
- Ning Ding, Yulin Chen, Bokai Xu, Yujia Qin, Zhi Zheng, Shengding Hu, Zhiyuan Liu, Maosong Sun, and Bowen Zhou. 2023. Enhancing chat language models by scaling high-quality instructional conversations. *arXiv preprint arXiv:2305.14233*.
- Yann Dubois, Chen Xuechen Li, Rohan Taori, Tianyi Zhang, Ishaan Gulrajani, Jimmy Ba, Carlos Guestrin, Percy S Liang, and Tatsunori B Hashimoto. 2024. Alpacafarm: A simulation framework for methods that learn from human feedback. *Advances in Neural Information Processing Systems*, 36.
- Tyna Eloundou, Sam Manning, Pamela Mishkin, and Daniel Rock. 2023. Gpts are gpts: An early look at the labor market impact potential of large language models. *arXiv preprint arXiv:2303.10130*.
- Chrisantha Fernando, Dylan Banarse, Henryk Michalewski, Simon Osindero, and Tim Rocktäschel. 2023. Promptbreeder: Self-referential self-improvement via prompt evolution. *arXiv preprint arXiv:2309.16797*.
- Shi Fu, Sen Zhang, Yingjie Wang, Xinmei Tian, and Dacheng Tao. 2024. Towards theoretical understandings of self-consuming generative models. *arXiv preprint arXiv:2402.11778*.
- Deep Ganguli, Amanda Askell, Nicholas Schiefer, Thomas I Liao, Kamilé Lukošiūtė, Anna Chen, Anna Goldie, Azalia Mirhoseini, Catherine Olsson, Danny Hernandez, et al. 2023. The capacity for moral self-correction in large language models. *arXiv preprint arXiv:2302.07459*.

- Yingqiang Ge, Wenyue Hua, Kai Mei, Juntao Tan, Shuyuan Xu, Zelong Li, Yongfeng Zhang, et al. 2024. Openagi: When llm meets domain experts. *Advances in Neural Information Processing Systems*, 36.
- Zelalem Gero, Chandan Singh, Hao Cheng, Tristan Naumann, Michel Galley, Jianfeng Gao, and Hoifung Poon. 2023. Self-verification improves few-shot clinical information extraction. *arXiv preprint arXiv:2306.00024*.
- Zhibin Gou, Zhihong Shao, Yeyun Gong, Yelong Shen, Yujiu Yang, Nan Duan, and Weizhu Chen. 2023. Critic: Large language models can self-correct with tool-interactive critiquing. *arXiv preprint arXiv:2305.11738*.
- Caglar Gulcehre, Tom Le Paine, Srivatsan Srinivasan, Ksenia Konyushkova, Lotte Weerts, Abhishek Sharma, Aditya Siddhant, Alex Ahern, Miaosen Wang, Chenjie Gu, et al. 2023. Reinforced self-training (rest) for language modeling. *arXiv preprint arXiv:2308.08998*.
- Yanzhu Guo, Guokan Shang, Michalis Vazirgiannis, and Chloé Clavel. 2023. The curious decline of linguistic diversity: Training language models on synthetic text. *arXiv preprint arXiv:2311.09807*.
- Anil K Gupta, Ken G Smith, and Christina E Shalley. 2006. The interplay between exploration and exploitation. *Academy of management journal*, 49(4):693–706.
- Joshua Hare. 2019. Dealing with sparse rewards in reinforcement learning. *arXiv preprint arXiv:1910.09281*.
- Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.
- John H Holland. 1992. *Adaptation in natural and artificial systems: an introductory analysis with applications to biology, control, and artificial intelligence*. MIT press.
- Or Honovich, Thomas Scialom, Omer Levy, and Timo Schick. 2022. Unnatural instructions: Tuning language models with (almost) no human labor. *arXiv preprint arXiv:2212.09689*.
- Arian Hosseini, Xingdi Yuan, Nikolay Malkin, Aaron Courville, Alessandro Sordoni, and Rishabh Agarwal. 2024. V-star: Training verifiers for self-taught reasoners. *arXiv preprint arXiv:2402.06457*.
- Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. 2021. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.
- Jianheng Huang, Leyang Cui, Ante Wang, Chengyi Yang, Xinting Liao, Linfeng Song, Junfeng Yao, and Jinsong Su. 2024a. Mitigating catastrophic forgetting in large language models with self-synthesized rehearsal. *arXiv preprint arXiv:2403.01244*.
- Jiaxin Huang, Shixiang Shane Gu, Le Hou, Yuexin Wu, Xuezhi Wang, Hongkun Yu, and Jiawei Han. 2022. Large language models can self-improve. *arXiv preprint arXiv:2210.11610*.
- Jie Huang, Xinyun Chen, Swaroop Mishra, Huaixiu Steven Zheng, Adams Wei Yu, Xinying Song, and Denny Zhou. 2023. Large language models cannot self-correct reasoning yet. In *The Twelfth International Conference on Learning Representations*.
- Shijue Huang, Wanjun Zhong, Jianqiao Lu, Qi Zhu, Jiahui Gao, Weiwen Liu, Yutai Hou, Xingshan Zeng, Yasheng Wang, Lifeng Shang, et al. 2024b. Planning, creation, usage: Benchmarking llms for comprehensive tool utilization in real-world complex scenarios. *arXiv preprint arXiv:2401.17167*.
- Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortzman, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. 2022. Editing models with task arithmetic. In *The Eleventh International Conference on Learning Representations*.
- Shuyang Jiang, Yuhao Wang, and Yu Wang. 2023. Self-evolve: A code evolution framework via large language models. *arXiv preprint arXiv:2306.02907*.
- Geunwoo Kim, Pierre Baldi, and Stephen McAleer. 2023. Language models can solve computer tasks. *arXiv preprint arXiv:2303.17491*.
- James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. 2017. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526.
- Kelvin JL Koa, Yunshan Ma, Ritchie Ng, and Tat-Seng Chua. 2024. Learning to generate explainable stock predictions using self-reflective large language models. *arXiv preprint arXiv:2402.03659*.
- Nicholas Lee, Thanakul Wattanawong, Sehoon Kim, Karttikeya Mangalam, Sheng Shen, Gopala Anumanchipali, Michael W Mahoney, Kurt Keutzer, and Amir Gholami. 2024. Llm2llm: Boosting llms with novel iterative data enhancement. *arXiv preprint arXiv:2403.15042*.
- Jan Leike and Ilya Sutskever. 2023. [Introducing super-alignment](#). Accessed: 2024-04-01.
- Jia Li, Ge Li, Xuanming Zhang, Yihong Dong, and Zhi Jin. 2024a. Evocodebench: An evolving code generation benchmark aligned with real-world code repositories. *arXiv preprint arXiv:2404.00599*.
- Ming Li, Lichang Chen, Juhai Chen, Shuai He, Jiuxiang Gu, and Tianyi Zhou. 2024b. Selective reflection-tuning: Student-selected data recycling for llm instruction-tuning. *arXiv preprint arXiv:2402.10110*.

- Ming Li, Yong Zhang, Zhitao Li, Juhai Chen, Lichang Chen, Ning Cheng, Jianzong Wang, Tianyi Zhou, and Jing Xiao. 2023a. From quantity to quality: Boosting llm performance with self-guided data selection for instruction tuning. *arXiv preprint arXiv:2308.12032*.
- Shimin Li, Tianxiang Sun, and Xipeng Qiu. 2024c. Agent alignment in evolving social norms. *arXiv preprint arXiv:2401.04620*.
- Xian Li, Ping Yu, Chunting Zhou, Timo Schick, Luke Zettlemoyer, Omer Levy, Jason Weston, and Mike Lewis. 2023b. Self-alignment with instruction back-translation. *arXiv preprint arXiv:2308.06259*.
- Xiaonan Li and Xipeng Qiu. 2023. Mot: Memory-of-thought enables chatgpt to self-improve. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6354–6374.
- Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.
- Yong Lin, Hangyu Lin, Wei Xiong, Shizhe Diao, Jianmeng Liu, Jipeng Zhang, Rui Pan, Haoxiang Wang, Wenbin Hu, Hanning Zhang, Hanze Dong, Renjie Pi, Han Zhao, Nan Jiang, Heng Ji, Yuan Yao, and Tong Zhang. 2024. [Mitigating the alignment tax of rlhf](#). In *arXiv*.
- Aiwei Liu, Haoping Bai, Zhiyun Lu, Xiang Kong, Simon Wang, Jilong Shan, Meng Cao, and Lijie Wen. 2024a. Direct large language model alignment through self-rewarding contrastive prompt distillation. *arXiv preprint arXiv:2402.11907*.
- Jiawei Liu, Chunqiu Steven Xia, Yuyao Wang, and Lingming Zhang. 2024b. Is your code generated by chatgpt really correct? rigorous evaluation of large language models for code generation. *Advances in Neural Information Processing Systems*, 36.
- Lei Liu, Xiaoyan Yang, Yue Shen, Binbin Hu, Zhiqiang Zhang, Jinjie Gu, and Guannan Zhang. 2023a. Think-in-memory: Recalling and post-thinking enable llms with long-term memory. *arXiv preprint arXiv:2311.08719*.
- Xiao Liu, Hao Yu, Hanchen Zhang, Yifan Xu, Xuyanyu Lei, Hanyu Lai, Yu Gu, Hangliang Ding, Kaiwen Men, Kejuan Yang, et al. 2023b. Agent-bench: Evaluating llms as agents. *arXiv preprint arXiv:2308.03688*.
- Yuqiao Liu, Yanan Sun, Bing Xue, Mengjie Zhang, Gary G Yen, and Kay Chen Tan. 2021. A survey on evolutionary neural architecture search. *IEEE transactions on neural networks and learning systems*, 34(2):550–570.
- Shayne Longpre, Le Hou, Tu Vu, Albert Webson, Hyung Won Chung, Yi Tay, Denny Zhou, Quoc V Le, Barret Zoph, Jason Wei, et al. 2023. The flan collection: Designing data and methods for effective instruction tuning. In *International Conference on Machine Learning*, pages 22631–22648. PMLR.
- Jianqiao Lu, Wanjun Zhong, Wenyong Huang, Yufei Wang, Fei Mi, Baojun Wang, Weichao Wang, Lifeng Shang, and Qun Liu. 2023. Self: Language-driven self-evolution for large language model. *arXiv preprint arXiv:2310.00533*.
- Keming Lu, Bowen Yu, Chang Zhou, and Jingren Zhou. 2024a. Large language models are superpositions of all characters: Attaining arbitrary role-play via self-alignment. *arXiv preprint arXiv:2401.12474*.
- Xinyu Lu, Bowen Yu, Yaojie Lu, Hongyu Lin, Haiyang Yu, Le Sun, Xianpei Han, and Yongbin Li. 2024b. Sofa: Shielded on-the-fly alignment via priority rule following. *arXiv preprint arXiv:2402.17358*.
- Ziyang Luo, Can Xu, Pu Zhao, Qingfeng Sun, Xubo Geng, Wenxiang Hu, Chongyang Tao, Jing Ma, Qingwei Lin, and Daxin Jiang. 2024. [Wizardcoder: Empowering code large language models with evol-instruct](#). In *The Twelfth International Conference on Learning Representations*.
- Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. 2023. Self-refine: Iterative refinement with self-feedback. *arXiv preprint arXiv:2303.17651*.
- Santiago Miret and NM Krishnan. 2024. Are llms ready for real-world materials discovery? *arXiv preprint arXiv:2402.05200*.
- Michael Noukhovitch, Samuel Lavoie, Florian Strub, and Aaron C Courville. 2024. Language model alignment with elastic reset. *Advances in Neural Information Processing Systems*, 36.
- Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.
- Charles Packer, Vivian Fang, Shishir G Patil, Kevin Lin, Sarah Wooders, and Joseph E Gonzalez. 2023. Memgpt: Towards llms as operating systems. *arXiv preprint arXiv:2310.08560*.
- Jing-Cheng Pang, Pengyuan Wang, Kaiyuan Li, Xiong-Hui Chen, Jiacheng Xu, Zongzhang Zhang, and Yang Yu. 2023. Language model self-improvement by reinforcement learning contemplation. *arXiv preprint arXiv:2305.14483*.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

- Keqin Peng, Liang Ding, Qihuang Zhong, Yuanxin Ouyang, Wenge Rong, Zhang Xiong, and Dacheng Tao. 2023. Token-level self-evolution training for sequence-to-sequence learning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 841–850.
- Ethan Perez, Saffron Huang, Francis Song, Trevor Cai, Roman Ring, John Aslanides, Amelia Glaese, Nat McAleese, and Geoffrey Irving. 2022. Red teaming language models with language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3419–3448.
- Cheng Qian, Shihao Liang, Yujia Qin, Yining Ye, Xin Cong, Yankai Lin, Yesai Wu, Zhiyuan Liu, and Maosong Sun. 2024. Investigate-consolidate-exploit: A general strategy for inter-task agent self-evolution. *arXiv preprint arXiv:2401.13996*.
- Shuofei Qiao, Ningyu Zhang, Runnan Fang, Yujie Luo, Wangchunshu Zhou, Yuchen Eleanor Jiang, Chengfei Lv, and Huajun Chen. 2024. Autoact: Automatic agent learning from scratch via self-planning. *arXiv preprint arXiv:2401.05268*.
- Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67.
- Alexandre Ramé, Nino Vieillard, Léonard Hussenot, Robert Dadashi, Geoffrey Cideron, Olivier Bachem, and Johan Ferret. 2024. Warm: On the benefits of weight averaged reward models. *arXiv preprint arXiv:2401.12187*.
- Baptiste Roziere, Jonas Gehring, Fabian Gloeckle, Sten Sootla, Itai Gat, Xiaoqing Ellen Tan, Yossi Adi, Jingyu Liu, Tal Remez, Jérémie Rapin, et al. 2023. Code llama: Open foundation models for code. *arXiv preprint arXiv:2308.12950*.
- William Saunders, Catherine Yeh, Jeff Wu, Steven Bills, Long Ouyang, Jonathan Ward, and Jan Leike. 2022. Self-critiquing models for assisting human evaluators. *arXiv preprint arXiv:2206.05802*.
- Philipp Schoenegger, Peter S Park, Ezra Karger, and Philip E Tetlock. 2024. Ai-augmented predictions: Llm assistants improve human forecasting accuracy. *arXiv preprint arXiv:2402.07862*.
- Donald A Schön. 2017. *The reflective practitioner: How professionals think in action*. Routledge.
- John R Searle. 1986. *Minds, brains and science*. Harvard university press.
- Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2024. Reflexion: Language agents with verbal reinforcement learning. *Advances in Neural Information Processing Systems*, 36.
- Noah Shinn, Beck Labash, and Ashwin Gopinath. 2023. Reflexion: an autonomous agent with dynamic memory and self-reflection. *arXiv preprint arXiv:2303.11366*.
- Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Yarin Gal, Nicolas Papernot, and Ross Anderson. 2023. The curse of recursion: Training on generated data makes models forget. *arXiv preprint arXiv:2305.17493*.
- David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. 2016. Mastering the game of go with deep neural networks and tree search. *nature*, 529(7587):484–489.
- David Silver, Thomas Hubert, Julian Schrittwieser, Ioannis Antonoglou, Matthew Lai, Arthur Guez, Marc Lanctot, Laurent Sifre, Dharshan Kumaran, Thore Graepel, et al. 2017. Mastering chess and shogi by self-play with a general reinforcement learning algorithm. *arXiv preprint arXiv:1712.01815*.
- Avi Singh, John D Co-Reyes, Rishabh Agarwal, Ankesh Anand, Piyush Patil, Peter J Liu, James Harrison, Jaehoon Lee, Kelvin Xu, Aaron Parisi, et al. 2023. Beyond human data: Scaling self-training for problem-solving with language models. *arXiv preprint arXiv:2312.06585*.
- Chenyang Song, Xu Han, Zheni Zeng, Kuai Li, Chen Chen, Zhiyuan Liu, Maosong Sun, and Tao Yang. 2023. Conpet: Continual parameter-efficient tuning for large language models. *arXiv preprint arXiv:2309.14763*.
- Yifan Song, Da Yin, Xiang Yue, Jie Huang, Sujian Li, and Bill Yuchen Lin. 2024. Trial and error: Exploration-based trajectory optimization for llm agents. *arXiv preprint arXiv:2403.02502*.
- Wolfgang Stammer, Felix Friedrich, David Steinmann, Hikaru Shindo, and Kristian Kersting. 2023. Learning by self-explaining.
- Zhiqing Sun, Yikang Shen, Hongxin Zhang, Qinhong Zhou, Zhenfang Chen, David Cox, Yiming Yang, and Chuang Gan. 2023. Salmon: Self-alignment with principle-following reward models. *arXiv preprint arXiv:2310.05910*.
- Zhiqing Sun, Yikang Shen, Qinhong Zhou, Hongxin Zhang, Zhenfang Chen, David Cox, Yiming Yang, and Chuang Gan. 2024. Principle-driven self-alignment of language models from scratch with minimal human supervision. *Advances in Neural Information Processing Systems*, 36.
- Yiming Tan, Dehai Min, Yu Li, Wenbo Li, Nan Hu, Yongrui Chen, and Guilin Qi. 2023. Evaluation of chatgpt as a question answering system for answering complex questions. *arXiv preprint arXiv:2303.07992*.

- Zhengwei Tao, Xiancai Chen, Zhi Jin, Xiaoying Bai, Haiyan Zhao, and Yiwei Lou. 2024a. [Evit: Event-oriented instruction tuning for event reasoning](#).
- Zhengwei Tao, Zhi Jin, Junqiang Huang, Xiancai Chen, Xiaoying Bai, Haiyan Zhao, Yifan Zhang, and Chongyang Tao. 2024b. [Meel: Multi-modal event evolution learning](#).
- Amir Taubenfeld, Yaniv Dover, Roi Reichart, and Ariel Goldstein. 2024. Systematic biases in llm simulations of debates. *arXiv preprint arXiv:2402.04049*.
- Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.
- Yongqi Tong, Dawei Li, Sizhe Wang, Yujia Wang, Fei Teng, and Jingbo Shang. 2024. Can llms learn from previous mistakes? investigating llms' errors to boost for reasoning. *arXiv preprint arXiv:2403.20046*.
- Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajwal Bhargava, Shruti Bhosale, et al. 2023a. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.
- Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajwal Bhargava, Shruti Bhosale, et al. 2023b. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.
- Tao Tu, Anil Palepu, Mike Schaekermann, Khaled Saab, Jan Freyberg, Ryutaro Tanno, Amy Wang, Brenna Li, Mohamed Amin, Nenad Tomasev, et al. 2024. Towards conversational diagnostic ai. *arXiv preprint arXiv:2401.05654*.
- Dennis Ulmer, Elman Mansimov, Kaixiang Lin, Justin Sun, Xibin Gao, and Yi Zhang. 2024. Bootstrapping llm-based task-oriented dialogue agents via self-talk. *arXiv preprint arXiv:2401.05033*.
- Fanqi Wan, Xinting Huang, Deng Cai, Xiaojun Quan, Wei Bi, and Shuming Shi. 2024. Knowledge fusion of large language models. In *The Twelfth International Conference on Learning Representations*.
- Bo Wang, Tianxiang Sun, Hang Yan, Siyin Wang, Qingyuan Cheng, and Xipeng Qiu. 2024a. In-memory learning: A declarative learning framework for large language models. *arXiv preprint arXiv:2403.02757*.
- Boshi Wang, Hao Fang, Jason Eisner, Benjamin Van Durme, and Yu Su. 2024b. Llms in the imaginarium: Tool learning through simulated trial and error. *arXiv preprint arXiv:2403.04746*.
- Haoyu Wang, Guozheng Ma, Ziqiao Meng, Zeyu Qin, Li Shen, Zhong Zhang, Bingzhe Wu, Liu Liu, Yatao Bian, Tingyang Xu, et al. 2024c. Step-on-feet tuning: Scaling self-alignment of llms via bootstrapping. *arXiv preprint arXiv:2402.07610*.
- Jiuniu Wang, Zehua Du, Yuyuan Zhao, Bo Yuan, Kexiang Wang, Jian Liang, Yaxi Zhao, Yihen Lu, Gengliang Li, Junlong Gao, et al. 2024d. Aesop-agent: Agent-driven evolutionary system on story-to-video production. *arXiv preprint arXiv:2403.07952*.
- Kuan Wang, Yadong Lu, Michael Santacroce, Yeyun Gong, Chao Zhang, and Yelong Shen. 2023a. Adapting llm agents through communication. *arXiv preprint arXiv:2310.01444*.
- Ruiyi Wang, Haofei Yu, Wenxin Zhang, Zhengyang Qi, Maarten Sap, Graham Neubig, Yonatan Bisk, and Hao Zhu. 2024e. Sotopia- π : Interactive learning of socially intelligent language agents. *arXiv preprint arXiv:2403.08715*.
- Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*.
- Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2023b. Self-instruct: Aligning language models with self-generated instructions. In *The 61st Annual Meeting Of The Association For Computational Linguistics*.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.
- Yixuan Weng, Minjun Zhu, Fei Xia, Bin Li, Shizhu He, Shengping Liu, Bin Sun, Kang Liu, and Jun Zhao. 2023. Large language models are better reasoners with self-verification. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 2550–2575.
- Mitchell Wortsman, Gabriel Ilharco, Samir Ya Gadre, Rebecca Roelofs, Raphael Gontijo-Lopes, Ari S Marcos, Hongseok Namkoong, Ali Farhadi, Yair Carmon, Simon Kornblith, et al. 2022. Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In *International conference on machine learning*, pages 23965–23998. PMLR.
- Shengguang Wu, Keming Lu, Benfeng Xu, Junyang Lin, Qi Su, and Chang Zhou. 2023. Self-evolved diverse data sampling for efficient instruction tuning. *arXiv preprint arXiv:2311.08182*.
- Tongtong Wu, Linhao Luo, Yuan-Fang Li, Shirui Pan, Thuy-Trang Vu, and Gholamreza Haffari. 2024. Continual learning for large language models: A survey. *arXiv preprint arXiv:2402.01364*.

- Can Xu, Qingfeng Sun, Kai Zheng, Xiubo Geng, Pu Zhao, Jiazhan Feng, Chongyang Tao, and Dixin Jiang. 2023a. Wizardlm: Empowering large language models to follow complex instructions. *arXiv preprint arXiv:2304.12244*.
- Can Xu, Qingfeng Sun, Kai Zheng, Xiubo Geng, Pu Zhao, Jiazhan Feng, Chongyang Tao, Qingwei Lin, and Dixin Jiang. 2024a. **WizardLM: Empowering large pre-trained language models to follow complex instructions.** In *The Twelfth International Conference on Learning Representations*.
- Jie Xu, Hanbo Zhang, Xinghang Li, Huaping Liu, Xuguang Lan, and Tao Kong. 2024b. Sinvig: A self-evolving interactive visual agent for human-robot interaction. *arXiv preprint arXiv:2402.11792*.
- Yuzhuang Xu, Shuo Wang, Peng Li, Fuwen Luo, Xiaolong Wang, Weidong Liu, and Yang Liu. 2023b. Exploring large language models for communication games: An empirical study on werewolf. *arXiv preprint arXiv:2309.04658*.
- Prateek Yadav, Derek Tam, Leshem Choshen, Colin A Raffel, and Mohit Bansal. 2024. Ties-merging: Resolving interference when merging models. *Advances in Neural Information Processing Systems*, 36.
- Kevin Yang, Dan Klein, Asli Celikyilmaz, Nanyun Peng, and Yuandong Tian. 2023a. Rlcld: Reinforcement learning from contrastive distillation for lm alignment. In *The Twelfth International Conference on Learning Representations*.
- Rui Yang, Lin Song, Yanwei Li, Sijie Zhao, Yixiao Ge, Xiu Li, and Ying Shan. 2024a. Gpt4tools: Teaching large language model to use tools via self-instruction. *Advances in Neural Information Processing Systems*, 36.
- Zeyuan Yang, Peng Li, and Yang Liu. 2023b. Failures pave the way: Enhancing large language models through tuning-free rule accumulation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1751–1777.
- Zhaorui Yang, Qian Liu, Tianyu Pang, Han Wang, Haozhe Feng, Minfeng Zhu, and Wei Chen. 2024b. Self-distillation bridges distribution gap in language model fine-tuning. *arXiv preprint arXiv:2402.13669*.
- Zonghan Yang, Peng Li, Ming Yan, Ji Zhang, Fei Huang, and Yang Liu. 2024c. React meets actre: Autonomous annotations of agent trajectories for contrastive self-training. *arXiv preprint arXiv:2403.14589*.
- Zonghan Yang, An Liu, Zijun Liu, Kaiming Liu, Fangzhou Xiong, Yile Wang, Zeyuan Yang, Qingyuan Hu, Xinrui Chen, Zhenhe Zhang, et al. 2024d. Towards unified alignment between agents, humans, and environment. *arXiv preprint arXiv:2402.07744*.
- Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. 2024. Tree of thoughts: Deliberate problem solving with large language models. *Advances in Neural Information Processing Systems*, 36.
- Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2022. React: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*.
- Le Yu, Bowen Yu, Haiyang Yu, Fei Huang, and Yongbin Li. 2023a. Language models are super mario: Absorbing abilities from homologous models as a free lunch. *arXiv preprint arXiv:2311.03099*.
- Longhui Yu, Weisen Jiang, Han Shi, Jincheng Yu, Zhengying Liu, Yu Zhang, James T Kwok, Zhen-guo Li, Adrian Weller, and Weiyang Liu. 2023b. Metamath: Bootstrap your own mathematical questions for large language models. *arXiv preprint arXiv:2309.12284*.
- Weizhe Yuan, Richard Yuanzhe Pang, Kyunghyun Cho, Sainbayar Sukhbaatar, Jing Xu, and Jason Weston. 2024. Self-rewarding language models. *arXiv preprint arXiv:2401.10020*.
- Zheng Yuan, Hongyi Yuan, Chengpeng Li, Guanting Dong, Chuanqi Tan, and Chang Zhou. 2023. Scaling relationship on learning mathematical reasoning with large language models. *arXiv preprint arXiv:2308.01825*.
- Eric Zelikman, Eliana Lorch, Lester Mackey, and Adam Tauman Kalai. 2023. Self-taught optimizer (stop): Recursively self-improving code generation. *arXiv preprint arXiv:2310.02304*.
- Eric Zelikman, Jesse Mu, Noah D Goodman, and Yuhuai Tony Wu. 2022. Star: Self-taught reasoner bootstrapping reasoning with reasoning. *Advances in Neural Information Processing Systems (NeurIPS)*.
- Ceyao Zhang, Kaijie Yang, Siyi Hu, Zihao Wang, Guanghe Li, Yihang Sun, Cheng Zhang, Zhaowei Zhang, Anji Liu, Song-Chun Zhu, et al. 2024a. Proagent: building proactive cooperative agents with large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 17591–17599.
- Dan Zhang, Ziniu Hu, Sining Zhoubian, Zhengxiao Du, Kaiyu Yang, Zihan Wang, Yisong Yue, Yuxiao Dong, and Jie Tang. 2024b. Sciglm: Training scientific language models with self-reflective instruction annotation and tuning. *arXiv preprint arXiv:2401.07950*.
- Wenqi Zhang, Yongliang Shen, Linjuan Wu, Qiuying Peng, Jun Wang, Yueting Zhuang, and Weiming Lu. 2024c. **Self-contrast: Better reflection through inconsistent solving perspectives.**
- Wenqi Zhang, Ke Tang, Hai Wu, Mengna Wang, Yongliang Shen, Guiyang Hou, Zeqi Tan, Peng Li,

- Yueting Zhuang, and Weiming Lu. 2024d. Agent-pro: Learning to evolve via policy-level reflection and optimization. *arXiv preprint arXiv:2402.17574*.
- Xiaoying Zhang, Baolin Peng, Ye Tian, Jingyan Zhou, Lifeng Jin, Linfeng Song, Haitao Mi, and Helen Meng. 2024e. Self-alignment for factuality: Mitigating hallucinations in llms via self-evaluation. *arXiv preprint arXiv:2402.09267*.
- Haoqi Zheng, Qihuang Zhong, Liang Ding, Zhiliang Tian, Xin Niu, Changjian Wang, Dongsheng Li, and Dacheng Tao. 2023. Self-evolution learning for mixup: Enhance data augmentation on few-shot text classification tasks. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 8964–8974.
- Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2024a. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36.
- Tianyu Zheng, Shuyue Guo, Xingwei Qu, Jiawei Guo, Weixu Zhang, Xinrun Du, Chenghua Lin, Wen-hao Huang, Wenhui Chen, Jie Fu, et al. 2024b. Kun: Answer polishing for chinese self-alignment with instruction back-translation. *arXiv preprint arXiv:2401.06477*.
- Li Zhong, Zilong Wang, and Jingbo Shang. 2024a. Ldb: A large language model debugger via verifying runtime execution step-by-step. *arXiv preprint arXiv:2402.16906*.
- Qihuang Zhong, Liang Ding, Juhua Liu, Bo Du, and Dacheng Tao. 2023. Self-evolution learning for discriminative language model pretraining. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 4130–4145.
- Wanjun Zhong, Lianghong Guo, Qiqi Gao, He Ye, and Yanlin Wang. 2024b. Memorybank: Enhancing large language models with long-term memory. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 19724–19731.
- Chunting Zhou, Pengfei Liu, Puxin Xu, Srinivasan Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, et al. 2024. Lima: Less is more for alignment. *Advances in Neural Information Processing Systems*, 36.
- Xuhui Zhou, Hao Zhu, Leena Mathur, Ruohong Zhang, Haofei Yu, Zhengyang Qi, Louis-Philippe Morency, Yonatan Bisk, Daniel Fried, Graham Neubig, et al. 2023a. Sotopia: Interactive evaluation for social intelligence in language agents. In *The Twelfth International Conference on Learning Representations*.
- Zhehua Zhou, Jiayang Song, Kunpeng Yao, Zhan Shu, and Lei Ma. 2023b. Isr-llm: Iterative self-refined large language model for long-horizon sequential task planning. *arXiv preprint arXiv:2308.13724*.
- Yuqi Zhu, Shuofei Qiao, Yixin Ou, Shumin Deng, Ningyu Zhang, Shiwei Lyu, Yue Shen, Lei Liang, Jinjie Gu, and Huajun Chen. 2024. Knowagent: Knowledge-augmented planning for llm-based agents. *arXiv preprint arXiv:2403.03101*.