LLMs Must Evaluate, Guide and Train Themselves: A Path to Self-Improvement.

Research Proposal

Viktor Moskvoretskii
vvmoskvoretskii@gmail.com

1 Research Problem

Large Language Models (LLMs) have emerged as powerful tools across a wide range of contemporary research applications. However, their development remains highly reliant on human intervention at multiple stages: pretraining on data scraped from the web, fine-tuning using human-generated prompts, and aligning with human-labeled preferences. Each of these stages presents fundamental challenges:

1.1 Data Quantity and Quality

Data Quality. Current research on LLMs heavily relies on scaling models by increasing the amount of training data. However, according to a report by EpochAI [58], we are nearing the exhaustion of available human-generated data. Furthermore, the use of closed-source data poses significant risks, including the potential leakage of personal information and privacy violations. While incorporating other modalities—such as images, audio, and video—could complement LLMs, this approach merely postpones the core issue rather than solving it.

A promising solution lies in the use of synthetic data and self-training methods, which reduce dependence on human-generated data. The need for further scaling is evident, as current models still require more data to improve [57]. However, the data currently available is insufficient, often noisy, and frequently too simplistic for LLMs. Self-training techniques offer a way to overcome both the limitations in data quantity and complexity. As demonstrated in self-play frameworks [51], models can surpass human knowledge by training on their own outputs, offering a potential path forward for LLM advancement.

Data Quantity. In addition to data complexity, another challenge lies in the limitations of the data's scope. Training data is often restricted to plain text or single-turn dialogues, which does not reflect realistic interactions or thought processes. Research has shown that models greatly benefit from labeling the reasoning process, not just the final outcome [35, 49]. Similarly, incorporating examples of self-correction, rather than only focusing on first-turn correct answers, has been shown to improve model performance [70].

These findings highlight the importance of creating more complex and nuanced data that allows models to learn a realistic process of thinking and reasoning. By providing guidance at each step—rather than just at the conclusion of an interaction—the model can better emulate human cognitive processes. Self-training and self-evaluation offer promising avenues for achieving this, especially given the rapid advancements in hardware and the increasing interest in scaling energy resources [57]. These developments make this direction not only feasible but also potentially cost-effective.

1.2 Biases

Reliance on Human Preference. Human-generated data not only suffers from limitations in quantity and scope but also introduces significant biases into models. While pre-training and supervised fine-tuning alone are insufficient for ensuring model safety, recent approaches like alignment with Reinforcement Learning from Human Feedback (RLHF) have emerged as potential solutions [45]. However, models trained on human preferences often act as "vibe-checkers," making them vulnerable to manipulation and exploitation [64, 21]. This issue is evident in reward models, which frequently align with annotators' beliefs rather than prioritizing objective truthfulness [50]. Additionally, these models exhibit biases across various dimensions, including societal, gender, race, age, cultural, and linguistic factors [27, 48].

All of these issues stem from the inherent reliance on human preferences during training. Aside from being costly and limited in availability, these preferences are highly influenced by the individual annotator's beliefs, cultural background, and linguistic nuances. Solutions such as increasing diversity in annotation or improving annotator training are temporary fixes, addressing symptoms rather than the core problem.

Self-Regulation. One creative approach, employed by Anthropic, involves providing models with a set of rules to follow, as demonstrated in their Constitutional AI framework [2]. However, this method is still vulnerable to exploitation by malicious users and remains limited by the constraints of human-authored rules.

A more robust solution would be to enable models to self-regulate, akin to Constitutional AI but with the ability to autonomously modify their own ”constitution”. Through accurate, thorough self-evaluation and bias correction, models could continuously refine themselves. This approach would leverage computational resources to enhance security and minimize susceptibility to biases or attacks, leading to a model that is far more resilient and adaptable.

1.3 Summary

The overall quick summary of problems is as follows:

- **Data Limits:**
  - Human-generated data is nearing exhaustion and is insufficient for further scaling [58, 57].
  - Models benefit from reasoning labels and self-correction examples, but these are scarce and costly in human-generated data [35, 49, 70].
- **Data Biases:**
  - Pre-training and supervised fine-tuning are not enough to ensure safety [45].
  - Models tend to align with annotator beliefs rather than objective truth, exhibiting biases related to societal, gender, race, age, and linguistic attributes [50, 27, 48].
  - Models often act as ”vibe-checkers,” making them vulnerable to manipulation [64, 21].

In alignment with Richard Sutton’s ”Bitter Lesson” [53], addressing these challenges requires shifting the focus from scaling human-annotated data to scaling compute resources, a method recently proven effective [52, 43]. Sutton stresses the importance of self-play as a general method, though it remains underutilized in LLMs. To make significant progress in areas such as reasoning, alignment, and trustworthiness, the focus must be on generalizable methods and compute scaling. The self-play framework, which has been shown to be effective in other domains [51], holds great promise for improving the future of LLMs.

2 Research Proposal

To address the challenges outlined in the Research Problems section, I will provide a proposal divided into logical steps. Each step may reference my previous or ongoing work, highlighted within a dedicated box in each subsection.

The steps are organized in a specific order, as each subsequent step depends on the completion of the previous ones. However, in some cases, initial steps can be conducted independently of this sequence.

2.1 Self-Evaluation and Self-Guidance

Self-Evaluation is a key step in Self-Play, Alignment and Step-Wise labeling. Without proper and unbiased Self-Evaluation, further steps would be pointless, as the labeling would be biased and uninformative.

2.1.1 Background

LLM-as-a-judge appeared and instantly became popular, as those LM evaluators are successfully used in every NLP area: Language Modeling [32, 3], Machine Translation [15], Human Arena [76] and even Text2Image Generation [14]. As well, Alignment with RL from AI Feedback was shown to be cheaper, faster and more controllable [25, 78]. Moreover, studies show that LLM are as good in evaluation as human [11].

The corollary is that if models are able to evaluate well, then they could guide the answers well. Studies propose zero-shot guidance by explicit text refinement [69] or even step-aware guidance [33]. Self-Correction could be also performed without explicit text evaluation with well results [24], however I believe that explicit evaluation would be more profitable. *The Scalable Oversight* studies have shown that even weaker models could be used in debates or consultant regime to enhance quality [22, 23, 5].

Further, Self-Rewarding LLM were shown to surpass DPO baseline with a just prompting itself [71, 34]. This technique could be easily enhanced with making a model better judge with meta-judge [65]. This Self-Guidance could be also applied for Data Curation with implicit model signal, such as gradients, as in my recent paper on Machine Translation, published at EMNLP [40].

The challenges however are serious, as self-evaluation should be evaluated too. From the most recent studies we know that LLMs are biased toward LLM generated output [54], assign themselves higher score and favor their own text style [67]. Moreover, current alignment is biased nearly everywhere: societal, gender, race, age, political, languages and dialects biases.

2.1.2 Self-Evaluation Research Proposals

In the following section I will draw the research plan, covering the problems, discussed above.

Proposal 1: Self-Evaluation Evaluation Framework. The main objective of this research is to develop a robust framework for assessing a model's self-evaluation capabilities across different dimensions, such as safety, reasoning, mathematics, agency, self-correction, instruction-following, and trustworthiness. As well, such framework should assess robustness of self-evaluations. A key aspect of this framework is the generation of reliable Ground Truth labels, which are necessary for evaluating the accuracy of the model's self-assessments.

Evaluation through Diverse Majority Voting. To address this, I propose using a "majority vote" system as the primary method for generating ground truth labels. This system involves either diverse sampling or sampling from multiple frontier models, ensuring model consistency (e.g., GPT-4, Claude) with further allowing them to debate their decision.

By applying such perturbations, the goal is to produce diverse and robust outputs, from which a majority consensus can be derived. The consensus across models could serve as the GT label. The approach generalizes and combines concepts from self-consistency [63] and debating frameworks [23], enhancing them and applying to the task of self-evaluation. Here, the system acts as a "meta-judge," evaluating other model assessments [65].

However, it is crucial to validate this majority vote system against human assessments. Correlating the results of the majority vote with evaluations by human annotators will ensure that the GT labels generated via perturbations reflect realistic and reliable judgments.

Assessing Robustness. – Beyond single-turn evaluations, it is essential to assess the stability of proposed "majority vote" and the model's self-evaluation across different type of perturbations. In particular, we will explore:

- **Multi-Turn Dialogue.** This research will investigate whether the model's self-evaluation abilities degrade or remain consistent over successive dialogue turns. This evaluation will assess the model's capacity for sustained multi-turn interaction, which will later be crucial for self-correction rounds and self-guidance processes.
- **Stability under Paraphrasing.** This test will determine whether the model's evaluations change when the same content is paraphrased. It will assess if the model remains consistent in its judgments across different formulations of the same task, helping to prevent potential exploitation where the model paraphrases to "hack" the evaluation process for reward maximization, a known issue in current reward models [21].
- **Stability under Rare Words and Language Variation.** This evaluation will test the model's performance when rare or uncommon words are introduced, as well as its response across different languages. Understanding the model's robustness in handling less frequent linguistic inputs and language variations will be key to mitigating biases, particularly those related to languages and dialects, as recently identified in reward models [48].
- **Length of Evaluated Span.** This analysis will investigate whether the model's self-assessment abilities are affected by the length of the content it evaluates. The goal is to ensure consistent performance across short, medium, and long inputs, addressing the common issue of length-based bias in evaluation systems [12].

Enforcing Diversity. To create more diverse and fair evaluation, I propose to use a variety of strategically designed role-playing prompts to elicit diverse perspectives, capitalizing on the strong role-playing capabilities of these models [62]. For instance, models might be asked to assume the role of a strict English-speaking judge from London. These roles could be either predefined by humans or generated by GPT itself to ensure robustness. Alternatively, roles could be dynamically sampled by GPT to introduce further variability. Role-Playing showed to be beneficial for alignment cases [46].

Prospective Results. The expected outcomes of this study include the creation of a dataset specifically designed for assessing self-evaluation abilities, the development of a framework for more diverse and fair meta-judging, and the establishment of a benchmark for evaluating current models based on this meta-judging system. Additionally, the robustness evaluations will shed light on the biases present in self-evaluation processes. This study represents a critical first step toward enhancing models self-evaluation abilities, which will, in turn, improve their self-correction, self-guidance, and overall self-training capabilities, ensuring that the reward functions are fair and unbiased.

Proposal 2: Intrinsic Self-Guidance. A promising direction for guiding the model is to examine its intrinsic processes, drawing inspiration from interpretability studies. In this branch In this branch, I propose to further explore how the internals of LLMs can guide models towards more truthful generation, accurate reasoning, and enhanced overall capabilities.

My recent papers, which focus on such techniques, have shown remarkable results. In collaboration with Professors Chris Biemann and Alexander Panchenko, my work on taxonomy learning, published at ACL and COLING [41, 39], demonstrated that LLMs' intrinsic uncertainty can produce state-of-the-art results across multiple taxonomy tasks, showing strong generalization. My paper on machine translation with Professors Samuel Horvath and Chris Biemann, published at EMNLP [40], revealed that LLM intrinsic information can guide data curation by filtering unrelated languages and prioritizing informative ones. Another paper, with Professor Evgeny Burnaev, currently under review, explores how intrinsic representations can identify important LLM weights, leading to a PEFT method based on this information [75]. I also have an ongoing project with Professor Alexander Panchenko focused on detecting LLM self-knowledge through intrinsic representations, with the goal of enhancing model trustworthiness.

Internal Dynamics. Building on my ongoing research and several recent papers [17, 13, 73, 4, 30], internal representations provide significant insights into trustworthiness and self-knowledge. However, typical research often overlooks the dynamics of these representations, as LLMs tend to perform incremental knowledge injection rather than entirely altering internal states. Moreover, studies frequently underestimate the value of token-wise information, which can introduce noise, as shown in a recent study [44].

Internal Guidance Generalization. Studies on internal states rarely demonstrate the generalization of these approaches across datasets and domains, often relying on a single dataset for both training and testing [37] or yielding controversial results [20]. Such linear probing methods typically exhibit limited generalization [44]. I propose to address this issue by investigating why generalization fails with diverse data, for example, by visualizing and classifying internal states across different datasets to gain insights into underlying mechanisms. Furthermore, the proposed examination of dynamics may address this limitation by focusing on relative patterns rather than specific values. Finally, like researchers from Meta and Anthropic [59, 56], I find the assumption of internal state linearity to be unfounded and advocate for an exploration of non-linear probing techniques.

Adaptive Internal Guidance. One of the drawbacks of current methods is the lack of adaptivity, as human-curated strategies, such as selecting specific layers, are typically fixed in studies to achieve better quality [4]. I propose exploring adaptive strategies for intrinsic state assessment, such as optimizing layer selection during inference by sampling multiple responses and using a meta-algorithm for mechanism selection. This approach could potentially address the generalization issue with out-of-distribution data.

Prospective Results. The expected outcome is a comprehensive study series providing insights and developing one or more methods that account for the dynamics of intrinsic representations, making them generalizable to out-of-distribution data. Additionally, investigating the reasons behind generalization failures will be a key focus, aiming to identify factors that hinder transferability. Finally, an adaptable method will be developed to adjust strategy at inference time, reducing the need for hyperparameter tuning and potentially enhancing generalization.

Proposal 3: Enhancing Explicit Self-Evaluation. Building on Proposal 1, which introduces a strong meta-judging framework, I acknowledge that such a procedure is a black-box approach and computationally expensive. Therefore, I propose distilling these abilities into a smaller, open-source model that can further refine itself through iterative self-training. This smaller model would be capable of continuous improvement by becoming a better judge through both off-policy and on-policy training settings and the addition of step-wise labeling.

> Currently, I am working on a Master's thesis project under the supervision of Chris Biemann and Irina Nikishina. This project aims to leverage a small model to evaluate the factuality of responses and iteratively revise answers based on this evaluation. The primary objective is to enhance the relevance and efficiency of the evaluator for effective revisions.

Off-Policy. A natural off-policy approach is to distill judging abilities from the meta-judge through regular SFT. This approach offers advantages for future research by eliminating the need for costly inference from the meta-judge and the model itself during training, resulting in faster and more cost-efficient training.

On-Policy. While SFT offers a more cost-effective solution, I hypothesize that it may underperform compared to other methods. Optimizing solely for Cross-Entropy may lead to overfitting and fail to address specific challenges, such as multi-turn dialogues, or the exclusion of the model's own generation, which can lead to performance collapse [24]. The advantage of using RL is that it incorporates the model's self-generations and the reward function can be shaped to account for the distribution of meta-judge evaluations or consider the number of revision steps in a multi-turn setting. To train preferences effectively, the model's own evaluations can be assessed by the meta-judge, with better evaluations receiving preference. Proper reward shaping could include applying a discount that depends on the number of multi-turn steps or reducing the KL-divergence regularization as evaluation steps increase, which has been shown to improve self-correction training [24].

Step-Wise Labeling. To improve the model's evaluation capabilities, I propose experimenting with step-wise labeling, which has demonstrated benefits in previous studies [35, 49]. Although this approach is likely to enhance evaluation abilities, this will be the first attempt to apply step-wise labeling with reasoning to tasks beyond mathematics, which have typically been used due to their easier verifiability. For training step-wise reasoning, per-step Direct Preference Optimization or its variations could be employed, such as leveraging Kahneman-Tversky Optimization [16] or Nash Equilibrium Mirror Descent Optimization [42].

Prospective Results. The primary outcome of this paper is the development of a technique for distilling self-evaluation skills into a model using knowledge from a majority of frontier closed-source models. Additionally, the results will include a comprehensive study comparing off-policy and on-policy training approaches for this task and investigation of step-wise labeling, along with ablation studies on different RL loss formulations to promote fairness.

Proposal 4: Explicit Self-Regulation. To improve the alignment, ethics, and safety of LLMs, the "Constitutional AI" approach appears to be one of the most promising solutions [2]. By using a constitution, this method ensures that the LLM follows predefined rules rather than simply maximizing rewards, as reward models are often susceptible to bias [27, 48, 50]. Even unbiased models may exploit the reward system, optimizing for reward without genuinely improving alignment [64].

However, human-written constitutions may become outdated, contain exploitable vulnerabilities, or lack comprehensive rules for regulating AI generations. To address these issues, I propose experimenting with self-refinement of the constitution, driven entirely by AI feedback. A single model or a consortium of models could engage in debate and collaboratively refine the constitution, continuously addressing vulnerabilities and updating guidelines. This approach draws inspiration from the recent IterAlign paper, which makes only a single step of constitution refinement, with its primary goal being to obtain safer responses for further tuning rather than modifying the constitution itself [7]. This highlights the potential for expanding upon this idea by enabling continuous refinement. Furthermore, this proposal builds on recent findings that frontier models demonstrate strong ethical reasoning abilities [47].

This idea shares similarities with Proposal 1 but differs significantly in its core application and the abilities being tested. While Proposal 1 emphasizes evaluation skills across diverse settings, Proposal 3 explores the capabilities and potential outcomes of Self-Regulation within the "Constitutional AI" paradigm.

Atomic Decisions. The first step is to evaluate whether models are capable of making correct atomic decisions. Before scaling the experiment, I propose constructing several settings that consist of an existing constitution (predefined by human assessors) and a prompt designed to exploit a known vulnerability. The primary goal of this experiment is to test whether models can refine the constitution to address and fix such vulnerabilities. This experiment builds on the established ability of models to identify unsafe behaviors, a problem that has been largely addressed in previous research [66].

Impact of Debate and Consortium. I propose utilizing a consortium of various frontier models in a consultancy setting, engaging in debate or debate sessions with meta-judge interventions [23]. Additionally, enforcing diversity through role-play or incorporating culturally specific LLMs could contribute to making more accurate decisions [29, 28, 46]. This experiment aims to determine whether these techniques lead to better overall decisions and more effective constitution refinements.

Cyclic Constitution Refinement. To further challenge the model in a real-world scenario, I propose using the previously identified best-performing setting to conduct multiple rounds of self-refinement. After several cycles of refinement, the model's performance under the refined constitution will be evaluated using established safety benchmarks [31]. The core objective of this experiment is to determine whether LLMs can continuously refine the constitution without performance degradation, and to assess whether repeated refinement is overall beneficial or leads to collapse.

Prospective Results. The expected outcomes of this research include the introduction of a new paradigm for Constitutional AI involving self-refinement, the creation of a dataset containing predefined constitutions and vulnerable prompts, and a series of thorough experiments to assess whether models can effectively refine their constitutions. Additionally, the research will provide a comprehensive evaluation of debate regimes and consortium performance, as well as multi-turn refinement experiments to test the model's ability for continuous improvement. Despite these advancements, many research questions remain within this paradigm that will need further exploration.

2.1.3 Summary

The section presents a research plan aimed at building the first key step for self-training of LLMs, focusing on enhancing their self-evaluation and self-regulation capabilities. This area of research is crucial because previous studies have shown that LLMs have strong potential for evaluation and guidance, providing an opportunity to create more effective, unbiased and scalable AI systems.

The proposals are structured into logical steps, each building on the prior one:

- **Proposal 1: Self-Evaluation Evaluation Framework** - Aims to create a robust framework for assessing a model's self-evaluation capabilities across various dimensions and includes the use of diverse majority voting for ground truth generation to ensure fairness.

- **Proposal 2: Intrinsic Self-Guidance** - Building on advancements from my recent papers at ACL, EMNLP, and COLING, as well as current ongoing projects, I propose to further explore intrinsic mechanisms for LLM guidance, particularly in the area of trustworthiness. This exploration would encompass the dynamics of internal states, considerations of non-linearity, out-of-distribution generalization testing, and adaptive strategies.

- **Proposal 3: Enhancing Explicit Self-Evaluation** - Builds on current Master's Thesis and the prospected evaluation framework by distilling self-evaluation capabilities into a smaller, open-source model for continuous refinement. This involves off-policy training through SFT, on-policy training with RL and experiments with step-wise labeling and reward shaping to boost evaluation effectiveness.

- **Proposal 4: Explicit Self-Regulation** - Focuses on improving alignment, ethics, and safety using a "Constitutional AI" approach with self-refinement. This involves allowing LLMs to refine their own constitutions based on feedback, conduct debates, and iteratively improve decision-making to enhance the robustness of ethical alignment.

2.2 Self-Play and Environment for LLM

Once the Self-Evaluation is properly explored, we could move further to building Self-Play for LLM and creating text environments to train LLM into, driven solely by LLM evaluators.

2.2.1 Background

Text Environment is one of the challenges for training LLM with RL as they are nearly absent. Most of previous attempts to build LLM environments focus on agency with web [68] or physical simulation [60]. The real natural language environment is multi-turn dialogue. Today the paradigm is far from real world language environment, as LLMs are trained with only single response, as well as LLM is not usually allowed to question the user with more precise directions. Some papers have proposed the LLM-powered multi-turn dialogue games as an environment [10, 1, 77], whether other propose to mitigate the problem with imagined dialogues [19], making use of LLM learned "world model", as in classic RL [18]. However, those approaches are build without proper evaluation of Self-Evaluation and only starting this research direction.

Self-Play paradigm could be seen as an environemnt play next steps, as we are making the model to play with itself within some rules, resembling to latest RL papers [51]. In recent studies, this method started gaining attention in NLP area [8], partially because it is easy to conduct with textual environment is initialized with human-written instructions and later revised by the model itself. This method has already proven to enhance truthfulness [6, 74, 9], alignment [34, 7, 36, 26], math [61, 26, 8], reasoning [8, 72, 10], classification [26]. Some papers propose a strong theoretical basement for Self-Play, also finding resembles to DPO, but with completely different assumptions and derivations [8].

2.2.2 Self-Play Research Proposals

In this section, I discuss the proposals related to applying the Self-Play method to LLMs.

Proposal 1: Simultaneous Learning. Previous efforts on self-training mechanisms have typically assumed that the judge model is pre-existing. However, it could be advantageous to simultaneously learn both the task and the judging criteria, with promising early results already demonstrated in initial attempts [65]. Future work should explore distinct judging strategies and focus on developing more complex judge learning methods to further enhance this approach.

This proposal is closely related to Proposal 2 from Self-Evaluation but differs in that it involves not only distilling self-evaluation knowledge but also learning self-play while simultaneously training the judge model. The goal is to leverage the synergy between training and evaluation to enhance both processes.

Robust Meta-Judge. Previously, the judge was learned through self-rewarding judgments, acting as a meta-judge to evaluate self-evaluation; however, there were no guarantees of effective meta-judging. Even the authors noted limitations, such as the judge's bias and collapsing performance towards the end of training. These issues could potentially be mitigated by employing a more sophisticated judge, as proposed in Proposal 2 from Self-Evaluation. Furthermore, optimizing both judge learning and self-play simultaneously may lead to a focus on "hacking" the judge rather than genuine evaluation. A possible solution is to adjust the optimization process, for example, by distinguishing between these losses, similar to the approach used in the Actor-Critic framework [38].

Adversarial Setup. As proposed in the SPIN framework [8], LLMs can be trained using a Generative-Adversarial setup, where the Critic attempts to distinguish between human-generated and model-generated answers, while the Generator aims to make its responses indistinguishable from human ones. I propose employing a similar paradigm for learning judgments, but instead of relying on human-generated responses, using a more sophisticated meta-judge to create ground truth responses.

Prospective Results. The expected outcome of this research is to develop a stronger self-training framework by simultaneously training the judge. To achieve this, I propose experimenting with the strength of the Meta-Judge, enhancing the loss function to prioritize desired behaviors, and using the adversarial approach to promote further scaling. This study aims to expand our understanding of whether simultaneous judge training is more effective than training the judge separately.

Proposal 2: Generated Self-Play Data. So far, Self-Play has only been used to modify the initial SFT dataset, without creating entirely new data that deviates from the original distribution. Methods such as SPIN, GRATH, and SK-Tuning sample answers from a fixed dataset [8, 6, 74], while LLM2LLM merely augments SFT data by rephrasing it [26]. I propose experimenting with the generation of new, unseen data that is not seeded with initial prompts.

Currently, I am leading a project focused on generating factual and counterfactual data for iterative self-play fine-tuning. This aims to enhance LLM verbalized confidence through DPO alignment with factuality, building on recent studies [6, 55].

Self-Evaluation for Assessment. The main challenge with this training strategy is the absence of ground truth answers. To address this, I propose leveraging Self-Evaluation as a foundational approach, which was the main goal of first step in my research proposal. Self-Evaluation can be utilized in several ways: it can select the best answer from multiple sampled responses to a generated question for subsequent preference or SFT tuning, or it can provide a raw score that can later be used for reward-based tuning.

Special Case of Truthfulness. Previous studies on self-play for truthfulness have employed a basic approach, using only existing questions and very basic approaches. This could be significantly enhanced by incorporating rephrased questions, compositional questions, and more. Additionally, generating plausible yet untrue responses that are challenging to distinguish from true ones remains an unexplored area. Furthermore, no research to date has employed a neural evaluator for self-play, which could elevate the approach by leveraging more synthetic data and enabling richer evaluation.

Prevent Hacking. A potential issue is the risk of hacking the self-evaluation model. To address this, I propose incorporating fine-tuning of the self-evaluation skills, enhancing robustness through agency (e.g., employing debates or using multiple models), and experimenting with the loss function to mitigate the risk of hacking.

Prospective Results. This study aims to develop a more robust and data-rich technique for self-play, which is crucial given the biases and limitations of seed data. Additionally, this research will investigate the model's ability to exploit its self-evaluation mechanisms and explore methods to prevent such vulnerabilities.

Proposal 3: Challenging Textual Environments for LLM Training. Current LLM training environments often lack complexity and diversity, limiting their effectiveness in cultivating nuanced reasoning, planning, and interaction abilities [1, 10]. To address this, I propose developing more challenging text-based game environments that can serve as a rich training ground for LLMs. These environments will cover a range of tasks that require deep reasoning, adaptability, and creativity, extending beyond standard datasets.

Spy Game. The Spy Game environment features multiple agents, adding complexity to both reasoning and interaction. All agents, except the learning agent, are assigned a specific location (e.g., a city, a room). However, they are not allowed to explicitly reveal this place. Each agent takes turns asking the next agent a closed-ended question about the location that could be ambiguous, such as asking, ”Is it cold there?” when the location is ”Alaska.” After each response, all agents collectively decide if the answer seems suspicious, potentially indicating that the agent is a spy unaware of the location. If the answer aligns well with the assigned place, suspicion decreases, and vice versa. The goal for the spy agent is to deduce the place quickly based on the questions and answers, while avoiding detection by the other agents. This environment forces the learning agent to develop advanced reasoning skills, deal with uncertainty, and navigate multi-agent dynamics.

Perudo. Perudo is a bluffing game where each agent attempts to predict the total count of dice showing a specific face value across all participants. Each agent starts with five virtual dice, with the outcomes generated randomly, and each agent can only view its own results. The objective is to estimate how many dice collectively show a certain face value. The first agent initiates with a bid, predicting the minimum quantity of dice showing a particular face (e.g., “five 3's”). Following this, agents can either raise the bid (increasing the quantity or face value) or challenge it. When a bid is challenged, all dice results are revealed. If the bid was accurate or underestimated, the challenger loses a die; if overstated, the bidder loses a die. The game continues until only one agent retains dice, making them the winner. Unlike the Spy Game, Perudo enables bluffing, introducing greater uncertainty. Additionally, this setting can serve as a test environment to assess an agent's ability to adapt to varied behaviors, such as facing agents who consistently bluff or never bluff at all.

Detective Scenarios. One environment involves solving detective stories, such as murder mysteries or town conspiracies, which can be sampled from LLMs or rewritten based on existing fictional narratives. These scenarios will require the model to gather clues, deduce relationships, and propose hypotheses, thereby fostering logical reasoning and critical thinking skills.

Guess the Person. Another proposed environment is a "guess the person" game, constructed using LLMs augmented with a graph of factual knowledge. This game would require the model to ask questions to identify a target individual, combining structured factual knowledge with an interactive dialogue to test the model's deductive capabilities.

Persuasion and Debates. A persuasive environment will focus on "telling a secret" or persuading an opponent of a specific viewpoint. Debates, which can be moderated by a meta-judge, will aim to test the LLM's ability to argue effectively and win debates, helping to improve its capabilities in logical reasoning, ethical considerations, and adaptability.

Planning and Execution. The planning environment will involve manipulating items, such as moving blocks in a simulated scenario. The goal is to test the model's ability to plan multi-step actions and adapt its strategy based on feedback, thereby enhancing its understanding of sequential decision-making.

Prospective Results. The anticipated outcome of this research is the development of a diverse set of complex text-based environments for LLM training, which will enable models to acquire more advanced reasoning, planning, persuasion, and problem-solving skills. These environments will also foster the model's ability to adapt to challenging scenarios and improve its robustness in generating coherent and accurate responses. Additionally, this proposal will help identify potential weaknesses in LLM capabilities across different types of tasks and suggest directions for further enhancement.

2.2.3 Summary

The section presents a research plan aimed at advancing self-training techniques for LLMs by focusing on self-evaluation, self-play, and creating complex training environments. This research is crucial as it leverages the potential of LLMs for learning through self-interaction and aims to establish more sophisticated training setups, ultimately leading to more adaptable and effective AI systems.

The proposals are structured into logical steps, each building on the previous one:

- **Proposal 1: Simultaneous Learning** - Proposes simultaneous learning of tasks and judging criteria, rather than relying on a pre-existing judge. This aims to leverage the synergy between task learning and evaluation to enhance both processes effectively.
- **Proposal 2: Generated Self-Play Data** - Introduces the generation of entirely new data for self-play that diverges from the original distribution, unlike previous methods that used existing datasets. A promising sub-branch is the application of this method to trustworthiness, which presents numerous intricacies and requires careful attention. This proposal also aimed at addressing the challenge of lacking ground truth through self-evaluation and aims to explore how to prevent model hacking via enhanced self-evaluation robustness.
- **Proposal 3: Challenging Textual Environments for LLM Training** - Suggests creating complex text-based game environments, such as detective scenarios, debates, and planning exercises, to challenge LLMs and foster advanced reasoning, adaptability, and problem-solving abilities. These environments might help LLMs acquire richer skills that are more aligned with real-world challenges.

3 Resources

To complete the aforementioned research, several key components are required:

- **Compute Power.** The trend towards smaller, compute-efficient models enables such experiments without excessive resources, requiring at least a 40-80GB GPU with Ampere architecture (e.g., A100, H100). Ideally, additional compute would allow for faster and more scalable experimentation. In the absence of local resources, cloud computing or collaboration with universities equipped with compute power (e.g., Skoltech) could provide alternatives.

- **Access to Frontier Models.** Recent trends indicate the necessity of access to advanced models like OpenAI's GPT, Claude, etc., to facilitate evaluations, knowledge distillation, and related tasks.
- **Supervision.** Ideally, at least one experienced supervisor should be available to guide the research.

4 Deliverables

Each research branch aims to deliver at least one A* paper and typically includes additional artifacts, such as datasets, models, or easily integrable methods. The deliverables are as follows:

- A benchmark and methodology for assessing LLM self-evaluation capabilities.
- A method and model for efficient, low-cost evaluation of LLMs, including fact-checking.
- In-depth research on models' ability to regulate behavior concerning factuality and ethical principles.
- Methods for applying self-play to trustworthiness challenges.
- A technique for enhancing self-play with synthetic data.
- A benchmark and environments utilizing text-based agent simulations.

References

[1] M. Abdulhai, I. White, C. Snell, C. Sun, J. Hong, Y. Zhai, K. Xu and S. Levine, LMRL Gym: Benchmarks for Multi-Turn Reinforcement Learning with Language Models, 2023. https://arxiv.org/abs/2311.18232.

[2] Y. Bai, S. Kadavath, S. Kundu, A. Askell, J. Kernion, A. Jones, A. Chen, A. Goldie, A. Mirhoseini, C. McKinnon, C. Chen, C. Olsson, C. Olah, D. Hernandez, D. Drain, D. Ganguli, D. Li, E. Tran-Johnson, E. Perez, J. Kerr, J. Mueller, J. Ladish, J. Landau, K. Ndousse, K. Lukosuite, L. Lovitt, M. Sellitto, N. Elhage, N. Schiefer, N. Mercado, N. DasSarma, R. Lasenby, R. Larson, S. Ringer, S. Johnston, S. Kravec, S.E. Showk, S. Fort, T. Lanham, T. Telleen-Lawton, T. Conerly, T. Henighan, T. Hume, S.R. Bowman, Z. Hatfield-Dodds, B. Mann, D. Amodei, N. Joseph, S. McCandlish, T. Brown and J. Kaplan, Constitutional AI: Harmlessness from AI Feedback, 2022. https://arxiv.org/abs/2212.08073.

[3] Y. Bai, J. Ying, Y. Cao, X. Lv, Y. He, X. Wang, J. Yu, K. Zeng, Y. Xiao, H. Lyu, J. Zhang, J. Li and L. Hou, Benchmarking Foundation Models with Language-Model-as-an-Examiner, in: *Advances in Neural Information Processing Systems*, Vol. 36, A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt and S. Levine, eds, Curran Associates, Inc., 2023, pp. 78142–78167. https://proceedings.neurips.cc/paper$_{f}iles/paper/2023/file/f64e55d03e2fe61aa4114e49cb654acb - Paper - Datasets_{a}nd_{B}enchmarks.pdf$.

[4] C. Burns, H. Ye, D. Klein and J. Steinhardt, Discovering latent knowledge in language models without supervision, *arXiv preprint arXiv:2212.03827* (2022).

[5] C. Burns, P. Izmailov, J.H. Kirchner, B. Baker, L. Gao, L. Aschenbrenner, Y. Chen, A. Ecoffet, M. Joglekar, J. Leike, I. Sutskever and J. Wu, Weak-to-Strong Generalization: Eliciting Strong Capabilities With Weak Supervision, 2023. https://arxiv.org/abs/2312.09390.

[6] W. Chen and B. Li, GRATH: Gradual Self-Truthifying for Large Language Models, *arXiv preprint arXiv:2401.12292* (2024).

[7] X. Chen, H. Wen, S. Nag, C. Luo, Q. Yin, R. Li, Z. Li and W. Wang, IterAlign: Iterative Constitutional Alignment of Large Language Models, 2024. https://arxiv.org/abs/2403.18341.

[8] Z. Chen, Y. Deng, H. Yuan, K. Ji and Q. Gu, Self-Play Fine-Tuning Converts Weak Language Models to Strong Language Models, 2024. https://arxiv.org/abs/2401.01335.

[9] Z. Chen, Y. Deng, H. Yuan, K. Ji and Q. Gu, Self-play fine-tuning converts weak language models to strong language models, *arXiv preprint arXiv:2401.01335* (2024).

[10] P. Cheng, T. Hu, H. Xu, Z. Zhang, Y. Dai, L. Han and N. Du, Self-playing Adversarial Language Game Enhances LLM Reasoning, *arXiv preprint arXiv:2404.10642* (2024).

[11] C.-H. Chiang and H.-y. Lee, Can large language models be an alternative to human evaluations?, *arXiv preprint arXiv:2305.01937* (2023).

[12] W.-L. Chiang, L. Zheng, Y. Sheng, A.N. Angelopoulos, T. Li, D. Li, H. Zhang, B. Zhu, M. Jordan, J.E. Gonzalez and I. Stoica, Chatbot Arena: An Open Platform for Evaluating LLMs by Human Preference, 2024. https://arxiv.org/abs/2403.04132.

[13] Y.-S. Chuang, L. Qiu, C.-Y. Hsieh, R. Krishna, Y. Kim and J. Glass, Lookback Lens: Detecting and Mitigating Contextual Hallucinations in Large Language Models Using Only Attention Maps, 2024. https://arxiv.org/abs/2407.07071.

[14] X. Cui, Q. Sun, W. Zhou and H. Li, Exploring GPT-4 Vision for Text-to-Image Synthesis Evaluation, in: *The Second Tiny Papers Track at ICLR 2024*, 2024. https://openreview.net/forum?id=xmQoodG82a.

[15] S. Dréano, D. Molloy and N. Murphy, Embed_Llama: using LLM embeddings for the Metrics Shared Task, in: *Proceedings of the Eighth Conference on Machine Translation*, 2023, pp. 738–745.

[16] K. Ethayarajh, W. Xu, N. Muennighoff, D. Jurafsky and D. Kiela, Kto: Model alignment as prospect theoretic optimization, *arXiv preprint arXiv:2402.01306* (2024).

[17] D. Gottesman and M. Geva, Estimating Knowledge in Large Language Models Without Generating a Single Token, 2024. https://arxiv.org/abs/2406.12673.

[18] D. Hafner, T. Lillicrap, J. Ba and M. Norouzi, Dream to control: Learning behaviors by latent imagination, *arXiv preprint arXiv:1912.01603* (2019).

[19] J. Hong, S. Levine and A. Dragan, Zero-shot goal-directed dialogue via rl on imagined conversations, *arXiv preprint arXiv:2311.05584* (2023).

[20] S. Kadavath, T. Conerly, A. Askell, T. Henighan, D. Drain, E. Perez, N. Schiefer, Z. Hatfield-Dodds, N. DasSarma, E. Tran-Johnson, S. Johnston, S. El-Showk, A. Jones, N. Elhage, T. Hume, A. Chen, Y. Bai, S. Bowman, S. Fort, D. Ganguli, D. Hernandez, J. Jacobson, J. Kernion, S. Kravec, L. Lovitt, K. Ndousse, C. Olsson, S. Ringer, D. Amodei, T. Brown, J. Clark, N. Joseph, B. Mann, S. McCandlish, C. Olah and J. Kaplan, Language Models (Mostly) Know What They Know, 2022. https://arxiv.org/abs/2207.05221.

[21] A. Karpathy, Tweet about AI and scaling, 2023, Accessed: 2024-09-01.

[22] Z. Kenton, N.Y. Siegel, J. Kramár, J. Brown-Cohen, S. Albanie, J. Bulian, R. Agarwal, D. Lindner, Y. Tang, N.D. Goodman and R. Shah, On scalable oversight with weak LLMs judging strong LLMs, 2024. https://arxiv.org/abs/2407.04622.

[23] A. Khan, J. Hughes, D. Valentine, L. Ruis, K. Sachan, A. Radhakrishnan, E. Grefenstette, S.R. Bowman, T. Rocktäschel and E. Perez, Debating with More Persuasive LLMs Leads to More Truthful Answers, 2024. https://arxiv.org/abs/2402.06782.

[24] A. Kumar, V. Zhuang, R. Agarwal, Y. Su, J.D. Co-Reyes, A. Singh, K. Baumli, S. Iqbal, C. Bishop, R. Roelofs, L.M. Zhang, K. McKinney, D. Shrivastava, C. Paduraru, G. Tucker, D. Precup, F. Behbahani and A. Faust, Training Language Models to Self-Correct via Reinforcement Learning, 2024. https://arxiv.org/abs/2409.12917.

[25] H. Lee, S. Phatale, H. Mansoor, T. Mesnard, J. Ferret, K. Lu, C. Bishop, E. Hall, V. Carbune, A. Rastogi and S. Prakash, RLAIF: Scaling Reinforcement Learning from Human Feedback with AI Feedback, 2023. https://arxiv.org/abs/2309.00267.

[26] N. Lee, T. Wattanawong, S. Kim, K. Mangalam, S. Shen, G. Anumanchipalli, M.W. Mahoney, K. Keutzer and A. Gholami, LLM2LLM: Boosting LLMs with Novel Iterative Data Enhancement, 2024. https://arxiv.org/abs/2403.15042.

[27] E.A. Lerner, F.E. Dorner, E. Ash and N. Goel, Whose Preferences? Differences in Fairness Preferences and Their Impact on the Fairness of AI Utilizing Human Feedback, *arXiv preprint arXiv:2406.05902* (2024).

[28] C. Li, M. Chen, J. Wang, S. Sitaram and X. Xie, CultureLLM: Incorporating Cultural Differences into Large Language Models, 2024. https://arxiv.org/abs/2402.10946.

[29] C. Li, D. Teney, L. Yang, Q. Wen, X. Xie and J. Wang, CulturePark: Boosting Cross-cultural Understanding in Large Language Models, *arXiv preprint arXiv:2405.15145* (2024).

[30] K. Li, O. Patel, F. Viégas, H. Pfister and M. Wattenberg, Inference-Time Intervention: Eliciting Truthful Answers from a Language Model, in: *Advances in Neural Information Processing Systems*, Vol. 36, A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt and S. Levine, eds, Curran Associates, Inc., 2023, pp. 41451–41530. https://proceedings.neurips.cc/paper$_{f}iles/paper/2023/file/81b8390039b7302c909cb769f8b6cd93 - Paper - Conference.pdf$.

[31] L. Li, B. Dong, R. Wang, X. Hu, W. Zuo, D. Lin, Y. Qiao and J. Shao, SALAD-Bench: A Hierarchical and Comprehensive Safety Benchmark for Large Language Models, 2024. https://arxiv.org/abs/2402.05044.

[32] T. Li, W.-L. Chiang, E. Frick, L. Dunlap, T. Wu, B. Zhu, J.E. Gonzalez and I. Stoica, From Crowdsourced Data to High-Quality Benchmarks: Arena-Hard and BenchBuilder Pipeline, 2024. https://arxiv.org/abs/2406.11939.

[33] Y. Li, Z. Lin, S. Zhang, Q. Fu, B. Chen, J.-G. Lou and W. Chen, Making Language Models Better Reasoners with Step-Aware Verifier, in: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, A. Rogers, J. Boyd-Graber and N. Okazaki, eds, Association for Computational Linguistics, Toronto, Canada, 2023, pp. 5315–5333. doi:10.18653/v1/2023.acl-long.291. https://aclanthology.org/2023.acl-long.291.

[34] Y. Liang, G. Zhang, X. Qu, T. Zheng, J. Guo, X. Du, Z. Yang, J. Liu, C. Lin, L. Ma et al., I-SHEEP: Self-Alignment of LLM from Scratch through an Iterative Self-Enhancement Paradigm, *arXiv preprint arXiv:2408.08072* (2024).

[35] H. Lightman, V. Kosaraju, Y. Burda, H. Edwards, B. Baker, T. Lee, J. Leike, J. Schulman, I. Sutskever and K. Cobbe, Let's verify step by step, *arXiv preprint arXiv:2305.20050* (2023).

[36] J. Lu, W. Zhong, W. Huang, Y. Wang, Q. Zhu, F. Mi, B. Wang, W. Wang, X. Zeng, L. Shang, X. Jiang and Q. Liu, SELF: Self-Evolution with Language Feedback, 2024. https://arxiv.org/abs/2310.00533.

[37] S. Marks and M. Tegmark, The geometry of truth: Emergent linear structure in large language model representations of true/false datasets, *arXiv preprint arXiv:2310.06824* (2023).

[38] V. Mnih, Asynchronous Methods for Deep Reinforcement Learning, *arXiv preprint arXiv:1602.01783* (2016).

[39] V. Moskvoretskii, A. Panchenko and I. Nikishina, Are Large Language Models Good at Lexical Semantics? A Case of Taxonomy Learning, in: *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti and N. Xue, eds, ELRA and ICCL, Torino, Italia, 2024, pp. 1498–1510. https://aclanthology.org/2024.lrec-main.133.

[40] V. Moskvoretskii, N. Tupitsa, C. Biemann, S. Horváth, E. Gorbunov and I. Nikishina, Low-Resource Machine Translation through the Lens of Personalized Federated Learning, 2024. https://arxiv.org/abs/2406.12564.

[41] V. Moskvoretskii, E. Neminova, A. Lobanova, A. Panchenko and I. Nikishina, TaxoLLaMA: WordNet-based Model for Solving Multiple Lexical Semantic Tasks, in: *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, L.-W. Ku, A. Martins and V. Srikumar, eds, Association for Computational Linguistics, Bangkok, Thailand, 2024, pp. 2331–2350. doi:10.18653/v1/2024.acl-long.127. https://aclanthology.org/2024.acl-long.127.

[42] R. Munos, M. Valko, D. Calandriello, M.G. Azar, M. Rowland, Z.D. Guo, Y. Tang, M. Geist, T. Mesnard, A. Michi et al., Nash learning from human feedback, *arXiv preprint arXiv:2312.00886* (2023).

[43] OpenAI, OpenAI o1-mini: Advancing cost-efficient reasoning, 2024, Accessed: 2024-09-26. https://openai.com/index/openai-o1-mini-advancing-cost-efficient-reasoning/.

[44] H. Orgad, M. Toker, Z. Gekhman, R. Reichart, I. Szpektor, H. Kotek and Y. Belinkov, LLMs Know More Than They Show: On the Intrinsic Representation of LLM Hallucinations, *arXiv preprint arXiv:2410.02707* (2024).

[45] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C.L. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, J. Schulman, J. Hilton, F. Kelton, L. Miller, M. Simens, A. Askell, P. Welinder, P. Christiano, J. Leike and R. Lowe, Training language models to follow instructions with human feedback, 2022. https://arxiv.org/abs/2203.02155.

[46] X. Pang, S. Tang, R. Ye, Y. Xiong, B. Zhang, Y. Wang and S. Chen, Self-alignment of large language models via monopolylogue-based social scene simulation, *arXiv preprint arXiv:2402.05699* (2024).

[47] A. Rao, A. Khandelwal, K. Tanmay, U. Agarwal and M. Choudhury, Ethical reasoning over moral alignment: A case and framework for in-context ethical policies in LLMs, *arXiv preprint arXiv:2310.07251* (2023).

[48] M.J. Ryan, W. Held and D. Yang, Unintended impacts of llm alignment on global representation, *arXiv preprint arXiv:2402.15018* (2024).

[49] A. Setlur, S. Garg, X. Geng, N. Garg, V. Smith and A. Kumar, RL on Incorrect Synthetic Data Scales the Efficiency of LLM Math Reasoning by Eight-Fold, *arXiv preprint arXiv:2406.14532* (2024).

[50] M. Sharma, M. Tong, T. Korbak, D. Duvenaud, A. Askell, S.R. Bowman, N. Cheng, E. Durmus, Z. Hatfield-Dodds, S.R. Johnston et al., Towards understanding sycophancy in language models, *arXiv preprint arXiv:2310.13548* (2023).

[51] D. Silver, J. Schrittwieser, K. Simonyan, I. Antonoglou, A. Huang, A. Guez, T. Hubert, L. Baker, M. Lai, A. Bolton et al., Mastering the game of go without human knowledge, *nature* **550**(7676) (2017), 354–359.

[52] C. Snell, J. Lee, K. Xu and A. Kumar, Scaling llm test-time compute optimally can be more effective than scaling model parameters, *arXiv preprint arXiv:2408.03314* (2024).

[53] R. Sutton, The bitter lesson, *Incomplete Ideas (blog)* **13**(1) (2019), 38.

[54] H. Tan, F. Sun, W. Yang, Y. Wang, Q. Cao and X. Cheng, Blinded by Generated Contexts: How Language Models Merge Generated and Retrieved Contexts When Knowledge Conflicts?, in: *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2024, pp. 6207–6227.

[55] S. Tao, L. Yao, H. Ding, Y. Xie, Q. Cao, F. Sun, J. Gao, H. Shen and B. Ding, When to Trust LLMs: Aligning Confidence with Response Quality, 2024. https://arxiv.org/abs/2404.17287.

[56] A. Templeton, T. Conerly, J. Marcus, J. Lindsey, T. Bricken, B. Chen, A. Pearce, C. Citro, E. Ameisen, A. Jones, H. Cunningham, N.L. Turner, C. McDougall, M. MacDiarmid, C.D. Freeman, T.R. Sumers, E. Rees, J. Batson, A. Jermyn, S. Carter, C. Olah and T. Henighan, Scaling Monosemanticity: Extracting Interpretable Features from Claude 3 Sonnet, *Transformer Circuits Thread* (2024). https://transformer-circuits.pub/2024/scaling-monosemanticity/index.html.

[57] P. Villalobos, A. Ho, J. Sevilla, T. Besiroglu, L. Heim and M. Hobbhahn, Can AI Scaling Continue Through 2030?, *Epoch AI Blog* (2024), Accessed: 2024-09-01. https://epochai.org/blog/can-ai-scaling-continue-through-2030synthetic-data.

[58] P. Villalobos, A. Ho, J. Sevilla, T. Besiroglu, L. Heim and M. Hobbhahn, Will We Run Out of Data? Limits of LLM Scaling Based on Human-Generated Data, *Epoch AI Blog* (2024), Accessed: 2024-09-01. https://epochai.org/blog/will-we-run-out-of-data-limits-of-llm-scaling-based-on-human-generated-data.

[59] E. Voita, J. Ferrando and C. Nalmpantis, Neurons in large language models: Dead, n-gram, positional, *arXiv preprint arXiv:2309.04827* (2023).

[60] R. Wang, P. Jansen, M.-A. Côté and P. Ammanabrolu, Scienceworld: Is your agent smarter than a 5th grader?, *arXiv preprint arXiv:2203.07540* (2022).

[61] T. Wang, S. Li and W. Lu, Self-Training with Direct Preference Optimization Improves Chain-of-Thought Reasoning, *arXiv preprint arXiv:2407.18248* (2024).

[62] X. Wang, Y. Fei, Z. Leng and C. Li, Does role-playing chatbots capture the character personalities? assessing personality traits for role-playing chatbots, *arXiv preprint arXiv:2310.17976* (2023).

[63] X. Wang, J. Wei, D. Schuurmans, Q. Le, E. Chi, S. Narang, A. Chowdhery and D. Zhou, Self-Consistency Improves Chain of Thought Reasoning in Language Models, 2023. https://arxiv.org/abs/2203.11171.

[64] J. Wen, R. Zhong, A. Khan, E. Perez, J. Steinhardt, M. Huang, S.R. Bowman, H. He and S. Feng, Language Models Learn to Mislead Humans via RLHF, 2024. https://arxiv.org/abs/2409.12822.

[65] T. Wu, W. Yuan, O. Golovneva, J. Xu, Y. Tian, J. Jiao, J. Weston and S. Sukhbaatar, Meta-rewarding language models: Self-improving alignment with llm-as-a-meta-judge, *arXiv preprint arXiv:2407.19594* (2024).

[66] Y. Xie, M. Fang, R. Pi and N. Gong, GradSafe: Detecting Jailbreak Prompts for LLMs via Safety-Critical Gradient Analysis, in: *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2024, pp. 507–518.

[67] W. Xu, G. Zhu, X. Zhao, L. Pan, L. Li and W.Y. Wang, Pride and Prejudice: LLM Amplifies Self-Bias in Self-Refinement, 2024. https://arxiv.org/abs/2402.11436.

[68] S. Yao, H. Chen, J. Yang and K. Narasimhan, Webshop: Towards scalable real-world web interaction with grounded language agents, *Advances in Neural Information Processing Systems* **35** (2022), 20744–20757.

[69] H. Ye and H.T. Ng, Preference-Guided Reflective Sampling for Aligning Language Models, *arXiv preprint arXiv:2408.12163* (2024).

[70] T. Ye, Z. Xu, Y. Li and Z. Allen-Zhu, Physics of Language Models: Part 2.2, How to Learn From Mistakes on Grade-School Math Problems, 2024. https://arxiv.org/abs/2408.16293.

[71] W. Yuan, R.Y. Pang, K. Cho, S. Sukhbaatar, J. Xu and J. Weston, Self-rewarding language models, *arXiv preprint arXiv:2401.10020* (2024).

[72] E. Zelikman, Y. Wu, J. Mu and N. Goodman, Star: Bootstrapping reasoning with reasoning, *Advances in Neural Information Processing Systems* **35** (2022), 15476–15488.

[73] S. Zhang, T. Yu and Y. Feng, TruthX: Alleviating Hallucinations by Editing Large Language Models in Truthful Space, 2024. https://arxiv.org/abs/2402.17811.

[74] X. Zhang, B. Peng, Y. Tian, J. Zhou, L. Jin, L. Song, H. Mi and H. Meng, Self-alignment for factuality: Mitigating hallucinations in llms via self-evaluation, *arXiv preprint arXiv:2402.09267* (2024).

[75] M. Zhelnin, V. Moskvoretskii, E. Shvetsov, E. Venediktov, M. Krylova, A. Zuev and E. Burnaev, GIFT-SW: Gaussian noise Injected Fine-Tuning of Salient Weights for LLMs, 2024. https://arxiv.org/abs/2408.15300.

[76] L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. Xing, H. Zhang, J.E. Gonzalez and I. Stoica, Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena, in: *Advances in Neural Information Processing Systems*, Vol. 36, A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt and S. Levine, eds, Curran Associates, Inc., 2023, pp. 46595–46623. https://proceedings.neurips.cc/paper$_{f}iles/paper/2023/file/91f18a1287b398d378ef22505bf41832-Paper-Datasets_{a}nd_{B}enchmarks.pdf$.

[77] Y. Zhou, A. Zanette, J. Pan, S. Levine and A. Kumar, Archer: Training language model agents via hierarchical multi-turn rl, *arXiv preprint arXiv:2402.19446* (2024).

[78] B. Zhu, E. Frick, T. Wu, H. Zhu and J. Jiao, Starling-7B: Improving LLM Helpfulness Harmlessness with RLAIF, 2023.