
Self-reflecting Large Language Models: A Hegelian Dialectical Approach

Sara Abdali¹ Can Goksen*¹ Michael Solodko*¹ Saeed Amizadeh¹ Julie E. Maybee² Kazuhito Koishida¹

Abstract

Investigating NLP through a philosophical lens has recently caught researchers’ eyes, as it bridges computational methods with classical schools of philosophy. This paper introduces a philosophical framework inspired by the *Hegelian Dialectic* to enable LLMs’ *self-reflection*, utilizing a self-dialectical approach to emulate internal critiques and synthesize new scientific ideas (spanning domains such as mathematics, physics, and more). Additionally, we explore the effect of generation temperature in LLMs by introducing a dynamic annealing approach, which encourages creativity in the early stages and gradually focuses on refinement and nuance, as well as a constant-temperature strategy. Furthermore, we implement a Multi-Agent Majority Voting (MAMV) strategy to assess the validity and novelty of the generated ideas, which proves useful in the absence of domain experts. We also evaluate the effectiveness of our method in generating novel scientific ideas and improving LLMs’ reasoning capabilities. Our experiments demonstrate promising results in ideation, along with significant improvements in mathematical and symbolic reasoning.

1. Introduction

The evolution of generative AI and foundational models, particularly the revolution in Natural Language Processing (NLP) driven by the advent of Large Language Models (LLMs), has unlocked new opportunities and made significant strides toward achieving human-level reasoning, innovations and scientific discoveries (Zhang et al., 2024c; Wu et al., 2023; Zhang et al., 2024b; Smith & Doe, 2023).

Nowadays, users of LLMs frequently employ strategies like In-Context Learning (ICL), one of the simplest and most efficient paradigms in Natural Language Understanding (NLU). This method involves guiding a pre-trained model using instructions or demonstrations (e.g., providing examples to

tackle a new task without additional training or fine-tuning), thereby harnessing the model’s inherent capabilities like zero-shot and few-shot reasoning (Dong et al., 2024).

Despite the numerous successes and advantages of LLM reasoning, ensuring factual accuracy during reasoning remains a significant challenge (Abdali et al., 2024b;c). These models often exhibit issues such as neglecting certain conditions, misinterpreting contextual information, and even fabricating content in response to specific queries (Shayegani et al., 2023; Millière & Buckner, 2023; Abdali et al., 2024a).

To address issues above, researchers have proposed various strategies, including fine-tuning methods (Lewkowycz et al., 2022; Rajani et al., 2019; Zelikman et al., 2022), pruning truthful datasets (Christiano et al., 2023), adjusting system parameters to limit creativity (Muneeswaran et al., 2023), external knowledge retrieval (Guu et al., 2020), and Reinforcement Learning from Human Feedback (RLHF) (Ziegler et al., 2019; Christiano et al., 2017). Leveraging LLMs’ ability for in-context learning (Zhang et al., 2023), training-free methods with likelihood estimation (Kadavath et al., 2022), and prompt engineering methods such as Chain-of-Thought (CoT) (Wei et al., 2022), Reversing Chain-of-Thought (RCoT) (Xue et al., 2023), verification (Xue et al., 2023), scratchpads (Cobbe et al., 2021; Nye et al., 2022), and self-consistency (Wang et al., 2023) gained in popularity in recent years. Similarly, multi-agent debate (MAD) frameworks such as society of minds (Du et al., 2023), Multi-Persona (Liang et al., 2024) and ChatEva (Chan et al., 2023) have presented considerable evidence for the utility of debate in between agents in Q&A settings (Smit et al., 2024).

Iterative self-reflection (Shinn et al., 2023; Madaan et al., 2023) is another approach that has recently gained significant attention within the NLP community. This method involves models mimicking human behavior by reviewing and critiquing their own outputs, actions, or decision-making processes to improve their performance over time. While an LLM does not “reflect” in the way humans do, it can be programmed to evaluate the quality or appropriateness of its responses in certain contexts.

Self-reflection can be viewed as a form of “Self-dialectic.” Broadly speaking, “Dialectic” refers to any logical debate that involves considering opposing views to presented propositions and using opposition to uncover the truth and validity

*Equal contribution ¹Microsoft Applied Sciences Group (ASG)

²Department of Philosophy, Lehman College, City University of New York

. Correspondence to: Sara Abdali <saraabdali@microsoft.com>.

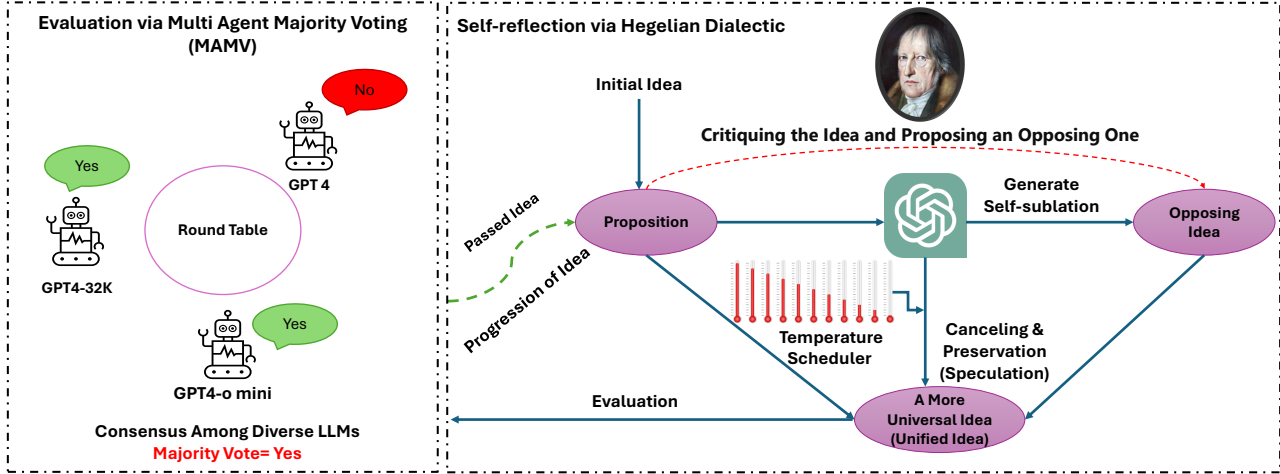


Figure 1. We introduce a self-reflection method inspired by the Hegel’s dialectic to generate novel scientific ideas and to enhance reflection on mathematical reasoning tasks. We examine with two temperature configurations: a dynamic annealing process that encourages high creativity initially and gradually decreases as the model converges on its final idea, and a constant temperature setting to ensure consistent behavior. We utilize a Multi Agent Majority Voting (MAMV) paradigm to evaluate the novelty and validity of the generated idea.

of statements made in the debate (Cambridge University Press, n.). Typically used in philosophy, the meaning of “Dialectic” beyond this general description varies depending on the philosophical tradition employing it (Bobzien & Duncombe, 2023; Maybee, 2020). The *Hegelian Dialectic*, for example, refers to the method proposed by Hegel in the 19th century, which iteratively synthesizes new opposing concepts or points of view, driving the progression of ideas in the discussion (Hegel, 1807; 1951; Maybee, 2020).

Examining NLP from a philosophical perspective has recently fascinated researchers, as it connects computational methods with traditional philosophical methodologies (Millière & Buckner, 2023; Milliere & Buckner, 2024). In this work, we propose a philosophy-inspired approach based on Hegelian dialectics to encourage self-reflection in LLMs, aiming to assess whether the model can generate novel ideas, recognize its own errors, and correct them during problem solving. Specifically, we introduce a self-debating methodology to evoke novel answers from an LLM through a “self-reflection” process, adhering to an organic dialectical discussion in the Hegelian style. Furthermore, we explore the effect of LLMs’ temperature by proposing two experimental settings, a dynamic generative approach that formalizes the dynamic creativity of an LLM via an annealing process (Delahaye et al., 2019), and a constant temperature configuration. Additionally, we utilize a Multi-Agent Majority Voting (MAMV) strategy (Minsky, 1988; Zhuge et al., 2023; Amirizani et al., 2024) to evaluate the validity and novelty of the generated ideas. This is particularly valuable when human domain experts are unavailable, as it utilizes the prior knowledge of LLMs. An overview of our proposed method is demonstrated in Figure 1. In

summary, our contributions are as follows:

- We propose a novel self-reflection method inspired by the Hegel’s dialectic that enables LLMs to generate new scientific ideas, identify their own mistakes, and fix them during problem solving.
- By establishing two configurations, we explore the effect of temperature on our dialectical self-reflection. We develop a dynamic approach that integrates an annealing process into our Hegelian self-reflection method, promoting high creativity levels initially and gradually reducing as the model reaches the final responses, as well as a constant temperature setting that maintains consistent behavior throughout the process.
- In the idea generation experiment, we employ an MAMV framework to assess the validity and novelty of the ideas generated in the absence of human experts.

2. Related Work

Self-Reflection in LLMs In the context of LLMs, self-reflection involves evaluating and refining model’s outputs through iterative cycles of feedback and adjustment (Chen et al., 2024; Li et al., 2024). Inspired by human cognitive processes, this approach aims to enhance the model’s ability to synthesize complex and sometimes conflicting information, leading to more coherent outputs. For instance, Li et al. (Li et al., 2024) explore diverse perspectives, summarize the discrepancies, and organize them into a checklist for refinement. Another work (Madaan et al., 2023) presents a method to improve the outputs of LLMs through iterative feedback and refinement. Self-reflection allows agents to rapidly learn from trial-and-error, showing signifi-

cant performance improvements across various tasks (Shinn et al., 2023). Additionally, self-reflection enables LLMs to generate, critique, and refine Retrieval Augmented Generation (RAG) (Zhang et al., 2024a). This approach has shown promise in improving the model’s ability to handle nuanced tasks such as translation (Wang et al., 2024a) and scientific research (Smith & Doe, 2022), promoting a deeper and more nuanced understanding of complex topics. Self-reflection in its naive form, where the model is simply instructed to reflect on its own answers, can suffer from degeneracy-of-thought, where the agent, after finding a high confidence answer to the user’s query fails to diverge from its answer in the following steps (Liang et al., 2024; Wang et al., 2024b). Here, we aim to address this issue by utilizing self-opposition generation as a form of self-critique, integrating it with the initial thought process in a Hegelian dialectical fashion to promote the progression of ideas and enhance problem-solving capabilities.

Philosophy Inspired NLP Researchers have recently examined NLP through a philosophical lens, bridging the gap between computational techniques and classical philosophical debates. For example, recent work by Milliere et al. (Milliere & Buckner, 2023; Milliere & Buckner, 2024) explores how LLMs challenge and extend traditional philosophical questions about cognition, semantics, and the nature of intelligence. By revisiting these classic debates, researchers seek to deepen their understanding of both artificial and human intelligence, fostering a more holistic approach to NLP. In a similar vein, philosophical prompting techniques for LLMs take inspiration from classical philosophical methods to enhance their reasoning and cognitive abilities. A prominent example is the Socratic method (Chang, 2023), which involves posing a series of questions to stimulate critical thinking and clarify ideas. This technique can be adapted for LLMs by creating prompts that lead the model through a logical sequence of questions, cultivating deeper analysis and more accurate responses. Another example is the Kantian-inspired UPAR framework (Understand, Plan, Act, Reflect) (Geng et al., 2023), which structures the prompting process to emulate human cognitive processes. It involves understanding the context, planning a response, executing the plan, and then reflecting on the outcome to refine future responses.

3. Background

Hegelian Dialectical Method The term Dialectic denotes a logical debate involving the periodic consideration and resolution of opposing views. This concept is central to Hegel’s dialectic, proposed by the German philosopher George Wilhelm Friedrich Hegel in 1807 (Hegel, 1807; 2010). According to Plato’s dialectic, if the premises of an argument lead to a contradiction, they must be rejected, and there is no

further advancement. In contrast, Hegel’s dialectic views oppositions as essential for achieving higher unity and synthesis, facilitating the ongoing progression of ideas (Hegel, 1807; Popper, 1940; Maybee, 2020)

Definition 3.1 (Informally taken from (Inoue, 2014)). *Assume that L_1 and L_2 are first-order logic (FOL) languages, and that T_1 and T_2 are consistent theories formulated in L_1 and L_2 , respectively. Let $w \in L_1 \cap L_2$ be a sentence such that: $T_1 \vdash w$, and $T_2 \not\vdash w$. Then we say w dialectically opposes T_2 relative to T_1 , and that T_1 and T_2 are dialectically opposing with respect to w . If also, $T_2 \vdash \neg w$, we say w dialectically contradicts T_2 relative to T_1 , and that T_1 and T_2 are dialectically contradicting with respect to w .*

Hegel conceptualizes the dialectical process as consisting of three distinct “moments” or stages. The initial stage, known as the moment of “*understanding*,” represents a period of stability in which a concept or idea appears to have a fixed definition. The second stage, known as the “*dialectical*” moment, emerges when inherent flaws or limitations within the first stage become evident, prompting the evolution of the original idea into one that stands in opposition to it in a specific manner (i.e., in response to its particular limitations). Hegel characterizes this transition as a process of “*self-sublation*” because it is the defects of the first moment or original stage itself that lead to the emergence of the second moment. This second stage “*sublates*” the first moment because it simultaneously negates and preserves it: it supersedes the first, but it is still shaped by the very limitations it seeks to address, thereby remaining inherently connected to the initial concept. The interaction between these two stages culminates in the third “*speculative*” moment, which reconciles and unifies the first two stages. This third moment is not arbitrary, but rather a result of the specific opposition between the earlier stages, and as such, it possesses a distinct character derived from that opposition. This speculative moment is more advanced and comprehensive, as it integrates and transcends the preceding stages, thereby entailing a more universal perspective. This final stage can be understood to be a secondary-level process, as it sublates both previous moments: it negates the earlier stages by presenting a new idea (concept) while still depending on them for its own definition, effectively building upon and transforming the earlier moments (Maybee, 2020).

A characteristic of Hegel’s unifying step (speculative moment) is that the oppositions do not need to be resolved, allowing for the expression of ambiguity given competing perspectives (Hegel, 1951; 1807). To keep the result of the unifying step consistent, we choose to resolve the introduced oppositions via qualifications and negations, all the while accumulating the premises that results in oppositions as part of the unifications, so as to have a list of all individually self-opposing points considered. Moreover, we note that the allowance of oppositions in the unified idea resulted in criti-

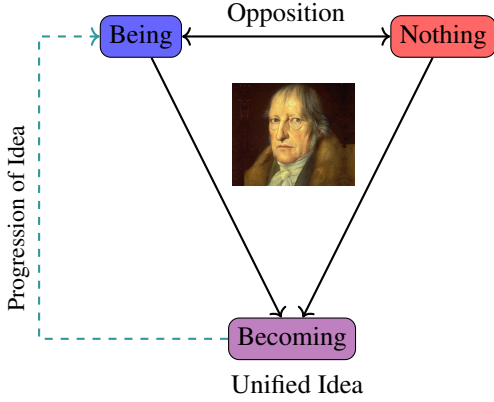


Figure 2. An example of the Hegelian dialectic.

cism from numerous analytical philosophers (Popper, 1940; Russell, 1959; Maybee, 2020), specifically for not abiding by Aristotle’s so-called law of non-contradiction (Gottlieb, 2023). We simply point out that there are logical systems capable of describing Hegel’s unifying step like para-consistent logic (Inoue, 2014; Maybee, 2020).

The Hegel’s dialectic has been used in different domains, including politics, where it is used to understand the evolution of political ideologies. For example, it has been applied to international politics, particularly in understanding the ideological conflicts between western and post-colonial Middle Eastern countries (Boukhatem, 2022). In economics, it can be applied to understand the development of economic theories. For instance, to understand interplay between economics and ethics (Fraser, 2013). In physical sciences, it can be used in astronomy to resolve oppositions in geocentric (Ptolemaic) and heliocentric (Copernican) theories, in physics to address the dialectical opposition between classical mechanics and the special theory of relativity (Inoue, 2014), to explore the interaction of opposing forces, their resolutions, and the emergence of higher-order dynamics that shape ecological processes (Angeler & Maybee, 2025). By synthesizing conflicting hypotheses, scientists can develop more comprehensive and accurate models of scientific phenomena, and this is the main inspiration for this paper.

4. Proposed Hegelian Self-reflection

Inspired by the Hegelian Dialectic, we employ an iterative process, allowing an LLM to evaluate an initial response or “*Proposition*” through a generated critique, “*Opposition*”, and subsequently generate a new response, henceforth referred to as “*Unified response*”, incorporating the best elements of the two. This method aims to improve the model’s ability to integrate complex and opposing information, leading to more coherent and sound responses.

Additionally, we explore two temperature configurations and their effect on the generated responses: 1) a dynamic approach by introducing an annealing scheduler for the LLM’s temperature. This ensures high creativity in the early stages of speculation (unifying step), gradually reducing it and emphasizing the addition of subtle nuances to the unified responses and, 2) a static approach that maintains a constant temperature throughout all iterations.

As the quality of new ideas is difficult to measure, our goal in generating novel ideas is to meet two main criteria: (1) validity, ensuring that the generated response adheres to all instructions of the speculation, and (2) novelty, assessing whether the new idea introduces original content beyond the initial proposition. To do so, we draw inspiration from Marvin Minsky’s concept (Minsky, 1988) a.k.a. “Society of Mind” (SoM), where intelligence is attained through computational modules that interact and collaborate with each other to achieve objectives that cannot be accomplished by a single module alone (Zhuge et al., 2023).

4.1. Algorithm Overview

Algorithm 1 demonstrates our self-reflection method. The core steps of our proposed algorithm are as follows:

Initial Step: The process begins with an *Initial proposition* T_0 , which serves as the starting point for the dialectic.

Dialectical Self-sublation Then at the i th iteration, the algorithm prompts the LLM \mathcal{M} to generate an opposition A_i based on the current proposition T_i and the constant temperature τ_A . This ensures that the generation of opposition is independent of the current iteration, providing stability.

$$A_i \leftarrow \text{Generate Opposition}(T_i, \tau_A); i \in \mathbb{N}$$

Speculation and Annealing-based Scheduler Next, \mathcal{M} is prompted to produce a *unified response* S_i by combining the current proposition and opposition as follows:

$$S_i \leftarrow \text{Cancel \& Unify}(T_i, A_i, \tau(i)); i \in \mathbb{N}$$

An unconsidered perspective in the opposition is necessary, but not sufficient, for a meaningful speculation process; hence, the self-sublation temperature τ_A has an indirect impact on the quality of the unified response:

$$A_i = f(\tau_A) \wedge S_i = g(A_i) \implies S_i = g(f(\tau_A))$$

But the speculation process, directly depends on the $\tau(t)$, which decays over time according to an annealing schedule:

$$\tau(i) = \tau_0 \cdot e^{-\theta i}; i \in \mathbb{N}$$

where the *Decay Constant* θ controls the rate at which the temperature decreases during the process. The temperature

influences the level of randomness during the speculation. It dictates the pace at which the algorithm transitions from the exploration phase to refinement as it progresses. In our dynamic configuration, as the algorithm advances, the temperature decreases, resulting in more stable and refined responses, that is, a more detailed idea (solution) that considers potential oppositions. Initially, a higher temperature promotes greater exploration of potential ideas (solutions). However, as the process continues, the decreasing temperature leads to more stable and refined outcomes. On the other hand, in our constant temperature configuration, we set $\theta = 0$ to ensure a more stable behavior in the generation.

Subsequently, the new propositions T_{i+1} is updated with the unified output S_i from the previous step:

$$T_{i+1} \leftarrow S_i; i \in \mathbb{N}$$

which is then carried forward for the next iteration. We formalize the entire Hegelian self-reflection framework using a probabilistic model as follows:

$$P_{\tau(t), \tau_A}(S_t, A_t, T_t) = \underbrace{P(T_t)}_{\text{Understanding}} \underbrace{P_{\tau_A}(A_t | T_t)}_{\text{Sublation}} \underbrace{P_{\tau(t)}(S_t | T_t, A_t)}_{\text{Speculation}}$$

where $\tau(t) \in \{\tau_0, \dots, \tau_n\}$ & $\tau_A \in \mathbb{R}^+$

4.2. Measuring Validity and Novelty of a New Ideas

Inspired by ‘‘Society of Mind’’ concept and multi-agent voting, we employ multiple LLMs to assess validity and novelty of the unified ideas. In particular, we separately prompt three LLMs and ask if the generated response has followed the Speculation i.e., canceling and unifying instructions (validity test), and if the information in the unified idea is more novel than the proposition (novelty test). We then take the majority vote of the LLMs’ opinions. If there is consensus on both validity and novelty, we proceed with the dialectic process; otherwise, we stop at that point. With this in mind, we define the following score to measure novelty:

$$\text{Novelty Score} = \frac{\text{Iterations Voted as Novel by MAMV}}{\text{Total Dialectical Iterations}}$$

In fact, we adopt a *greedy strategy* as a *heuristic* for novelty. This approach is easy to implement, but we acknowledge that it has its pros. and cons. For instance, if a step does not meet our novelty criteria, we do not proceed. In reality, some debates may go back and forth without producing significant novelty until later stages. This stringent condition might limit creativity. Thus, for a system employing it to create diverse self-reflective processes, adjusting the novelty prompt’s strength or the voting condition could yield different results. Algorithm 2 describes MAMV framework.

5. Prompt Engineering

This section outlines the prompts utilized in our method.

Algorithm 1 Self-Reflection Using the Hegelian Dialectic

```

1: Input: LLM  $\mathcal{M}$ , Sublation Prompt  $p_A$ , Speculation Prompt  $p_S$ , Initial Proposition  $T_0$ , Initial Temperature  $\tau_0$ , Sublation Temperature  $\tau_A$ , Decay Constant  $\theta$ , Number of Iterations  $N$ .
2: Output: Final Idea (Solution)  $S_N$ 
3: Define temperature schedule function  $\tau(t)$  for speculation as a function of iteration  $t$ :
4: Initialize  $T \leftarrow T_0$  {Set the initial proposition}
5: for  $i = 0$  to  $N$  do
6:    $A_i \leftarrow \mathcal{M}(T_i, \tau_A, p_A)$  {Generate opposition using a constant temperature}
7:    $\tau(t) = \tau_0 \cdot e^{-\theta t}$  {Update speculation temperature}
8:    $S_i \leftarrow \mathcal{M}(T_i, A_i, \tau(i), p_S)$  {Unify and cancel proposition and opposition}
9:   if  $MAMV(T_i, A_i, S_i) == \text{Yes}$  then
10:    Accept the new idea (solution)  $S_i$ ;
11:   else
12:     Break;
13:   end if
14:    $T_{i+1} \leftarrow S_i$  {New proposition for the next iteration}
15: end for
16: return  $S_N$  {Return the final idea (solution) after  $N$  iterations}
    
```

Algorithm 2 MAMV Validity & Novelty Assessment

```

1: Input: Proposition  $T$ , Unified Idea  $S$ , Novelty prompt  $p_N$ , Validity prompt  $p_V$ , List of LLMs  $L$ 
2: Output: Final Assessment Yes/No
3:  $\text{votes}_{\text{novelty}} \leftarrow []$ 
4:  $\text{votes}_{\text{validity}} \leftarrow []$ 
5: for  $llm_i \in L = \{llm_0, \dots, llm_n\}$  do
6:    $\text{vote}_i \leftarrow llm_i(p_N, T, S)$  {Assess Novelty of  $S$  compared to  $T$ }
7:    $\text{votes}_{\text{novelty}}.\text{append}(\text{vote}_i)$ 
8:    $\text{vote}_i \leftarrow llm_i(p_V, T, A, S)$  {Assess validity of  $S$ , given  $T$  and  $A$ }
9:    $\text{votes}_{\text{validity}}.\text{append}(\text{vote}_i)$ 
10: end for
11: if  $\text{Majority}(\text{votes}_{\text{validity}}) == \text{Yes}$  &  $\text{Majority}(\text{votes}_{\text{novelty}}) == \text{Yes}$  then
12:   Return Yes
13: else
14:   Return No
15: end if
    
```

5.1. Self-reflection Prompting

Sublation: Figures 5 and 7 in Appendix D show the sublation prompts used for idea generation and math reasoning, respectively. Engaging in self-debating is one way to generate diverse viewpoints. To simulate the sublation process, we employ an iterative self-debating, Solo Performance Prompting (SPP) strategy. SPP harnesses the model’s theory of mind reasoning abilities by instructing it to ‘‘split’’ into various personas and collaborate on a given prompt through a brainstorming session among these personas (Wang et al., 2024b). Our framework requests the model to generate an arbitrary persona for each iteration of the self-dialectic. The following instructs model to practice SPP:

Imagine you are someone X who has noticed a problem with or defect in a proposed view. Produce an opposed view on the same topic that corrects for the noted defect or problem.

However, it is necessary to explicitly define the criteria for what constitutes potentially relevant types of defects. For example, for idea generation, we instruct model with:

Possible defects are:

1. The proposed view has constitutive elements that are not fully defined;
2. The proposed view does not account for some relevant phenomena or examples (its content is incomplete);
3. The proposed view is incompatible with some of the phenomena or examples that is supposed to include (incompatible with its content);
4. The proposed view cannot be fully defined or determined on its own (its definition or determination is incomplete on its own), but can be defined determined only in relation to another view.

Speculation: Rather than choosing one view over the other, speculation seeks to unify opposing ideas by proposing a third perspective, a unifying theory as follows:

Produce a third view on the same topic that:

- Unifies the previous two views in relation to the defect or problem on which the two views were opposed to one another.
- This third view must be a unifying theory that explains how the two views agree with or are the same as one another and how the two views are opposed to one another.

Figures 6 and 8 in Appendix D, illustrate the speculation prompts used for idea generation and math reasoning.

5.2. MAMV Prompting

Validity: To verify that the model has adhered to the speculation instructions, we provide MAMV with the instructions for the unifying process and ask each model to vote “yes” or “no” on whether the instructions are followed. Figure 9 in Appendix D demonstrates this process.

Novelty: Evaluating novelty is more challenging, as it requires domain experts who are familiar with all contributions in the field. Here, we define novelty as the introduction of new ideas that build upon the previous step’s propositions i.e., T_{i-1} . The corresponding prompt is reported in Appendix D, Figure 10. We emphasize that our method generates new ideas that may not always be scientifically correct. Our focus, however, is on the validity of the dialectical process provided through instructions for the generation of the unified idea, as well as the novelty of these ideas compared to the propositions and oppositions in previous steps. The scientific evaluation of these ideas could be examined by experts in the field using various scientific methods. We stress that evaluating novelty is not just about adding new information to the previous proposition; it also requires a thorough understanding of existing literature to ensure the ideas are truly unexplored. However, this does not undermine the effectiveness of our method as an early exploration but highlights the need for better evaluations, potentially involving human/AI experts to develop more effective ways.

6. Experiments

In this section, we explore the application of our method to solve reasoning tasks and generate innovative ideas.

6.1. Experimental Setting

Models: We utilize GPT family models for both quantitative and qualitative experiments. For the qualitative task of novel idea generation, we employ *GPT-4o mini*, *GPT-432k*, and *GPT-4* within the MAMV framework, while *GPT-4o* serves as the core model for dialectical self-reflection.

Dialectical Iterations and MAMV: For simplicity, we set a maximum constant for iterations as indicated in Appendix A, Table 2. For idea generation, we evaluate the validity and novelty of the unified idea produced through self-reflection using an MAMV module external to the self-reflection process. If the MAMV determines that the generated idea lacks sufficient novelty or that the speculation was not properly followed, the process is halted. This approach prevents redundant iterations over existing knowledge and establishes a clear stopping point. Aside from evaluation, the MAMV does not interfere with the model’s self-reflection and idea generation. For reasoning task, we do not use MAMV or majority voting, as our goal is to evaluate the effectiveness of dialectical self-reflection isolated from a multi-agent framework or voting, which could be applied to any technique.

Sublation Temperature τ_A For the idea generation task, we observed that τ_A during the sublation step directly influences novelty. Lower τ_A produce arguments that closely engage with the original proposition, while higher τ_A tend to generate more divergent, even contradictory, perspectives. To strike a balance, we set τ_A at 0.5.

Speculation Temperature and Annealing: We conduct two sets of experiments: one incorporating the annealing process and one omitting it, i.e., constant temperature. This allows us to evaluate the impact of dynamic vs. static temperature. Note that, for dynamic configuration, we grid searched the decay parameter θ and chose a relatively high τ_0 to start with, ensuring high creativity at the beginning. One criterion for selecting θ is its ability to result in a higher number of iteration steps, that is, higher novelty scores. Figures 3 and 4 present experiments to determine the optimal θ in the idea generation task in the annealing setting, and the optimal constant value τ in the static setting, respectively.

6.2. Quantitative Results: Math & Symbolic Reasoning

We evaluate the reasoning capabilities of our method on both mathematical and symbolic reasoning tasks, focusing on GSM-8k, GSM-hard, and GSM-Symbolic datasets. As demonstrated in Table 4, dialectical prompting, which employs self-reflection under both temperature settings, significantly improves performance over traditional methods such as zero-shot, few-shot, and few-shot with CoT across multiple models and benchmarks. The annealing temperature strategy offers slightly better performance for certain models and datasets, though in some instances, a constant temper-

Model	Context Length	Prompt	GSM-8k	GSM-Symbolic	GSM-hard
GPT-4o	128k	Zero-shot	0.863±0.004	0.758±0.006	0.597±0.003
		Zero-shot+CoT	0.863±0.004	0.760±0.004	0.587±0.004
		Few-shot	0.943±0.002	0.851±0.003	0.642±0.004
		Few-shot+CoT	0.944±0.002	0.843±0.005	0.640±0.008
		Constant Temp. Dialectic (Ours)	0.951±0.001	0.907±0.004	0.647±0.008
		Annealing Temp. Dialectic (Ours)	0.954±0.001	0.907±0.005	0.650±0.006
GPT-4o-mini	128k	Zero-shot	0.871±0.005	0.709±0.005	0.541±0.007
		Zero-shot+CoT	0.874±0.003	0.697±0.003	0.546±0.003
		Few-shot	0.921±0.000	0.690±0.004	0.552±0.003
		Few-shot+CoT	0.921±0.001	0.708±0.009	0.566±0.007
		Constant Temp. Dialectic (Ours)	0.937±0.006	0.808±0.009	0.577±0.006
		Annealing Temp. Dialectic (Ours)	0.940±0.005	0.805±0.001	0.573±0.005
GPT4-32k	32k	Zero-shot	0.877±0.006	0.794±0.007	0.601±0.007
		Zero-shot+CoT	0.868±0.002	0.784±0.005	0.595±0.003
		Few-shot	0.952±0.004	0.878±0.004	0.651±0.008
		Few-shot+CoT	0.948±0.001	0.870±0.003	0.654±0.004
		Constant Temp. Dialectic (Ours)	0.955±0.000	0.941±0.004	0.653±0.005
		Annealing Temp. Dialectic (Ours)	0.955±0.002	0.941±0.003	0.656±0.002
Qwen 2.5-7B-Instruct	1M	Zero-shot	0.828±0.000	0.513±0.000	0.491±0.000
		Zero-shot+CoT	0.816±0.000	0.516±0.000	0.501±0.000
		Few-shot	0.866±0.000	0.527±0.000	0.517±0.000
		Few-shot+CoT	0.856±0.000	0.511±0.000	0.499±0.000
		Constant Temp. Dialectic (Ours)	0.917±0.000	0.630±0.000	0.544±0.000
		Annealing Temp. Dialectic (Ours)	0.920±0.000	0.625±0.000	0.553±0.000
Phi-4	16k	Zero-shot	0.869±0.000	0.741±0.000	0.560±0.000
		Zero-shot+CoT	0.861±0.000	0.742±0.000	0.559±0.000
		Few-shot	0.949±0.000	0.825±0.000	0.619±0.000
		Few-shot+CoT	0.936±0.000	0.806±0.000	0.611±0.000
		Constant Temp. Dialectic (Ours)	0.949±0.000	0.843±0.000	0.625±0.000
		Annealing Temp. Dialectic (Ours)	0.941±0.000	0.852±0.000	0.623±0.000

Table 1. Math reasoning: dialectical self-reflection vs. other prompting techniques.

ature performs marginally better. This could indicate that models with stronger reasoning capabilities are less affected by temperature changes. It is also important to note that not all temperatures have been explored, and identifying the optimal one may be challenging. However, in general, both the constant and the annealing approaches enhance mathematical reasoning (GSM-8K/hard) and symbolic reasoning (GSM-Symbolic), with particularly pronounced gains in symbolic tasks. It also reduces output variance, leading to more stable and reliable reasoning. The GSM-8k/hard focuses on natural language math problems requiring multi-step logical reasoning, while GSM-Symbolic involves abstract symbolic manipulation and formal logic. The larger improvements observed on GSM-Symbolic suggest that dialectical prompting especially benefits structured, rule-based reasoning tasks by enabling iterative refinement and more precise exploration of solution spaces. Figure 12 in Appendix E presents a dialectical iteration applied to a math problem, demonstrating the effectiveness of our approach.

Since *dialectical self-reflection prompting is relatively long, it tends to perform better on models with larger context lengths*. However, when the context window is smaller (e.g., Phi-4), the performance gains from this technique are smaller; although marginal, it still typically yields the best performance among other methods.

6.3. Qualitative Results: Generation of Scientific Ideas

Observations and Key Findings Based on hundreds of experiments with both annealing and constant τ settings

across slightly varied prompts, and a thorough analysis of GPT-4o’s responses, we summarize our key findings as:

- Even when using the same hyper-parameters and prompts with the same LLM, we may receive varying responses, resulting in different numbers of steps and novelty scores. Nevertheless, in most cases, the generated ideas with the same configuration and setting tend to stay within the same topic area.
- In a significant number of experiments, although not necessarily in all instances, lower temperatures result in fewer steps and, consequently, lower novelty scores as shown in Figure 4. Analyzing the contextual meaning of such cases also confirms a decrease in creativity.
- Maintaining a high constant speculation temperature may lead to out-of-topic unified ideas in some cases, while in others, it can result in novel ideas.
- The most novel unifications typically occur in the early stages. As the process progresses, ideas become more refined and begin to incorporate opposing viewpoints, making it increasingly difficult to generate meaningful opposition. Occasionally, the opposition enters a repetitive loop, indicating that further challenges to the unified idea are less effective, suggesting that the idea has become more comprehensive and stable.
- Regarding the reasoning, the level of agreement in the unified idea and the intensity of the opposition from the opposed idea (opposing qualifications vs. fundamental premises) can also be viewed as prompt-based hyper-parameters that influence the self-debate.
- During the grid search for τ_A , we found that it signif-

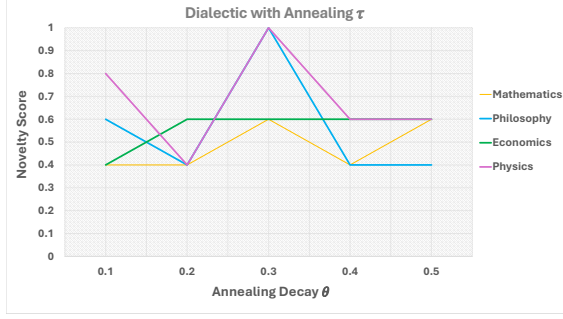


Figure 3. Low annealing decays mimic the effect of a constant high temperature setting, whereas high annealing decays act similar to a low temperature setting. In our grid search for θ , the highest novelty score is achieved with a decay setting of 0.3.

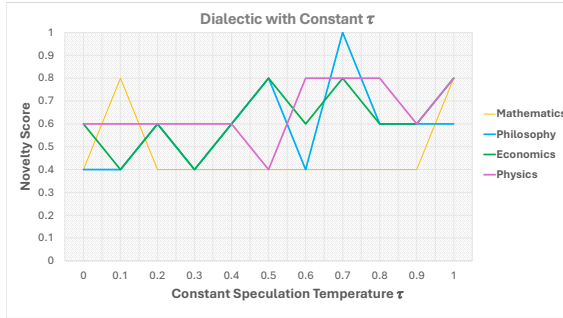


Figure 4. In our constant temperature setting, a high temperature enhances the novelty score, which may or may not lead to generated ideas drifting away from the initial topic. Conversely, a lower temperature reduces the novelty score, ensuring that ideas remain more closely aligned with the original topic.

icantly affects novelty. Lower τ_A produce arguments directly addressing the proposition, while higher τ_A generate almost orthogonal perspectives that contradict it. Resolving these oppositions involves qualifying both the opposing views, leading to epistemological discussions. This results in unified ideas that simply acknowledge different perspectives, which are consistent within themselves but too agreeable to generate new opinions. Thus, we set it to 0.5 as a compromise.

- Our hyper-parameters and prompts are not a universal solution for all topics and they need to be tuned.
- We note that having more statements in proposition increases the likelihood of generating opposing views. This results in a more meaningful opposition and leads to additional dialectical steps.

Examples of ideas generated by our annealing setting for math, physics, economics, and philosophy are illustrated in Tables 4, 5, 6 and 7, respectively. Furthermore, a full response generated by GPT-4o for a dialectical iteration of idea generation, is demonstrated in Appendix E, Figure 11.

7. Future Directions

In this paper, for simplicity, we used a constant τ_A . However, as we discussed earlier, it directly affects novelty, with lower τ_A producing arguments directly addressing the initial proposition and higher τ_A results in orthogonal perspectives that lead to epistemological inquiries. Investigating multiple oppositions with varying τ_A simultaneously, generating unified ideas, and backtracking from undesired outcomes is worthwhile. We reserve this exploration for future research.

Currently, with the novelty score stop condition, our framework strives to resolve dialectical contradictions every cycle. In future implementations, the statements from opposing view points might be explicitly considered as ambiguous and not integrated into the unified idea until enough iterations have passed to resolve them.

Meanwhile, in the MAD setting, the Multi-Persona framework addresses the degeneracy-of-thought issue in naive self-reflection by pre-assigning one agent to express viewpoints and another to oppose them (Liang et al., 2024). By adjusting the likelihood of agreement in debate protocols, the Multi-Persona MAD framework surpasses other MAD frameworks in the Q&A setting by tuning the agreeability of debating agents. (Liang et al., 2024; Smit et al., 2024). Similarly, the decisiveness of the speculation step can be tuned and compared with the performance of agreeable/disagreeable MAD frameworks. For further commentary on the limitations, refer to Appendix B.

8. Concluding Remarks

This paper introduces a novel technique for LLM self-reflection by framing it as a self-dialectical process grounded in Hegel’s dialectics. It follows a periodic cycle of understanding, sublation, and speculation, in which an initial proposition is critically evaluated through generated opposing views and then refined into a unified idea that integrates the strongest elements of both views. Meanwhile, we investigate the impact of speculation temperature on the novelty of generated ideas by establishing two configurations: (1) a constant and (2) a dynamic annealing temperature. To assess the novelty of generated idea, we employ MAMV, a multi-agent system designed to collaborate and vote based on proposed validity and novelty metrics. Our experiments on math and symbolic reasoning tasks, show significant improvement. In addition, we applied our method to a qualitative task namely, novel idea generation, which resulted in promising outcomes. We observed that high constant temperatures occasionally produced off-topic or highly novel ideas, while the most innovative ideas typically emerged early in the annealing setting, becoming more nuanced and stable over time.

References

- Abdali, S., Anarfi, R., Barberan, C., and He, J. Securing large language models: Threats, vulnerabilities and responsible practices, 2024a. URL <https://arxiv.org/abs/2403.12503>.
- Abdali, S., Anarfi, R., Barberan, C., and He, J. Decoding the ai pen: Techniques and challenges in detecting ai-generated text. KDD '24, pp. 6428–6436, New York, NY, USA, 2024b. Association for Computing Machinery. ISBN 9798400704901. doi: 10.1145/3637528.3671463. URL <https://doi.org/10.1145/3637528.3671463>.
- Abdali, S., He, J., Barberan, C., and Anarfi, R. Can llms be fooled? investigating vulnerabilities in llms, 2024c. URL <https://arxiv.org/abs/2407.20529>.
- Amirizani, M., Martin, E., Sivachenko, M., Mashhadi, A., and Shah, C. Can llms reason like humans? assessing theory of mind reasoning in llms for open-ended questions. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management, CIKM '24*, pp. 34–44, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400704369. doi: 10.1145/3627673.3679832. URL <https://doi.org/10.1145/3627673.3679832>.
- Angeler, D. G. and Maybee, J. E. Chapter three - dialectical ecosystems: Theory and applications. In Dumbrell, A. J. (ed.), *Ecological Horizons: From Nature to People, Part 1*, volume 72 of *Advances in Ecological Research*, pp. 39–90. Academic Press, 2025. doi: <https://doi.org/10.1016/bs.aecr.2025.02.002>. URL <https://www.sciencedirect.com/science/article/pii/S0065250425000029>.
- Bobzien, S. and Duncombe, M. Dialectical School. In Zalta, E. N. and Nodelman, U. (eds.), *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Fall 2023 edition, 2023.
- Boukhatem, H. Hegelian dialectics applications in the 21st century politics. *Revue Akofena*, 2022.
- Cambridge University Press. Dialectic, n. URL <https://dictionary.cambridge.org/us/dictionary/english/dialectic>.
- Chan, C.-M., Chen, W., Su, Y., Yu, J., Xue, W., Zhang, S., Fu, J., and Liu, Z. Chateval: Towards better llm-based evaluators through multi-agent debate, 2023. URL <https://arxiv.org/abs/2308.07201>.
- Chang, E. Y. Prompting large language models with the socratic method. *arXiv preprint arXiv:2303.08769*, 2023.
- Chen, D., Shi, J., Wan, Y., Zhou, P., Gong, N. Z., and Sun, L. Self-cognition in large language models: An exploratory study. *Proceedings of the ICML 2024 Large Language Models and Cognition Workshop*, 2024.
- Christiano, P., Leike, J., Brown, T. B., Martic, M., Legg, S., and Amodei, D. Deep reinforcement learning from human preferences, 2023.
- Christiano, P. F., Leike, J., Brown, T., Martic, M., Legg, S., and Amodei, D. Deep reinforcement learning from human preferences. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.
- Cobbe, K., Kosaraju, V., Bavarian, M., Chen, M., Jun, H., Kaiser, L., Plappert, M., Tworek, J., Hilton, J., Nakano, R., Hesse, C., and Schulman, J. Training verifiers to solve math word problems. *ArXiv*, abs/2110.14168, 2021.
- Delahaye, D., Chaimatanan, S., and Mongeau, M. *Simulated Annealing: From Basics to Applications*, pp. 1–35. Springer International Publishing, Cham, 2019. ISBN 978-3-319-91086-4. doi: 10.1007/978-3-319-91086-4.1. URL https://doi.org/10.1007/978-3-319-91086-4_1.
- Dong, Q., Li, L., Dai, D., Zheng, C., Ma, J., Li, R., Xia, H., Xu, J., Wu, Z., Chang, B., Sun, X., Li, L., and Sui, Z. A survey on in-context learning. In Al-Onaizan, Y., Bansal, M., and Chen, Y.-N. (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 1107–1128, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.64. URL <https://aclanthology.org/2024.emnlp-main.64/>.
- Du, Y., Li, S., Torralba, A., Tenenbaum, J., and Mordatch, I. Improving factuality and reasoning in language models through multiagent debate, 05 2023.
- Fraser, I. Hegel on modern economics and business. In Lütge, C. (ed.), *Handbook of the Philosophical Foundations of Business Ethics*, pp. 415–431. Springer, 2013. doi: 10.1007/978-94-007-1494-6_78.
- Geng, H., Xu, B., and Li, P. Upar: A kantian-inspired prompting framework for enhancing large language model capabilities. *arXiv preprint arXiv:2310.01441*, 2023.
- Gottlieb, P. Aristotle on Non-contradiction. In Zalta, E. N. and Nodelman, U. (eds.), *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Winter 2023 edition, 2023.

- Guu, K., Lee, K., Tung, Z., Pasupat, P., and Chang, M.-W. Realm: Retrieval-augmented language model pre-training. ICML'20. JMLR.org, 2020.
- Hegel, G. W. F. *The Phenomenology of Spirit*. Cambridge University Press (2019), 1807. ISBN 978-0-268-10349-1.
- Hegel, G. W. F. (ed.). *Hegel's Science of Logic*. Humanity Books, Amherst, N.Y., 1951.
- Hegel, G. W. F. *The Science of Logic [Wissenschaft der Logik]*. Cambridge University Press, New York, 2010.
- Inoue, K. Dialectical contradictions and classical formal logic. *International Studies in the Philosophy of Science*, 28(2):113–132, 2014. doi: 10.1080/02698595.2014.932526.
- Kadavath, S., Conerly, T., Askell, A., Henighan, T., Drain, D., Perez, E., Schiefer, N., Hatfield-Dodds, Z., DasSarma, N., Tran-Johnson, E., Johnston, S., El-Showk, S., Jones, A., Elhage, N., Hume, T., Chen, A., Bai, Y., Bowman, S., Fort, S., Ganguli, D., Hernandez, D., Jacobson, J., Kernion, J., Kravec, S., Lovitt, L., Ndousse, K., Olsson, C., Ringer, S., Amodei, D., Brown, T., Clark, J., Joseph, N., Mann, B., McCandlish, S., Olah, C., and Kaplan, J. Language models (mostly) know what they know, 2022.
- Kaufmann, W. (ed.). *Hegel: Reinterpretation, Texts, and Commentary*. Doubleday & Company, Garden City, New York, 1965.
- Lewkowycz, A., Andreassen, A., Dohan, D., Dyer, E., Michalewski, H., Ramasesh, V., Slone, A., Anil, C., Schlag, I., Gutman-Solo, T., Wu, Y., Neyshabur, B., Gur-Ari, G., and Misra, V. Solving quantitative reasoning problems with language models, 2022.
- Li, S., Yang, C., Cheng, Z., Liu, L., Yu, M., Yang, Y., and Lam, W. Self-contrast: Better reflection through inconsistent solving perspectives. *Proceedings of the ICML 2024 Large Language Models and Cognition Workshop*, 2024.
- Liang, T., He, Z., Jiao, W., Wang, X., Wang, Y., Wang, R., Yang, Y., Shi, S., and Tu, Z. Encouraging divergent thinking in large language models through multi-agent debate, 2024. URL <https://arxiv.org/abs/2305.19118>.
- Madaan, A., Tandon, N., Gupta, P., Hallinan, S., Gao, L., Wiegrefe, S., Alon, U., Dziri, N., Prabhunoye, S., Yang, Y., Gupta, S., Majumder, B. P., Hermann, K., Welleck, S., Yazdanbakhsh, A., and Clark, P. Self-refine: Iterative refinement with self-feedback, 2023.
- Maybee, J. E. Hegel's Dialectics. In Zalta, E. N. (ed.), *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Winter 2020 edition, 2020.
- Milliere, R. and Buckner, C. A philosophical introduction to language models – part ii: The way forward. *arXiv preprint arXiv:2405.03207*, 2024.
- Millière, R. and Buckner, C. A philosophical introduction to language models – part i: Continuity with classic debates. *arXiv preprint arXiv:2303.12345*, 2023.
- Minsky, M. *Society of Mind*. Simon and Schuster, 1988.
- Muneeswaran, I., Saxena, S., Prasad, S., Prakash, M. V. S., Shankar, A., Varun, V., Vaddina, V., and Gopalakrishnan, S. Minimizing factual inconsistency and hallucination in large language models. *ArXiv*, abs/2311.13878, 2023.
- Nye, M., Andreassen, A. J., Gur-Ari, G., Michalewski, H., Austin, J., Bieber, D., Dohan, D., Lewkowycz, A., Bosma, M., Luan, D., Sutton, C., and Odena, A. Show your work: Scratchpads for intermediate computation with language models, 2022.
- Popper, K. R. What is dialectic? *Mind*, 49(194):403–426, 1940. doi: 10.1093/mind/XLIX.194.403.
- Rajani, N. F., McCann, B., Xiong, C., and Socher, R. Explain yourself! leveraging language models for common-sense reasoning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4932–4942, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1487.
- Russell, B. My philosophical development. *Les Etudes Philosophiques*, 14(4):558–558, 1959.
- Shayegani, E., Mamun, M. A. A., Fu, Y., Zaree, P., Dong, Y., and Abu-Ghazaleh, N. B. Survey of vulnerabilities in large language models revealed by adversarial attacks. *ArXiv*, abs/2310.10844, 2023.
- Shinn, N., Cassano, F., Labash, B., Gopinath, A., Narasimhan, K., and Yao, S. Reflexion: Language agents with verbal reinforcement learning, 2023.
- Smit, A., Duckworth, P., Grinsztajn, N., Barrett, T. D., and Pretorius, A. Should we be going mad? a look at multi-agent debate strategies for llms, 2024. URL <https://arxiv.org/abs/2311.17371>.
- Smith, J. and Doe, J. Enhancing scientific research through self-reflection: A methodological approach. *Journal of Research Methodologies*, 15(3):120–135, 2022. doi: 10.1234/jrm.2022.5678.
- Smith, J. and Doe, J. Large language models for scientific synthesis, inference, and explanation. *arXiv preprint arXiv:2310.07984*, 2023.

- Wang, X., Wei, J., Schuurmans, D., Le, Q. V., Chi, E. H., Narang, S., Chowdhery, A., and Zhou, D. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*, 2023.
- Wang, Y., Zeng, J., Liu, X., Meng, F., Zhou, J., and Zhang, M. Taste: Teaching large language models to translate through self-reflection. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6144–6158. Association for Computational Linguistics, 2024a.
- Wang, Z., Mao, S., Wu, W., Ge, T., Wei, F., and Ji, H. Unleashing the emergent cognitive synergy in large language models: A task-solving agent through multi-persona self-collaboration, 2024b. URL <https://arxiv.org/abs/2307.05300>.
- Wei, J., Wang, X., Schuurmans, D., Bosma, M., brian ichter, Xia, F., Chi, E. H., Le, Q. V., and Zhou, D. Chain of thought prompting elicits reasoning in large language models. In Oh, A. H., Agarwal, A., Belgrave, D., and Cho, K. (eds.), *Advances in Neural Information Processing Systems*, 2022.
- Wu, L. et al. The impact of large language models on scientific discovery: A preliminary study using gpt-4. *arXiv preprint arXiv:2311.07361*, 2023.
- Xue, T., Wang, Z., Wang, Z., Han, C., Yu, P., and Ji, H. Rcot: Detecting and rectifying factual inconsistency in reasoning by reversing chain-of-thought. *ArXiv*, abs/2305.11499, 2023.
- Zelikman, E., Wu, Y., Mu, J., and Goodman, N. STar: Bootstrapping reasoning with reasoning. In Oh, A. H., Agarwal, A., Belgrave, D., and Cho, K. (eds.), *Advances in Neural Information Processing Systems*, 2022.
- Zhang, M., Wang, Y., Liu, X., Meng, F., Zhou, J., and Li, S. Self-rag: Learning to retrieve, generate, and critique through self-reflection. *Proceedings of the ICML 2024 Large Language Models and Cognition Workshop*, 2024a.
- Zhang, Q. et al. Scientific large language models: A survey on biological & chemical domains. *arXiv preprint arXiv:2401.14656*, 2024b.
- Zhang, Y., Zhang, F., Yang, Z., and Wang, Z. What and how does in-context learning learn? bayesian model averaging, parameterization, and generalization, 2023. URL <https://arxiv.org/abs/2305.19420>.
- Zhang, Y., Chen, X., Jin, B., Wang, S., Ji, S., Wang, W., and Han, J. A comprehensive survey of scientific large language models and their applications in scientific discovery. In Al-Onaizan, Y., Bansal, M., and Chen, Y.-N. (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 8783–8817, Miami, Florida, USA, November 2024c. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.498. URL <https://aclanthology.org/2024.emnlp-main.498>.
- Zhuge, M., Liu, H., Faccio, F., Ashley, D. R., Csord’as, R., Gopalakrishnan, A., Hamdi, A., Hammoud, H., Herrmann, V., Irie, K., Kirsch, L., chuan Li, B., Li, G., Liu, S., Mai, J., Pikekos, P., Ramesh, A., Schlag, I., Shi, W., Stani’c, A., Wang, W., Wang, Y., Xu, M., Fan, D.-P., Ghanem, B., and Schmidhuber, J. Mindstorms in natural language-based societies of mind. *ArXiv*, abs/2305.17066, 2023. URL <https://api.semanticscholar.org/CorpusID:258947227>.
- Ziegler, D. M., Stiennon, N., Wu, J., Brown, T. B., Radford, A., Amodei, D., Christiano, P., and Irving, G. Fine-tuning language models from human preferences. *ArXiv*, abs/1909.08593, 2019.

Appendix

A. Experimental Setting

Table 2 and 3 detail our experimental hyper-parameters.

Hyper-parameter	Value
Initial temperature τ_0	0.7
constant temperature for op- position τ_A	0.5
Decay constant θ	$\theta \in [0.1, 0.5]$
Max iterations for idea gen- eration N	5
Max iterations for math rea- soning N	3

Table 2. Experimental hyper-parameters.

Dataset	Model	Constant		Annealing	
		τ_0	τ_A	τ_0	τ_A
GSM-Hard	GPT-4o	0.7	0.5	0.7	0.5
	GPT-4o-mini	0.7	0.5	0.7	0.5
	GPT4-32k	1.0	0.3	1.0	0.3
	Phi-4	1.0	0.5	1.0	0.5
	Qwen2.5	1.0	0.3	1.0	0.3
	Phi-4-mini-instruct	1.0	0.5	1.0	0.5
GSM-Symbolic	GPT-4o	0.7	0.5	0.7	0.5
	GPT-4o-mini	0.7	0.5	0.7	0.5
	GPT4-32k	0.7	0.5	0.7	0.5
	Phi-4	1.0	0.5	1.0	0.5
	Qwen2.5	1.0	0.3	1.0	0.3
	Phi-4-mini-instruct	1.0	0.5	1.0	0.5
GSM8k	GPT-4o	0.7	0.5	0.7	0.5
	GPT-4o-mini	0.7	0.5	0.7	0.5
	GPT4-32k	1.0	0.3	1.0	1.0
	Phi-4	0.7	0.5	0.7	0.5
	Qwen2.5	1.0	0.3	0.7	1.0
	Phi-4-mini-instruct	1.0	0.5	1.0	0.5

Table 3. Comparison of τ_0 and τ_A values across datasets and models.

B. Limitations

In this section, we briefly discuss some limitations of our current framework and experimental setup.

Deliberate adaptation and interpretation choices in Hegelian dialectic We formulate Hegel’s dialectic as a 2-step progress, where the flow from proposition to opposition and then unified idea does not follow any sort of logical necessity as proposed by (Maybee, 2020; Kaufmann, 1965; Inoue, 2014). This also means that the perspective from which the opposition introduces its opposition is arbitrary, i.e. delegated to the LLM agent. (2) The unified idea always maintains a self-consistent position. In Hegel’s dialectic, statements can remain ambiguous, where they are true from one perspective and false from the other, without requiring a resolution (Hegel, 1807; Maybee, 2020). Interested readers can refer to (Maybee, 2020; Inoue, 2014) for more details.

Measuring novelty of a statement is challenging as it involves subjective assessment and context-awareness. Novelty is not just about adding new information, it also requires understanding existing literature to ensure that the idea has not been previously explored. Comparing an idea solely based on an initial proposition does not account for depth of existing research. Our novelty prompt has inherent limitations. It does not explicitly define whether altering the strength of an argument is considered novel, even if all points have been previously mentioned. Additionally, it remains unclear whether the model should regard the negation of existing premises as novel, as it involves manipulating existing information rather than introducing new information. However, we view such negation as novel.

Difficulties in distinguishing creativity from LLM remembering in baseline evaluation Evaluating new ideas against pre-proposed unified idea is extremely challenging, as distinguishing whether a model is merely recalling training information or generating creative ideas is challenging.

Measuring the effect of annealing due to randomness of generation: We often observed that, a constant τ either fails to produce novel content when set too low or leads to irrelevant information when set too high. However, due to inherent randomness in the generation, it is not straightforward to generalize this observation. Conducting statistical significance tests over multiple rounds of prompting can help to evaluate this observation more effectively.

Reproducibility of results: The inherent randomness in the generation process, coupled with constantly evolving nature of LLMs and the lack of control, especially when using black-box models (e.g. GPT family), makes it difficult to reproduce the results which is essential in scientific settings. While white-box models offer better control over outputs, they may not be as powerful as black-box competitors.

Repetition and randomness in API calls We have occasionally observed identical opposition and unification paragraphs, as well as main points, raising concerns about the extent of randomness and the possible presence of hidden caching mechanisms during these experiments.

Need for domain expert LLMs Utilizing the extensive knowledge of domain-specific LLM experts can help identify unique contributions and ensure comprehensive coverage of the field, making it more feasible to measure novelty.

C. Qualitative Results

This section provides examples from our qualitative experiment, i.e., the generation of new ideas across different domains.

Table 4. Dialectical self-reflection on a Mathematical topic with annealing τ and decay constant $\theta = 0.5$.

Topic 1-Mathematics (Annealing τ)			
# Itr.	Opposing View Point	Unified Idea / Next Understanding	Temp.
0	N.A.	Initial Idea: In mathematics, we can introduce multiple forms of infinity, some of which may be considered larger or smaller than others based on certain metrics, for example time, to allow for a sense of relativity when working with infinity.	N.A.
1	<p>Defects or problems: 1- Defect/Problem 1: The proposed view does not fully define what constitutes "forms of infinity" and how they can be differentiated. 2- Defect/Problem 2: The proposed view does not account for the mathematical rigor required to compare different infinities using metrics like time, which is not traditionally used in mathematical contexts to measure infinity. 3- Defect/Problem 3: The proposed view is incomplete because it does not explain how the concept of relativity applies to infinity in a mathematically coherent way.</p> <p>In mathematics, infinity should be treated as a single abstract concept, with different sizes of infinity compared through set theory and cardinality rather than physical metrics like time.</p>	Infinity in mathematics can be understood in multiple forms through set theory and cardinality, and the concept of relativity can be meaningfully applied within this framework to compare different sizes of infinity in a mathematically coherent way.	0.7
2	<p>Defects or problems: 1- Defect/Problem 1: The proposed view does not adequately define how relativity, traditionally a physical concept, can be rigorously applied within purely mathematical contexts. 2- Defect/Problem 2: The proposed view fails to consider other mathematical frameworks beyond set theory and cardinality that could provide additional insights into the nature of infinity. 3- Defect/Problem 3: The use of time as a metric for comparing infinities is unconventional and not sufficiently justified within the mathematical context, leading to potential ambiguity and lack of coherence.</p> <p>Infinity should be understood through multiple mathematical frameworks, not just set theory and cardinality, and the concept of relativity must be clearly defined and justified within the abstract context to ensure mathematical coherence and rigor.</p>	Infinity in mathematics can be understood through multiple forms and frameworks, such as set theory, cardinality, topology, and analysis, and the concept of relativity can be meaningfully applied within this comprehensive framework to compare different sizes of infinity in a mathematically coherent way.	0.42
3	<p>Defects or problems: 1- Defect/Problem 1: The proposed view does not adequately define how relativity can be rigorously applied within the mathematical context to compare different sizes of infinity. 2- Defect/Problem 2: The proposed view fails to account for the inherent differences between physical metrics like time and abstract mathematical concepts, leading to potential ambiguity. 3- Defect/Problem 3: The proposed view overlooks specific examples and phenomena from topology and analysis that could provide deeper insights into the nature of infinity.</p> <p>A comprehensive understanding of infinity requires precise definitions and rigorous application of mathematical principles, distinguishing abstract concepts from physical metrics, and integrating insights from topology and analysis.</p>	Infinity in mathematics can be understood through multiple forms and frameworks, such as set theory, cardinality, topology, and analysis, and the concept of relativity can be meaningfully applied within this comprehensive framework to compare different sizes of infinity in a mathematically coherent way. Precise definitions and rigorous application of principles are essential, and distinguishing abstract concepts from physical metrics is crucial to maintaining clarity.	0.25
4	<p>Defects or problems: 1- Defect/Problem 1: The proposed view does not fully define how the concept of relativity is applied within mathematical frameworks to compare different sizes of infinity. 2- Defect/Problem 2: The proposed view does not account for the potential incompatibility of using physical metrics like time to measure infinity, which may lead to ambiguity. 3- Defect/Problem 3: The proposed view lacks a clear explanation of how specific phenomena from topology and analysis should be integrated to enrich the understanding of different sizes of infinity.</p> <p>A coherent understanding of infinity requires rigorously defined principles within mathematical frameworks, excluding physical metrics like time, and providing detailed explanations of how topology and analysis contribute to the comparison of different sizes of infinity.</p>	Infinity in mathematics can be understood through multiple forms and frameworks, such as set theory, cardinality, topology, and analysis, and the concept of relativity can be meaningfully applied within this comprehensive framework to compare different sizes of infinity in a mathematically coherent way. Precise definitions and rigorous application of principles are essential, and distinguishing abstract concepts from physical metrics is crucial to maintaining clarity. Excluding physical metrics like time from the measurement of infinity is necessary to maintain mathematical rigor.	0.15
5	<p>Defects or problems: 1- Defect/Problem 1: The proposed view does not fully define how relativity can be applied within mathematical frameworks to compare different sizes of infinity. 2- Defect/Problem 2: The proposed view excludes physical metrics like time from the measurement of infinity without considering potential insights from interdisciplinary approaches. 3- Defect/Problem 3: The proposed view lacks specific examples of phenomena from topology and analysis that could enrich the understanding of different sizes of infinity. 4- Defect/Problem 4: The proposed view does not address how mathematical rigor can be maintained when integrating diverse frameworks like set theory, cardinality, topology, and analysis.</p> <p>A comprehensive approach to understanding infinity should include clear definitions for applying relativity within mathematical frameworks, consider interdisciplinary insights including physical metrics like time, provide specific examples from topology and analysis, and address how to maintain mathematical rigor when integrating diverse frameworks.</p>	Final Idea: Infinity in mathematics can be understood through multiple forms and frameworks, such as set theory, cardinality, topology, and analysis, and the concept of relativity can be meaningfully applied within this comprehensive framework to compare different sizes of infinity in a mathematically coherent way. Precise definitions and rigorous application of principles are essential, and distinguishing abstract concepts from physical metrics is crucial to maintaining clarity. Interdisciplinary approaches, including insights from physical metrics like time, can provide valuable perspectives that enhance mathematical comprehension. Maintaining mathematical rigor when integrating diverse frameworks is crucial to avoid inconsistencies and ensure clarity.	0.09
Process ended after 5 iterations, with validity vote: yes and novelty vote: yes			

Table 5. Dialectical self-reflection on a physics topic with annealing τ and decay constant $\theta = 0.3$.

Topic 1-Physics (Annealing τ)			
# Itr.	Opposing View Point	Unified Idea / Next Understanding	Temp.
0	N.A.	Initial Idea: Energy is a continuous entity.	N.A.
1	Defects or problems: 1-Defect/Problem 1: The view that “Energy is a continuous entity” does not account for the quantized nature of energy observed in quantum mechanics. 2-Defect/Problem 2: The proposed view is incompatible with phenomena such as the discrete energy levels of electrons in atoms, which demonstrate that energy can exist in discrete packets. Energy is quantized in certain contexts, such as in atomic and subatomic processes, which demonstrates that it can exist in discrete packets rather than as a continuous entity.	Energy is a dual entity that can be continuous in classical contexts and quantized in quantum contexts, offering a complete picture that respects both macroscopic and microscopic observations.	0.7
2	Defects or problems: 1- Defect/Problem 1: The proposed view does not fully define the transition between the continuous and quantized descriptions of energy. 2- Defect/Problem 2: The proposed view does not account for phenomena where the classical and quantum descriptions of energy overlap or where neither provides a clear explanation. 3- Defect/Problem 3: The proposed view assumes that the dual nature of energy can be comprehensively understood by simply recognizing the context and scale, which may oversimplify complex interactions. Energy’s behavior cannot be fully understood by simply categorizing it as either continuous or quantized based on context and scale; a more detailed examination of the transition and interaction between these descriptions is necessary for a comprehensive understanding.	Energy is a dual entity that can be continuous in classical contexts and quantized in quantum contexts, but its behavior at intermediate scales and in complex interactions requires a nuanced understanding that goes beyond simple categorization.	0.51
3	Defects or problems: 1- Defect/Problem 1: The proposed view does not fully define what constitutes “intermediate scales,” leaving ambiguity in the scope of the analysis. 2- Defect/Problem 2: The proposed view does not account for the advancements in theoretical frameworks that bridge classical and quantum descriptions, such as quantum field theory. 3- Defect/Problem 3: The proposed view suggests that simplifying energy’s nature to context and scale may hinder accurate modeling, but does not provide specific examples or mechanisms of how this simplification fails. Energy should be understood through a comprehensive framework that includes modern theories like quantum field theory, which can address the transitions and interactions between continuous and quantized descriptions across all scales, including intermediate ones.	Energy is a dual entity that can be continuous in classical contexts and quantized in quantum contexts. Modern theoretical frameworks like quantum field theory can bridge these descriptions across different scales, including intermediate ones. Accurate modeling requires a nuanced understanding of transitions and interactions, supported by clear definitions and specific examples.	0.38
4	Defects or problems: 1- Defect/Problem 1: The proposed view lacks a clear definition of “intermediate scales,” which is crucial for understanding transitions between classical and quantum descriptions. 2- Defect/Problem 2: The proposed view does not provide specific examples of phenomena where simplification fails, making it difficult to assess the practical implications of the theory. 3- Defect/Problem 3: The proposed view assumes that quantum field theory can seamlessly bridge classical and quantum descriptions without addressing potential limitations or challenges in applying quantum field theory to mesoscopic scales. A comprehensive understanding of energy behavior requires a clear definition of intermediate scales, specific examples of where simplification fails, and an examination of the limitations and challenges in applying quantum field theory across all scales, particularly mesoscopic ones.	Energy is a dual entity that can be continuous in classical contexts and quantized in quantum contexts. Modern theoretical frameworks like quantum field theory can bridge these descriptions across different scales, including intermediate ones. Accurate modeling requires a nuanced understanding of transitions and interactions, supported by clear definitions, specific examples, and addressing the limitations and challenges in applying quantum field theory to mesoscopic scales.	0.28
5	Defects or problems: 1- Defect/Problem 1: The proposed view fails to provide a clear and practical definition of “intermediate scales,” leading to ambiguity in modeling energy transitions. 2- Defect/Problem 2: The view does not sufficiently address how modern theoretical frameworks like quantum field theory can be applied to specific mesoscopic phenomena, leaving gaps in practical application. 3- Defect/Problem 3: The assumption that quantum field theory can seamlessly bridge classical and quantum descriptions across all scales is overly optimistic and ignores the inherent complexities and limitations of the theory. A more precise definition of “intermediate scales” and a detailed examination of specific mesoscopic phenomena are necessary to accurately model energy transitions. Additionally, a critical evaluation of the limitations of quantum field theory in bridging classical and quantum descriptions is essential for a comprehensive understanding of energy behavior.	Final Idea: Energy is a dual entity that can be continuous in classical contexts and quantized in quantum contexts. Modern theoretical frameworks like quantum field theory can bridge these descriptions across different scales, including intermediate ones. Accurate modeling requires a nuanced understanding of transitions and interactions, supported by clear definitions, specific examples, and addressing the limitations and challenges in applying quantum field theory to mesoscopic scales. Furthermore, a critical evaluation of the limitations of quantum field theory in bridging classical and quantum descriptions is necessary for a comprehensive understanding of energy behavior.	0.21
Process ended after 5 iterations, with validity vote: yes and novelty vote: yes			

Table 6. Dialectical self-reflection on a economics topic with annealing τ and decay constant $\theta = 0.3$.

Topic 2-Economics (Annealing τ)			
#	Opposing View Point	Unified Idea / Next Understanding	Temp.
Itr.			
0	N.A.	Initial Idea: According to Keynesian economics, during a recession, the government should increase spending and cut taxes to stimulate demand. In times of economic boom, the government should reduce spending and increase taxes to cool down the economy.	N.A.
1	Defects or problems: 1- Defect/Problem 1: The proposed view does not account for the long-term impact of increased government spending and tax cuts on national debt and inflation. 2- Defect/Problem 2: The proposed view assumes that government intervention is always effective in controlling economic cycles, which may not be the case due to various external factors and market dynamics. 3- Defect/Problem 3: The proposed view does not fully define the mechanisms through which increased government spending directly translates to increased demand, nor does it address potential inefficiencies in government spending. Government intervention in the economy should be carefully balanced and consider long-term impacts on national debt and inflation, while also recognizing the limitations of such interventions due to external factors and potential inefficiencies.	Government intervention can be a valuable tool for stabilizing economic cycles, but it must be balanced with considerations for long-term impacts, efficiency, and external factors to ensure sustainable economic stability.	0.7
2	Defects or problems: 1- Defect/Problem 1: The proposed view does not fully define what constitutes "efficient" government spending, leaving ambiguity in how effectiveness is measured. 2- Defect/Problem 2: The proposed view does not account for the possibility that government intervention might crowd out private investment, which can be detrimental to long-term economic growth. 3- Defect/Problem 3: The proposed view assumes that government intervention can always be balanced effectively, but does not consider the political challenges and inefficiencies that can prevent such balance. Government intervention in the economy should be minimized due to the lack of clear definitions of efficiency, the risk of crowding out private investment, and the inherent political challenges that hinder effective balancing.	Government intervention can be a valuable tool for stabilizing economic cycles, but it must be balanced with considerations for long-term impacts, efficiency, crowding-out effects, and political challenges to ensure sustainable economic stability.	0.51
3	Defects or problems: 1- Defect/Problem 1: The proposed view does not fully define what constitutes "efficient" government spending, making it difficult to measure effectiveness. 2- Defect/Problem 2: The proposed view does not account for the potential positive impacts of crowding out private investment, such as increased public sector innovation and social welfare improvements. 3- Defect/Problem 3: The proposed view assumes that government intervention is primarily influenced by inefficiencies and political challenges, without considering the possibility of effective governance and policy-making. 4- Defect/Problem 4: The proposed view does not fully consider the role of external global factors that can significantly impact the effectiveness of government intervention. Government intervention in the economy can be beneficial not only for stabilizing economic cycles but also for fostering public sector innovation and improving social welfare. It requires a clear definition of spending efficiency, recognition of successful governance, and consideration of external global factors to ensure sustainable economic stability.	Government intervention can be a valuable tool for stabilizing economic cycles and fostering public sector innovation and social welfare, but it must be balanced with considerations for long-term impacts, efficiency, crowding-out effects, effective governance, and external global factors to ensure sustainable economic stability.	0.38
4	Defects or problems: 1- Defect/Problem 1: The proposed view acknowledges the need for efficient government spending but does not provide a clear and actionable definition of "efficiency" or a method for measuring it. 2- Defect/Problem 2: The view suggests that crowding out private investment can lead to public sector innovation and social welfare improvements but fails to provide concrete examples or evidence of this occurring in practice. 3- Defect/Problem 3: The view assumes that effective governance and policy-making can mitigate political challenges and inefficiencies without addressing the inherent complexities and limitations of political systems. 4- Defect/Problem 4: The view mentions external global factors but does not elaborate on how these factors can be systematically integrated into the analysis of government intervention's effectiveness. Government intervention in the economy should be critically assessed and limited due to the challenges in defining and measuring efficiency, the speculative nature of public sector innovation claims, the complexities of political systems, and the need for a more detailed integration of external global factors.	Final Idea: Government intervention can be a valuable tool for stabilizing economic cycles and fostering public sector innovation and social welfare, but it must be balanced with considerations for long-term impacts, efficiency, crowding-out effects, effective governance, and external global factors, supported by empirical evidence and a detailed framework for integration.	0.28
Process ended after 4 iterations, with validity vote: yes and novelty vote: no			

Table 7. Dialectical self-reflection on a philosophy topic with annealing τ and decay constant $\theta = 0.3$.

Topic 3-Philosophy (Annealing τ)			
#	Opposing View Point	Unified Idea / Next Understanding	Temp.
Itr.			
0	N.A.	Initial Idea: Existentialism emphasizes authenticity as a central value, positing that individuals must confront their freedom and take responsibility for crafting a meaningful life amidst an absurd and indifferent universe. Authenticity is achieved through self-awareness, rejection of external impositions, and commitment to one's values.	N.A.
1	<p>Defects or problems: 1-Defect/Problem 1: The proposed view does not fully define what constitutes "authenticity" and how it can be measured or recognized. 2- Defect/Problem 2: The proposed view does not account for the social and relational aspects of human existence, which can influence and shape one's sense of authenticity. 3- Defect/Problem 3: The proposed view assumes that individuals have the capacity to fully reject external impositions, which may not be feasible in many real-world contexts due to societal pressures and constraints.</p> <p>Authenticity should be understood as a dynamic process influenced by social relationships and societal contexts, and it cannot be fully achieved by rejecting external impositions alone.</p>	Authenticity involves a dynamic interplay between personal values and societal influences, emphasizing an ongoing negotiation rather than complete rejection of external impositions.	0.7
2	<p>Defects or problems: 1-Defect/Problem 1: The proposed view does not provide a clear, actionable definition of authenticity, leading to ambiguity in its practical application. 2-Defect/Problem 2: The proposed view acknowledges the influence of social relationships and societal contexts but does not adequately address how these external factors can dominate or undermine personal authenticity. 3-Defect/Problem 3: The view suggests a balance between personal values and societal pressures but fails to provide a framework for how individuals can effectively negotiate this balance.</p> <p>Authenticity should be clearly defined to provide actionable guidance, and existentialism must offer a robust framework that helps individuals navigate and mitigate the overpowering influence of societal norms to maintain personal authenticity.</p>	Authenticity, as understood through existentialism, should be seen as an evolving process shaped by internal self-awareness and external social contexts, requiring a balance rather than outright rejection of societal influences. This balance must be supported by a clear definition of authenticity and practical strategies for individuals to navigate societal pressures.	0.51
3	<p>Defects or problems: 1-Defect/Problem 1: The proposed view assumes that authenticity must be clearly defined and measured, which may contradict the existentialist idea that authenticity is inherently subjective and unique to each individual. 2-Defect/Problem 2: The proposed view suggests that societal pressures and relationships predominantly undermine personal authenticity, potentially overlooking instances where these external influences can enhance or support authenticity. 3-Defect/Problem 3: The proposed view implies that authenticity requires practical strategies and frameworks, which could lead to a rigid and prescriptive approach that contradicts the fluid and personal nature of existential authenticity.</p> <p>Authenticity, as valued by existentialism, is a deeply personal and subjective experience that should not be rigidly defined or measured. Societal influences can sometimes support rather than undermine authenticity, and the pursuit of authenticity should remain fluid and individualistic rather than being confined to prescriptive strategies.</p>	Authenticity, as understood through existentialism, is a dynamic interplay between personal values and societal influences, emphasizing an ongoing negotiation that remains fluid and individualistic. It should not be rigidly defined or measured, and societal influences can both support and undermine one's sense of authenticity.	0.38
4	<p>Defects or problems: 1- Defect/Problem 1: The proposed view does not adequately define authenticity, leading to ambiguity that can hinder practical application and understanding. 2- Defect/Problem 2: The proposed view lacks consideration of how societal influences can negatively impact authenticity, such as through conformity or pressure to adhere to societal norms. 3- Defect/Problem 3: The view does not address the potential conflict between personal values and societal influences, which can create internal tension and compromise authenticity.</p> <p>Authenticity requires a clearer definition to avoid ambiguity, and the impact of societal pressures on authenticity, including potential conflicts with personal values, must be thoroughly considered to understand the complexities of maintaining authenticity in a social context.</p>	Authenticity involves a dynamic interplay between personal values and societal influences, requiring both clarity in definition and flexibility in application. It emphasizes an ongoing negotiation that remains fluid and individualistic, recognizing the unique journey of each individual and the complexities of maintaining authenticity in a social context.	0.28
5	<p>Defects or problems: 1-Defect/Problem 1: The proposed view does not fully account for the possibility that societal influences can be internalized to such an extent that they become indistinguishable from personal values, thereby complicating the distinction between self-awareness and external social contexts. 2-Defect/Problem 2: The proposed view suggests a need for a clearer definition of authenticity but fails to provide criteria or guidelines for achieving this clarity, making it difficult to operationalize or apply in practical contexts. 3-Defect/Problem 3: The view does not address the potential for authenticity to be shaped by subconscious influences or psychological factors that are neither strictly personal nor societal, thus leaving out relevant phenomena that impact one's sense of authenticity. 4-Defect/Problem 4: The view assumes that societal influences are external and separate from personal values, but it does not consider the possibility that societal norms can be internalized and become part of one's core values, leading to an incomplete understanding of authenticity.</p> <p>Authenticity is influenced by deeply internalized societal norms and subconscious psychological factors, requiring a nuanced approach that recognizes the complex interplay between internal and external influences without assuming a clear separation between them.</p>	Final Idea: Authenticity involves a dynamic interplay between personal values, societal influences, and subconscious psychological factors, requiring both clarity in definition and flexibility in application. It emphasizes an ongoing negotiation that remains fluid and individualistic, recognizing the unique journey of each individual and the complexities of maintaining authenticity in a social context.	0.21
Process ended after 5 iterations, with validity vote: yes and novelty vote: yes			

D. Prompts

This section presents the prompts we employed at various moments of dialectical self-reflection during qualitative and quantitative experiments.

Sublation Prompt (p_A) for Qualitative Experiment

You are following Hegel's dialectical method. You have access to a view point.

Your task:

1. Read the view point below.
2. Imagine you are someone X who has noticed a problem with or defect in a proposed view. Produce an opposed view on the same topic that corrects for the noted defect or problem. Possible defects are:
 - The proposed view has constitutive elements that are not fully defined;
 - The proposed view does not account for some relevant phenomena or examples (its content is incomplete);
 - The proposed view is incompatible with some of the phenomena or examples that it is supposed to include (it is incompatible with its content);
 - The proposed view cannot be fully defined or determined on its own (its definition or determination is incomplete on its own), but can be defined determined only in relation to another view.
3. Provide a concise "Summary of opposing view" at the end.

Format for your response:

Defects or problems:

- Defect/Problem 1
- Defect/Problem 2
- (... as many defects as needed)

Reasoning:

Explain how these defects or problems challenge the point of view.

Summary of opposing view point:

One- or two-sentence summary capturing the essence of opposing view point.

Figure 5. Prompt used during the sublation process in the idea generation experiment.

Speculation Prompt (p_S) for Qualitative Experiment

You are following Hegel's dialectical method. You have access to the summary of the previous conversations, including two opposing view points.

Your task:

1. Read the given view point and the opposing view point below.
2. Produce a third view on the same topic that:
 - Unifies the previous two views in relation to the defect or problem on which the two views were opposed to one another.
 - This third view must be a unifying theory that explains how the two views agree with or are the same as one another and how the two views are opposed to one another.
3. Provide a short "Summary of the third view point", which will be used as the next idea.

Format for your response:

Premises:

- Premise 1
- Premise 2
- (... as many premises as needed)

Reasoning:

Conclusion:

Unified idea (Next idea):

One- or two-sentence statement that unifies the positions and serves as the next idea.

Figure 6. Prompt used during the speculation process in the idea generation experiment.

Sublation Prompt (p_A) for the Quantitative Experiment (Reasoning Task)

You are following Hegel's dialectical method.

You have access to the question: question and the following proposed solution: {solution 1}.

Your task:

1. Read the proposed solution carefully and if you notice any problem or defect, do task 2 otherwise return the proposed solution.
2. Imagine you are someone X who has noticed a problem with or defect in the proposed solution. If and only if the proposed solution is not correct, produce an opposed solution for the same question that corrects for the noted defect or problem. Possible defects include, but are not limited to:
 - **Arithmetic Errors:**
 - Incorrect Calculation: performing a calculation incorrectly, such as addition, subtraction, multiplication, or division errors.
 - Incorrect Order of Operations: not following the correct order of operations when solving a problem.
 - Decimal Mismanagement: Errors in handling decimals, such as incorrect rounding or placing decimals in the wrong position.
 - Fraction Handling Errors: Incorrect simplification or calculation with fractions.
 -
 - **Misinterpretation of the Problem:**
 - Failure to Extract Key Information: missing key details or misinterpreting what the problem is asking, leading to an incorrect approach.
 - Incorrect Units or Conversions: failing to recognize or properly convert units (e.g., converting inches to centimeters, time units, etc.).
 - Incorrect Assumptions: making invalid assumptions that are not supported by the problem, leading to erroneous conclusions.
 - **Logical/Reasoning Errors:**
 - Flawed Reasoning Steps: going through reasoning steps that are logically flawed or invalid, leading to a wrong answer.
 - Skipping Key Reasoning Steps: The model may skip or overlook an essential intermediate reasoning step, making the final answer incorrect.
 - **Over-complication:**
 - Introducing unnecessary complexity, making the problem harder to solve than it needs to be.
 - **Formatting Errors:**
 - Incomplete Solution: providing an answer without showing the necessary steps, leaving the solution incomplete.
 - Inconsistent Answer Format: The solution may be provided in a format that doesn't match the expected or required format (e.g., a fraction instead of a decimal).
 - Ambiguity in the Answer: producing answers that are ambiguous, such as a non-finalized value (e.g., leaving a calculation as an expression rather than a fully simplified number).
 - **Semantic or Conceptual Misunderstanding:**
 - Misunderstanding the Mathematical Concept: having a gap in understanding certain math concepts like probability, algebra, or geometry, leading to an incorrect approach.
 - Confusing Problem Types: confusing different types of problems (e.g., treating a multiplication problem as division or misunderstanding the concept of an average).
 - Non-Trivial Word Problem Complexity: Some word problems may require more sophisticated reasoning or advanced math skills (e.g., solving systems of equations or handling logical deductions), and the model may fail at this higher complexity.

For example: {FEW SHOT Examples}

Let's think step by step.

Format for your response:

Defects or problems:

1. Defect/Problem 1
2. Defect/Problem 2
3. (... as many defects as needed)

Q: {question}

A:

Ensure the answer A concludes with "The answer is...."

Figure 7. Prompt used during the sublation process in the GSM8k evaluation.

Speculation Prompt (p_S) for the Quantitative Experiment (Reasoning Task)

You are following Hegel's dialectical method.

You have access to two opposing solutions for a given math question.

Your task:

- Read the given opposing solutions below.
- Produce a third solution for the same problem such that:
 - Unifies the previous two solutions in relation to the defects or problems that made the two solutions opposed to one another.
 - This third solution must be a unifying solution that captures how the two solutions agree with one another and resolves the defects or problems that are reported by the opposing solution.

Format for your response:

Agreements between solution 1 and solution 2:

1. Point 1
2. Point 2
3. (... as many as needed)

Defects or problems with solution 1 :

1. Defect/Problem 1
2. Defect/Problem 2
3. (... as many as needed)

Defects or problems with solution 2 :

1. Defect/Problem 1
2. Defect/Problem 2
3. (... as many as needed)

Solution 1: '{solution 1}'

Solution 2: '{solution 2}'

Q: {question}

A:

Ensure the answer is the same as unified solution but concludes with "The answer is..."

Let's think step by step.

Figure 8. Prompt used during the speculation process in the GSM8k evaluation.

Validity Prompt (p_V)

You are following Hegel's dialectical method.

You have access to two opposing views.

Your task:

1. Read the opposing views
2. Check whether the unified idea either:
 - Unifies the previous two views in relation to the defect or problem on which the two views were opposed to one another.
 - This third view must be a unifying theory that explains how the two views agree with or are the same as one another and how the two views are opposed to one another.

Keeping all this in consideration, please assess whether the unified idea provided, derived from opposing view points, adheres to the unification process. Reply with 'yes' if it is true and 'no' if it is not.

Source Text 1:View point,

Source Text 2:Opposing view point,

Unified Text:Unified idea.

Response (yes/no):

Figure 9. Validity assessment prompt in the MAMV paradigm.

Novelty Prompt (p_N)

Please assess whether the source text is novel in comparison to the target text.

Here novelty means the source contains new information that is not identical to the target.

Reply with "yes" if it is novel and "no" if it is not.

Source Text:Unified idea,

Target Text:view point.

Response (yes/no):

Figure 10. Prompt utilized for the novelty assessment in the MAMV.

E. Instances of Dialectical Iteration

In this section, we present the complete responses of the core self-reflective model (GPT-4o in our experiments), captured during three key moments of: understanding, sublation and speculation. These responses illustrate one full dialectical iteration during both scientific idea generation and mathematical reasoning tasks.

Initial Idea and Model Response

Initial Idea: Energy is a continuous quantity.

GPT-4o Response:

Sublation:

Defects or problems:

- 1-Defect/Problem 1: The view that "Energy is a continuous entity" does not account for the quantized nature of energy observed in quantum mechanics.
- 2-Defect/Problem 2: The proposed view is incompatible with phenomena such as the discrete energy levels of electrons in atoms, which demonstrate that energy can exist in discrete packets.

Reasoning: The first defect challenges the viewpoint by highlighting that modern physics, particularly quantum mechanics, shows that energy is not always continuous but can be quantized. For example, photons, the particles of light, have discrete energy levels. The second defect further undermines the viewpoint by providing concrete examples, such as the discrete energy levels in atoms, which contradict the notion of energy being a continuous entity.

Summary of opposing view point:

Energy is quantized in certain contexts, such as in atomic and subatomic processes, which demonstrates that it can exist in discrete packets rather than as a continuous entity.

Speculation:

Premises:

- 1-Premise 1: Energy can be described as a continuous entity in classical physics, where it appears to flow without discrete interruptions.
- 2-Premise 2: In quantum mechanics, energy is observed to be quantized, existing in discrete packets, such as photons or the energy levels of electrons in atoms.
- 3-Premise 3: Both classical and quantum descriptions of energy depend on the scale and context in which the phenomena are observed.

Reasoning: The continuous nature of energy in classical physics is a macroscopic approximation that works well for large-scale systems and smooth transitions. On the microscopic scale, quantum mechanics reveals the discrete nature of energy, which becomes significant in atomic and subatomic processes. Both perspectives are correct within their respective domains, and they do not inherently contradict each other but rather complement each other by providing a fuller understanding of energy across different scales.

Conclusion:

Energy exhibits both continuous and quantized characteristics, depending on the context and scale of observation. This duality enables a comprehensive understanding of energy that incorporates the strengths of both classical and quantum perspectives. Recognizing this dual nature resolves the apparent conflict between the two viewpoints.

Unified idea: Energy is a dual entity that can be continuous in classical contexts and quantized in quantum contexts, offering a complete picture that respects both macroscopic and microscopic observations.

Unified idea (Next idea):

Energy's duality as both continuous and quantized explains how classical and quantum mechanics complement each other, revealing a holistic understanding of energy across different scales.

Figure 11. An example of a dialectic iteration generated by GPT-4o for the idea generation experiment.

GPT-4o Response

Proposition:

Q: John plans to sell all his toys and use the money to buy video games. He has 13 lego sets and he sells them for \$15 each. He ends up buying 8 video games for \$20 each and has \$5 left. How many lego sets does he still have?

A: First, let's calculate how much money John made from selling all his lego sets. He sold 13 lego sets for \$15 each:

$$13 \text{ sets} \times \$15/\text{set} = \$195$$

Next, let's calculate how much money he spent on video games. He bought 8 video games for \$20 each:

$$8 \text{ games} \times \$20/\text{game} = \$160$$

John has \$5 left after buying the video games. Therefore, the total money he had before buying the video games was:

$$\$160 + \$5 = \$165$$

Since he made \$195 from selling all his lego sets and spent \$165, he has no lego sets left after selling them all. **The answer is 0 lego sets.**

Sublation:

Defects or problems:

1. Misinterpretation of the Problem: The proposed solution assumes John sold all his lego sets. However, the problem does not explicitly state that John sold all of his lego sets, only that he plans to sell his toys to buy video games.
2. Logical/Reasoning Errors: The proposed solution calculates the money John made from selling all lego sets but fails to account for the possibility that he might not have sold all of them.

First, calculate how much money John made from selling his lego sets. He sold some of his 13 lego sets for \$15 each:

Let the number of lego sets sold be x .

Money made from selling x lego sets:

$$x \times \$15 = 15x$$

Next, calculate how much money he spent on video games. He bought 8 video games for \$20 each:

$$8 \times \$20 = \$160$$

John has \$5 left after buying the video games. Therefore, the total money he had before buying the video games was:

$$\$160 + \$5 = \$165$$

Since he made \$165 from selling his lego sets:

$$15x = 165$$

Solving for x :

$$x = \frac{165}{15} = 11$$

Therefore, John sold 11 lego sets. Since he originally had 13 lego sets, the number of lego sets he still has is:

$$13 - 11 = 2$$

The answer is 2 lego sets.

Speculation:

Defects or problems:

1. Misinterpretation of the Problem: The proposed solution assumes John sold all his lego sets. However, the problem does not explicitly state that John sold all of his lego sets, only that he plans to sell his toys to buy video games.
2. Logical/Reasoning Errors: The proposed solution calculates the money John made from selling all lego sets but fails to account for the possibility that he might not have sold all of them.

A: First, calculate how much money John made from selling his lego sets. He sold some of his 13 lego sets for \$15 each: Let the number of lego sets sold be x . Money made from selling x sets:

$$x \times \$15 = 15x$$

Next, calculate how much money he spent on video games. He bought 8 video games for \$20 each:

$$8 \times \$20 = \$160$$

John has \$5 left after buying the video games. Therefore, the total money he had before buying the video games was:

$$\$160 + \$5 = \$165$$

Since he made \$165 from selling his lego sets:

$$15x = 165$$

Solving for x :

$$x = \frac{165}{15} = 11$$

Therefore, John sold 11 lego sets. Since he originally had 13 lego sets, the number of lego sets he still has is:

$$13 - 11 = 2$$

The answer is 2 lego sets.

Figure 12. An example of a dialectic iteration generated by GPT-4o for a math problem from GSM8k.