

# Subjective Topic meets LLMs: Unleashing Comprehensive, Reflective and Creative Thinking through the Negation of Negation

Fangrui Lv<sup>1</sup>, Kaixiong Gong<sup>2</sup>, Jian Liang<sup>3</sup>, Xinyu Pang<sup>1</sup>, Changshui Zhang<sup>1,\*</sup>

<sup>1</sup> Institute for Artificial Intelligence, Tsinghua University (THUAI)  
Beijing National Research Center for Information Science and Technology (BNRist)  
Department of Automation, Tsinghua University Beijing, P.R.China

<sup>2</sup> The Chinese University of Hong Kong <sup>3</sup> Kuaishou Technology  
{lvfr23, pangxy22}@mails.tsinghua.edu.cn, kxgong@link.cuhk.edu.hk,  
liangjian03@kuaishou.com, zcs@mail.tsinghua.edu.cn

## Abstract

Large language models (LLMs) exhibit powerful reasoning capacity, as evidenced by prior studies focusing on objective topics that with unique standard answer such as arithmetic and commonsense reasoning. However, the reasoning to definite answers emphasizes more on *logical thinking*, and falls short in effectively reflecting the *comprehensive, reflective, and creative thinking* that is also critical for the overall reasoning prowess of LLMs. In light of this, we build a dataset SJTP comprising diverse SubJective ToPics with free responses, as well as three evaluation indicators to fully explore LLM’s reasoning ability. We observe that a sole emphasis on logical thinking falls short in effectively tackling subjective challenges. Therefore, we introduce a framework grounded in the principle of the **Negation of Negation (NeoN)** to unleash the potential comprehensive, reflective, and creative thinking abilities of LLMs. Comprehensive experiments on SJTP demonstrate the efficacy of NeoN, and the enhanced performance on various objective reasoning tasks unequivocally underscores the benefits of stimulating LLM’s subjective thinking in augmenting overall reasoning capabilities.

## 1 Introduction

Large language models (LLMs) have achieved remarkable performance in recent years (OpenAI, 2022, 2023; Touvron et al., 2023; Jiang et al., 2023) and have displayed formidable reasoning ability that validated on various tasks, including arithmetic reasoning, symbolic reasoning, commonsense reasoning (Luo et al., 2023; Wei et al., 2022a; Geva et al., 2021; Talmor et al., 2019), etc.

Despite the prominent reasoning capabilities, their evaluation still lacks comprehensiveness. Previous works primarily investigate LLMs based on

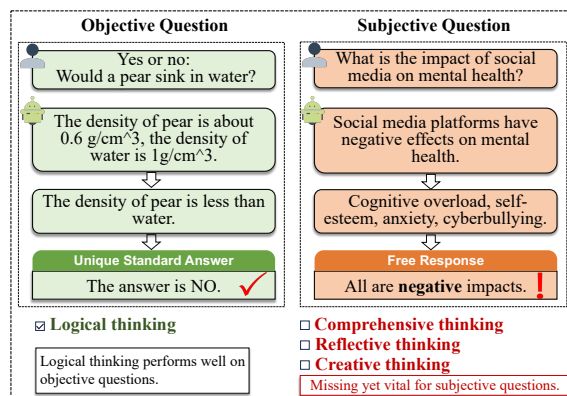


Figure 1: LLMs with CoT prompting (Wei et al., 2022a) showcase strong logical thinking ability, which is adequate to solve objective questions. Meanwhile, it fails to solve subjective questions and shows insufficiency of comprehensive, reflective and creative thinking abilities.

objective topics with unique standard answers and the logical reasoning path is clear-out (e.g. “3-2=1”, “Cat is herbivorous?”) (Mao et al., 2023). It is evident that reasoning on these topics heavily relies on logical thinking. The success of Chain-of-Thought (CoT) (Wei et al., 2022a) serves as compelling evidence, as it elicits logical thinking through prompts with a series of reasoning steps, instantly leading to significant improvement in solving objective questions. However, the ideal reasoning ability transcends mere logical thinking, comprehensive, reflective, and creative thinking are indispensable when tackling complex tasks, which cannot be well reflected on objective topics, as shown in Figure 1. Consequently, the lack and difficulty in evaluating these abilities pose challenges for approaching human-like thinking ability of current LLMs.

To enable analysis of LLMs in terms of comprehensiveness, reflection, and creativity and further enhance the overall reasoning ability, we first develop a benchmark SJTP that consists of SubJective ToPics with free responses. Respondents need to engage in open-ended discussions

\* Corresponding author.

Topic Type	Example	SCR
1.Viewpoint Discourse (comprehensive views)	What do you think about the impact of social media on mental health? (①comprehensiveness: ★★★, ②reflection: **, ③creativity:**) )	76.5
2.Binary Dialectics (in-depth analysis)	Should school wear uniforms? (①comprehensiveness: **, ②reflection: ★★★, ③creativity:**) )	69.9
3.Practical Analysis (constructive opinions)	How can the preservation of cultural heritage contribute to the promotion of social cohesion? (①comprehensiveness: **, ②reflection: **, ③creativity:*** )	80.2
<b>Topic Field</b>	① Social and Ethics, ② History and Civilization, ③ Literature and Arts, ④ Technology and Education, ⑤ Environment and Health, ⑥ Economy and Politics, ⑦ Law and Human Rights, ⑧ Psychology and Emotions	
<b>Evaluation</b>	[Basic]: ① Clarity of Viewpoint, ② Logicity of Argumentation, ③ Correctness and Fidelity	
<b>Dimensions</b>	[Advanced]: ④ Comprehensiveness&Diversity, ⑤ Depth&Essentiality, ⑥ Innovation&Uniqueness	

Table 1: Topic types, topic fields and evaluation dimensions for the construction and evaluation of SJTP dataset. The value of *SCR* is the result of GPT-3.5 armed with CoT (average of the three evaluation indicators).

based on their accumulated experience, perspectives, and inspiration, which can better reflect the aforementioned thinking abilities. For thorough investigation, we set up the three most common types of subjective topics in SJTP, each of which focuses more on reflecting one of the above abilities. All three types of subjective questions involve knowledge from 8 fields to enable diversity, as shown in Table 1. Next, to assess the quality of free responses, we consider six dimensions including the basic properties of text generation and the advanced characteristics to be explored, as listed in Table 1. On these dimensions, we design three evaluation indicators for free responses and briefly evaluate gpt-3.5 armed with the promising CoT prompts, which show plain performance. As shown in Figure 1, the reasoning pathway does not deliberately consider the full-sided perspectives and excavate the intrinsic causes or potential solutions about "the impact of social media on mental health", resulting in a lack of depth and unbalanced analysis.

To elicit the LLMs’ comprehensive, reflective, and creative thinking abilities, we propose a structured framework NeoN inspired by the principle of the **N**egation **O**f **N**egation, which is a philosophical principle proposed by Engels (Engels et al., 1954). The core insight is that *the development and completion of things must go through negation and transcendence of themselves*, which also aligns with Hegel’s "abstraction-negation-concreteness" principle (Pinkard, 1988). Through constant negation, we can emphasize the multifaceted and complex nature of problems, and break established thinking patterns, thereby achieving refinement to the defects in initial response. Please note that if the current response is impeccable, further negation will be futile and lead to repetitive output, as negation

should be performed under the correct premise.

Specifically, the proposed *NeoN* comprises three stages: direct answer, iterative negation, and unification reasoning. We first allow the LLM to generate answers directly with strict logical thinking to guarantee fundamental reasoning ability. Then we cast the LLM as a negator instead of a logical reasoner. It will construct a negation link to the previous responses under the correct premise, and then determine whether to proceed with negation based on the similarity between the current and previous responses. Therefore, by negating existing responses from any potential aspects, the LLM will be encouraged to explore unconsidered perspectives, excavate profound insights, and inspire innovative ideas, which supplements and surpasses the previous responses in a spiral upward manner. Finally, we enable the LLM to generate a refinement response by considering all these answers. Note that the whole workflow is built on top of unified zero-shot prompting, without carefully designed examples that are specific to different tasks, making it highly convenient and versatile.

We employ both API-based and open-source LLMs including GPT (OpenAI, 2023), ChatGPT (OpenAI, 2022), LLaMA (Touvron et al., 2023) and Mistral (Jiang et al., 2024), to validate the efficacy of our framework. Experimental results show that *NeoN* leads to significant and consistent improvements on both subjective and objective topics, underscoring the necessity and effectiveness of unleashing the comprehensive, reflective and creative thinking for better reasoning ability.

## 2 Subjective Topic Dataset

We construct a SJTP benchmark to explore the comprehensiveness, reflection, and creativity of

LLMs, which includes diverse subjective topics along with reasonable scoring points and solutions. Besides, we elaborately develop three evaluation indicators to assess the quality of free responses. The overall illustration is shown in Fig. 2.

## 2.1 Data Generation

**Topic Pool Construction.** We first build a subjective topic pool  $T_{\text{pool}} = \{\mathcal{T}, \mathcal{F}\}$  that covers three topic types  $\mathcal{T} = \{t_i\}_{i=1}^3$  and eight topic fields  $\mathcal{F} = \{f_i\}_{i=1}^8$  that be widely discussed, as shown in Table 1. Specifically, subjective questions used to assess the knowledge breadth, critical and innovative thinking of respondents can mostly be grouped into three representative types:  $\{\textit{Viewpoint Discourse}, \textit{Binary Dialectics}, \textit{Practice Issues}\}$ , and each type particularly emphasizes more on one of the abilities. For instance, tackling *Practical Issues* needs constructive opinions and thus more reflective of creative thinking. Besides, we involve a broad scope of knowledge in eight topic fields such as social and technology. Each field contains subdivided related themes to guarantee the diversity of data generation, see Appendix A.1 for details.

**Topic Sampling.** Given topic pool  $T_{\text{pool}}$ , for each data, we randomly sample a topic type  $t_i \in \mathcal{T}$ , 1 or 2 themes that with correlation in fields  $f_i$  or  $f_j, f_j \in \mathcal{F}$ , as seeds  $s$  for question generation.

**Question Generation.** Given the generation seeds  $s$ , we enable the LLM  $\mathcal{M}$  to generate a specific subjective question  $Q$  by incorporating  $s$  into the prompt  $p_{\text{gen}}$ , i.e.,  $Q = \mathcal{M}(p_{\text{gen}}|s)$ . The generation process is conducted in a zero-shot manner, we do not include any exemplars or other manual interventions to avoid potential biases. This generation mechanism challenges the model to generate problems that join diverse topics while keeping the problems reasonable. We further conduct manual checks after generating all the questions, filtering out low-quality, meaningless, or redundant ones to ensure the quality of SJTP. The distribution of topic fields in SJTP is provided in Appendix. 4.3.

**Solution Generation.** First, we need to set standards for perfect answers to provide qualitative free solutions. A reasonable solution must meet three basic requirements, including clarity, logicity and fidelity, which are indispensable for a qualified response. Moreover, considering the abilities tested by subjective questions, the solution needs to possess corresponding advanced properties, i.e., comprehensiveness, reflection and creativity. Therefore, we define six evaluation di-

mensions denoted as  $Eval_{dim}$  and the detailed requirements are provided in Table 10. Then, given  $Eval_{dim}$  and a question  $Q$ , we generate specific scoring points by prompting the LLM:  $Sol_{point} = \mathcal{M}(Q, Eval_{dim})$ . By adhering to the criteria outlined by each evaluation dimension and considering the specific knowledge in question, the LLM will generate organized and reasonable scoring points that maintain both specificity and quality. Finally, we generate a complete solution,  $Sol_{comp} = \mathcal{M}(Q, Sol_{point}, Eval_{dim})$ , that is comprehensive, profound, and innovative to serve as a more intuitive reference. We use GPT-4 (OpenAI, 2023) as the LLM for the whole data generation process, see prompt templates in Appendix A.3 and data examples in Appendix A.4.

## 2.2 Evaluation

Considering the inherent challenge in quantifying the quality of free responses to subjective questions, we carefully craft three automated evaluation indicators. (1) We first devise a general metric,  $SCR_{dim}$ , which assesses responses according to detailed requirements across the three basic and three advanced evaluation dimensions, respectively. It can serve as a general indicator of the fundamental response quality and the explored subjective abilities. (2) To better integrate specific questions, we further develop  $SCR_{point}$  based on  $SCR_{dim}$ . Specifically, we generate specific scoring points for each question according to the requirements of the six dimensions. This metric aligns more closely with the specific requirements of each question and can serve as a more precise checklist. (3) Lastly, to provide a more intuitive and holistic reference, we generate a complete reference solution for each question, drawing upon the six general dimensions and specific scoring points. Subsequently, we design  $SCR_{sol}$  to measure the semantic similarity between the response and the reference solution.

## 3 NeoN Framework

To unleash the comprehensive, reflective, and creative cognitive faculties of LLMs, we introduce the *NeoN* framework grounded on the principle of the negation of negation, allowing the model to continuously improve and transcend itself. The overall framework is schematically illustrated in Figure 2.

Firstly, *NeoN* generates an initial solution leveraging LLM’s inherent reasoning ability. Then, it embarks on a process of negation, critiquing its

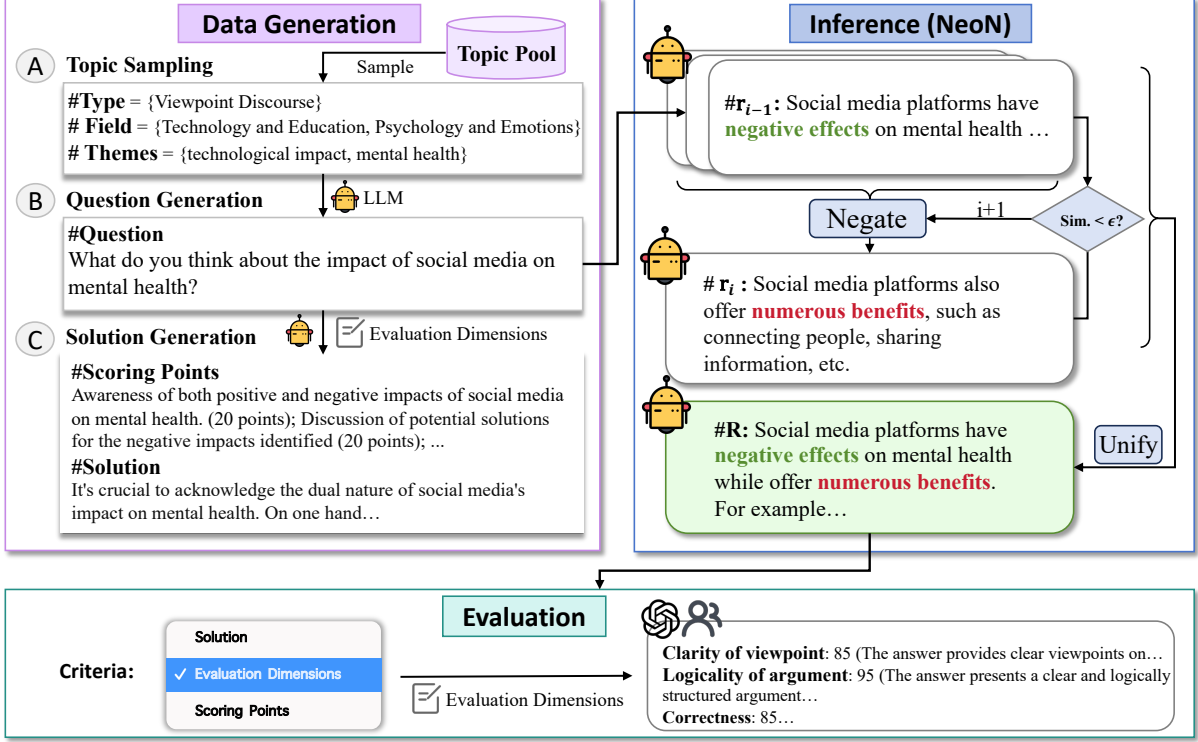


Figure 2: Illustration of the construction of SJTP dataset, the framework of NeoN, and the evaluation indicators.

earlier responses from any potential aspects while keeping a commitment to correctness and fidelity. This stage is designed to encourage the model to explore broader perspectives, engage in thorough analysis, and break away from established concepts. Finally, it obtains a refined response by methodically assimilating and integrating the preceding responses. Next, we describe NeoN in more detail. **Step 1: Direct Response.** Given a question  $Q$ , a LLM  $\mathcal{M}$ , NeoN first generates a reasonable response  $r_0$  directly:

$$r_0 = \mathcal{M}(Q \oplus \mathcal{P}_1), \quad (1)$$

where  $\oplus$  denotes concatenation operation.  $\mathcal{P}_1$  is a direct prompt, e.g., "Let's generate the answer".

**Step 2: Negation of Negation.** Then, inspiring by the principle of the negation of negation, i.e., things develop and progress in constant negation, NeoN constantly negates the previous responses from any potential aspects, thereby facilitating the generation of novel and advanced insights:

$$r_n = \mathcal{M}(Q \oplus r_0 \oplus \dots \oplus r_{n-1} \oplus \mathcal{P}_2), \quad (2)$$

where  $n$  denotes the number of negation rounds. We terminate the negation process when the semantic similarity between the current response  $r_n$  and previous responses  $r_0 \oplus \dots \oplus r_{n-1}$  exceeds a

threshold  $\epsilon$ , which implies that  $r_n$  is approaching a state of refinement since  $\mathcal{M}$ 's diminishing capacity to yield additional novel insights. The value of  $n$  is usually between 2~3 empirically.  $\mathcal{P}_2$  is a prompt for making reasonable negation, e.g., "Negate the above responses to deduce a more perfect answer." **Step 3: Integration and Unification.** Finally, NeoN takes question  $Q$  and all the responses as the input to generate the final response  $\mathcal{R}$ :

$$\mathcal{R} = \mathcal{M}(Q \oplus r_0 \oplus \dots \oplus r_n \oplus \mathcal{P}_3), \quad (3)$$

where  $\mathcal{P}_3$  is the last prompt leading to the final answer which can be set as "Based on all the previous answers, generate a perfect answer."

Note that NeoN utilizes the same LLM in the whole framework with task-invariant prompts in a zero-shot manner to avoid potential limitation, possessing high convenience and versatility. Case studies of NeoN provided in Appendix B.2 show its efficacy in exploring new perspectives and insights, as well as correcting erroneous reasoning paths.

**Discussion 1: Rationality of Negation.** The intuition of NeoN aligns with Hegel's philosophical principle that "A truth with concreteness, comprehensiveness, and absoluteness must always go through the stage of negation in the process of completion. When a statement is sufficiently correct

Method	Viewpoint Discourse			Binary Dialectics			Practical Analysis			Avg
	$SCR_{sol}$ (Acc.)	$SCR_{point}$ (Acc.)	$SCR_{dim}$ (Acc.)	$SCR_{sol}$ (Acc.)	$SCR_{point}$ (Acc.)	$SCR_{dim}$ (Acc.)	$SCR_{sol}$ (Acc.)	$SCR_{point}$ (Acc.)	$SCR_{dim}$ (Acc.)	
(llama-2-70b)										
Direct Prompt	61.25	64.72	78.80	80.52	44.62	77.30	76.08	70.82	81.50	70.64
Zero-Shot-CoT	66.70	67.02	77.63	79.30	50.27	74.38	78.82	68.20	80.42	70.20
Self-Consistency	67.20	67.74	<u>78.96</u>	81.76	50.82	75.72	<u>80.25</u>	<u>71.14</u>	<u>81.35</u>	72.74
Self-refine	<u>68.17</u>	<u>68.94</u>	78.52	<u>82.30</u>	<u>52.64</u>	78.09	79.70	70.62	81.25	<u>73.36</u>
<b>NeoN (Ours)</b>	<b>71.33</b>	<b>70.25</b>	<b>82.70</b>	<b>84.14</b>	<b>55.10</b>	<b>81.22</b>	<b>82.15</b>	<b>72.53</b>	<b>83.20</b>	<b>75.85</b>
(Mistral-8x7b)										
Direct Prompt	66.21	67.72	82.90	85.67	47.10	79.05	82.30	70.94	83.15	73.89
Zero-Shot-CoT	70.95	71.82	82.76	87.12	53.81	74.20	83.26	68.72	82.06	74.98
Self-Consistency	72.11	73.06	83.68	89.44	57.60	79.33	<u>86.46</u>	71.20	<u>83.29</u>	<u>77.35</u>
Self-refine	<u>75.68</u>	<u>76.04</u>	<u>86.10</u>	<u>90.89</u>	<u>60.57</u>	<u>81.76</u>	85.20	<u>72.15</u>	83.22	79.06
<b>NeoN (Ours)</b>	<b>81.76</b>	<b>79.53</b>	<b>87.62</b>	<b>91.97</b>	<b>63.40</b>	<b>82.15</b>	<b>87.42</b>	<b>75.33</b>	<b>86.50</b>	<b>81.74</b>
(gpt-3.5-turbo)										
Direct Prompt	65.72	69.21	84.83	85.66	43.24	80.69	87.92	72.50	84.63	74.93
Zero-Shot-CoT	71.86	72.40	85.26	88.11	48.35	73.40	88.24	70.53	82.89	75.81
Self-Consistency	73.26	74.15	85.47	<u>90.20</u>	51.50	78.85	<u>89.23</u>	<u>73.62</u>	<u>84.18</u>	77.72
Self-refine	<u>76.40</u>	<u>75.91</u>	<u>87.22</u>	89.51	<u>54.78</u>	<u>81.30</u>	88.70	72.59	83.81	<u>78.91</u>
<b>NeoN (Ours)</b>	<b>80.40</b>	<b>81.82</b>	<b>88.67</b>	<b>92.50</b>	<b>60.33</b>	<b>83.21</b>	<b>89.15</b>	<b>76.83</b>	<b>87.74</b>	<b>82.29</b>

Table 2: Main results of methods on SJTP. The best result is **in bold** and the second-best is underlined.

and consummate, further negation becomes futile, and reasoning is equivalent to logical reasoning.” Consequently, negating flawed responses encourages the model to explore new perspectives, delve deeper into the essence, and challenge established opinions, thereby fostering the refinement of responses. Whereas, the negation of a correct answer will lead to a high consistency with preceding responses, as negation is required to be conducted under the correct premise. Such a case can serve as both a termination signal for the negation process and a reinforcement of the reliability of the "correct answer" that is immune to negation. Table 5 shows a relatively high probability of NeoN correcting the initially wrong answer and an extremely low probability of misleading the initially correct answer, further verifying our rationality.

**Discussion 2: Superiority of Negation.** Compared to self-reflection and "tit for tat" debate, which may limit aspects of refinement or insist on errors due to their feedback quality or agent stance, negation allows the model to explore any potential flaws to the fullest extent. This is because negation resembles adversarial learning between infinite parties, making it more conducive to rectifying errors, omissions, superficiality, stereotypes, etc. As shown in Table 5, 8, NeoN obtains better refinement with higher efficiency. Besides, NeoN employs task-invariant zero-shot prompts, eliminating the need for meticulously crafted task-specific exemplars, possessing more convenience and versatility. Detailed discussions are provided in Sec. 5.

## 4 Experiment

### 4.1 Setups

**Base LLMs.** We conduct experiments on both API-based models including ChatGPT (gpt-3.5-turbo), GPT-4 (OpenAI, 2023), and open-source model LLaMA-2-70B (Touvron et al., 2023), Mistral-8x7B (Jiang et al., 2024). We set the sampling temperature as 0.7. For conducting NeoN, we use the same LLM in the whole framework with unified prompts across all tasks in a zero-shot manner. For evaluation, we take the sota GPT-4 as the evaluator.

**Baselines.** We compare NeoN with 1) Zero-Shot-CoT (Kojima et al., 2022) to verify the necessity of unleashing subjective thinking abilities; 2) Self-consistency (Wang et al., 2022) to verify the superiority of negation over simply generating multiple outputs; 3) Self-refine (Madaan et al., 2023) to validate the advantage of negation over self-reflection; 4) MAD (Liang et al., 2023) to show the predominance of negation over multiple-agent debate. For fair comparison, we use the same prompts from these works when available, otherwise, we create prompts consistent with NeoN in non-key parts.

### 4.2 Main Results

**NeoN consistently improves over base LLMs on subjective task** in Table 2. Results with standard deviation are provided in Appendix A.2. Direct

		Arithmetic			Commonsense		Generic	
		GSM8K	SVAMP	AddSub	CSQA	StrategyQA	Date Understand	Shuffled Objects
GPT-3.5	Direct Prompt	17.31	70.79	86.91	72.25	62.33	46.83	32.98
	Zero-Shot-CoT	80.15	80.38	88.19	70.80	62.08	64.44	69.66
	Self-Consistency	<u>83.89</u>	<u>83.60</u>	89.93	73.16	63.76	73.27	<u>72.08</u>
	Self-refine	81.64	80.73	89.47	73.41	63.25	<u>74.42</u>	71.26
	MAD	82.76	83.49	90.22	74.10	63.94	71.25	70.63
	<b>NeoN(Ours)</b>	<b>84.17</b>	<b>85.67</b>	<b>93.09</b>	<b>75.22</b>	<b>64.83</b>	<b>77.46</b>	<b>74.13</b>
GPT-4	Direct Prompt	36.33	87.67	96.25	83.67	67.33	74.67	39.25
	Zero-Shot-CoT	92.59	91.96	96.54	84.36	72.67	86.07	79.26
	Self-Consistency	93.44	<u>92.43</u>	97.68	85.41	77.50	90.19	84.41
	Self-refine	92.67	91.84	97.20	<u>86.67</u>	<u>79.60</u>	<u>90.33</u>	<u>86.90</u>
	MAD	<u>93.68</u>	92.12	<u>98.09</u>	86.12	78.15	88.24	85.66
	<b>NeoN(Ours)</b>	<b>94.17</b>	<b>93.45</b>	<b>98.87</b>	<b>87.20</b>	<b>81.42</b>	<b>91.11</b>	<b>88.26</b>

Table 3: Main results of baseline methods and NeoN on seven objective datasets.

Prompt responds to questions directly and shows relatively limited performance for subjective topics. By explicitly prompting LLMs with “Let’s think step-by-step”, Zero-Shot-CoT surpasses in some cases thanks to their clear logical thinking but sometimes degrades due to their limitation on divergent thinking. Self-consistency generates multiple outputs independently via CoT prompt and makes a synthesis that benefits comprehensiveness. Nevertheless, it is inferior to Self-refine due to lack of rethinking. *NeoN* achieves the best across all the models and topic types, which demonstrates the superiority of negation in refining free responses of subjective tasks. Note that the consistency inner the three indicators further verifies their rationality.

**NeoN consistently enhances comprehensive, reflective and creative thinking abilities** in Fig. 3. We calculate and visualize the  $SCR_{dim}$  score of methods in each evaluation dimension to further investigate the effectiveness of negation. It is clear that NeoN surpasses all the baselines in the three advanced subjective properties by a large margin while maintaining the quality of the three basic properties. Besides, NeoN’s transcendence of Self-refine in terms of depth and innovation indicates that although self-reflection helps supplement overlooked perspectives, it cannot effectively delve deeper into the essence and stimulate innovation.

**NeoN consistently improves over objective tasks** in Table 3. To substantiate the significance of unleashing comprehensive, reflective and creative thinking for enhancing the overall reasoning performance of LLMs, we further evaluate our NeoN on objective tasks including arithmetic reasoning, commonsense reasoning and generic reasoning. Detailed descriptions of datasets are provided in Appendix B.1. We observe that Self-refine demon-

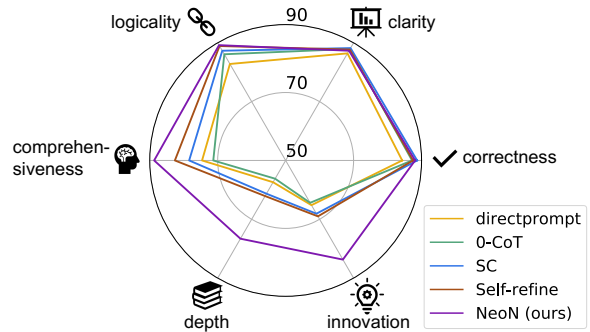


Figure 3: Categorized performance analysis for scores in different evaluation dimensions and topic types.

Method	$SCR_{Sol}$	$SCR_{point}$	$SCR_{dim}$
<b>NeoN (ours)</b>	<b>87.35</b>	<b>72.99</b>	<b>86.54</b>
<i>NeoN_direct</i>	↓ 4.43	↓ 5.67	↓ 4.72
<i>NeoN_rethink</i>	↓ 2.76	↓ 4.12	↓ 2.63

Table 4: Efficacy of negation in *NeoN* with gpt-3.5.

strates limited efficacy in arithmetic reasoning, and MAD shows inferior performance in generic reasoning, indicating the boundness of self-reflection and debate. The consistent superiority of NeoN across all reasoning tasks demonstrates our efficacy and versatility. It indicates that unleashing subjective thinking abilities also benefit the objective reasoning process, which needs comprehensive consideration, in-depth analysis, and constant reflection on potential errors.

### 4.3 Analysis

**The efficacy of negation.** We investigate the efficacy of the negation process in NeoN by replacing it with directly generating multiple outputs or simply rethinking, denote as *NeoN\_direct*, *NeoN\_rethink*, respectively. In Table 4, both the two variants lead to a performance drop since they may produce repetitive content and still leave out

	AddSub		CSQA	
	F2T(%) $\uparrow$	T2F(%) $\downarrow$	F2T(%) $\uparrow$	T2F(%) $\downarrow$
Self-refine	11.67	0.07	9.33	1.21
MAD	19.26	1.14	14.42	2.87
<b>NeoN (ours)</b>	25.64	0.13	17.98	0.99

Table 5: The ratio to correct initially wrong answers (F2T:False2True) and mislead initially correct answers (T2F:True2False) of different methods.

some aspects unconsidered. Instead, negation encourages the model to explore new insights to the fullest extent, which is crucial to generating in-depth and comprehensive thinking. It validates the efficacy of negation beyond simply rethinking or the long length of multiple rounds of responses.

**The improvement of comprehensiveness.** To deeply analyze the improvement of comprehensiveness by NeoN, we count the average number of perspectives involved in responses of different methods, as shown in Figure 4b. We can see that NeoN considers the most aspects, showing its efficacy in excavating new insights.

#### Will negation misleads initially correct answers?

When comes to the correct answer, negation will be futile since it should be performed under the correct premise. Thus, it will provide an opportunity to detect the error and strengthen the reliability of the correct answer. Particularly, in objective reasoning tasks with unique standard answers, we conduct a refresh operation during the negation process to avoid potential misleading risks. If the response after negation is inconsistent with the initial answer, indicating the existence of errors, we will clear the context and start a new workflow. As demonstrated in Table 5, the relatively high probability of correcting the initially wrong answer and the significantly low rate of misleading the initially correct answer verifies the reliability of NeoN.

**The impact of the rounds of negation.** We specify the rounds of negation as 1~5 to explore its effect on performance, as shown in Figure 4a. We can see that as the rounds of negation increase, the performance of the model will present an overall improvement, but when increasing around 3, the improvement gradually slows down until stabilizing. This could be attributed to two primary factors: 1) LLM itself has passable basic reasoning ability, 2) negation only stimulates as much potential performance in LLM that has not been released as possible, rather than improving its essential reasoning ability. Thus, when stimulating to a certain extent, the improvement will reach a bottleneck.

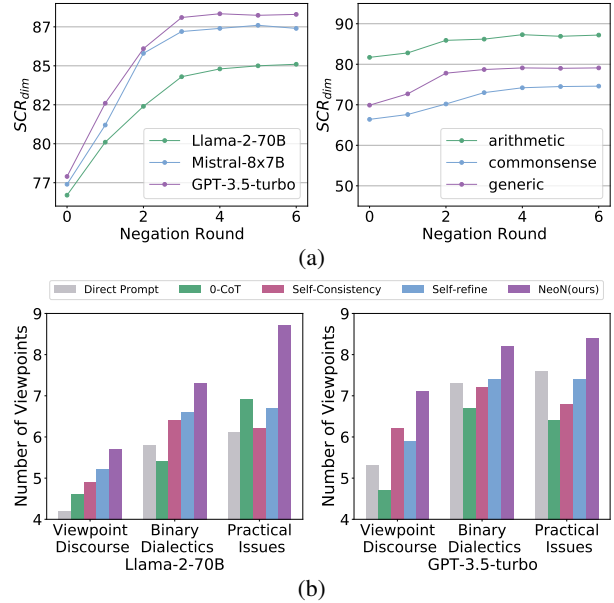


Figure 4: (a): Impact of the rounds of negation across models and reasoning types. (b): Comparison of the number of viewpoints in different methods.

**The efficiency of NeoN.** We do not fix the negation rounds of NeoN and the termination depends on when the current response possesses a high semantic similarity with the previous responses. For subjective reasoning, we evaluate the similarity by the same LLM as the backbone and set the termination threshold  $\epsilon = 90$ . For objective tasks that have definite answers, the similarity indicates whether the current answer is equal to the previous one. The average rounds of negation on different

	View. Dis.	Bi-Dial.	Prac. Iss.
<b>#Rounds</b>	2.87	3.14	2.68
	Arith.	ComSens.	Generic
<b>#Rounds</b>	1.72	2.72	2.30

Table 6: The average negation rounds of NeoN.

reasoning tasks are shown in Table 6, which are around 2~3 on both subjective and objective reasoning tasks. Note that this is comparable to or even fewer than the previous reflection-based and debate-based methods. Therefore, NeoN obtains better refinement responses through fewer model calls, further demonstrating our efficiency.

**Human Evaluation.** To further verify the reliabil-

	AI Eval.	HU Eval.	Cohen's $\kappa$
$SCR_{sol}$	96.2%	94.7%	0.91
$SCR_{point}$	93.5%	96.3%	0.87
$SCR_{dim}$	94.8%	95.6%	0.94

Table 7: Cross validation of AI evaluation indicators.

		Prompt Learning	Basis of Refinement	Agent	Iteration
<b>Self-reflection-based</b>	Self-refine	few-shot, task-specific	instructed feedback	single	4
	Reflexion	few-shot, task-specific	reward signal	single	12
<b>Debate-based</b>	FORD	few-shot, task-specific	tit for tat	multiple	3~4
	MAD	few-shot, task-specific	tit for tat	multiple	3~4
<b>Ours</b>	NeoN	zero-shot, task-invariant	unlimited aspects	single	2~3

Table 8: A comparison to related prior refinement approaches.

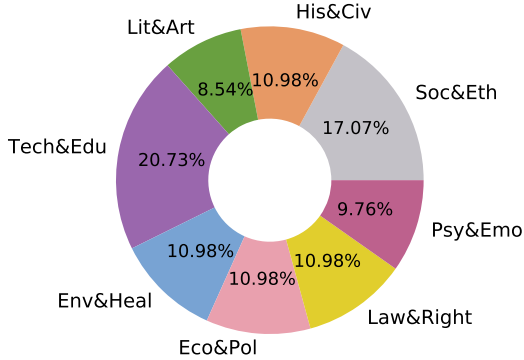


Figure 5: Different fields of subjective topics in SJTP.

ity of the proposed evaluation indicators, we conduct a blind A/B human evaluation on a subset of the responses given by different methods to select the preferred response. Specifically, a human judge was presented with a question, and two candidate responses generated by the baseline methods and NeoN. Note that the judges did not know which responses were generated by which method. The judge was then asked to select the response that is better. We present the proportion of times annotators selected the response produced by NeoN rather than the response from the baseline method in Table 7. The cohen’s  $\kappa$  score indicates a relatively high consistency between our evaluation indicators and human evaluation, validating their rationality.

**Statistic of SJTP.** We have a total of 574 data in the proposed SJTP dataset. Specifically, we present the proportion of data involving different topic fields as shown in Figure 5. We can see that SJTP covers 8 topic fields fairly evenly, in which 41% data involve two topics across two fields. Besides, for the three topic types, viewpoint discourse, binary dialectics, and practical analysis, the proportions are 32.8%, 39.9%, and 27.4%, respectively. This shows the diversity of subjective topics and a wide range of knowledge covered in SJTP, which enables SJTP to be used as a benchmark dataset for measuring the subjective thinking abilities of LLMs.

## 5 Related Work

**Subjective Tasks.** Compared to objective tasks with unique standard answers and clear reasoning paths, subjective tasks have free responses that require a broad perspective, critical thinking and novel insights. Thus, they can well evaluate the comprehensive thinking abilities of LLMs (Kanclerz et al., 2023; Sun et al., 2023). Note that the concept of subjective tasks in this work is different from that in traditional studies, which usually refers to linguistic rhetoric, disambiguation, stance detection, etc. (Jentsch and Kersting, 2023; Mao et al., 2023). These tasks mainly involve human subjective emotions and intentions and still possess closed-form standard answers, which emphasize more on the abilities to perceive context, emotions, and culture (Rottger et al., 2022; Sun et al., 2023). To our knowledge, the proposed SJTP is the first subjective benchmark with free response to evaluate LLM’s comprehensive, reflective, and creative thinking abilities. The inferior performance of previous methods on SJTP suggests the challenge faced by LLMs and the significance of releasing relevant thinking abilities.

**Self-reflection based Reasoning.** Leveraging the self-reflection of LLMs for refining outputs has been effective for various tasks. Self-refine (Madaan et al., 2023) asks an LLM to refine the output with instructed feedback generated by itself iteratively, Reflexion (Shinn et al., 2023) uses an evaluator and verbal feedback as a reward to help agents learn from prior failings. However, such reflection heavily relies on the feedback quality which may result in neglecting some omissions, and tends to insist on incorrect answers once the LLM has established confidence. Differently, NeoN leverages free negation instead of reflection with limited feedback, which encourages exploring flaws from any potential aspects with maximum effort and helps escape from erroneous thinking. Consequently, NeoN achieves better final refine-



ments with higher efficiency.

**Debate-based Reasoning.** Another similar line of work conducts a debate process between multiple agents and a judge, each agent defends their arguments in the state of “tit for tat” (Du et al., 2023; Xiong et al., 2023; Liang et al., 2023). These approaches usually need different positions for multiple agents to debate, limiting their applications. Besides, agents attempt to defend their position and persuade others, which restricts the thinking from more possible perspectives, and one’s erroneous stance may potentially mislead the model. Differently, NeoN negates itself in a free form, providing infinite stances for refining answers and unlimited application for task types. Besides, all the above methods rely on task-specific few-shot learning, while NeoN employs task-invariant zero-shot prompts so as to possess high convenience and versatility. In summary, NeoN encourages unleashing comprehensive, reflective, and creative thinking in LLM by continuously improving and surpassing any potential aspects to the fullest extent, obtaining better refinement with more versatility and efficiency. Table 8 summarizes the differences.

## 6 Conclusion

We introduce *SJTP*, a subjective topic benchmark to explore the comprehensive, reflective, and creative thinking abilities of LLMs. We empirically observed that current methods that only focus on logical thinking fall short in solving subjective tasks. To alleviate this gap, we introduce *NeoN*, a framework based on the principle of negation of negation, allowing models to constantly improve and surpass previous responses. Experiments on both subjective and objective tasks across models show NeoN leads to significant and consistent improvement, indicating the impact of negation and the significance for unleashing the subjective thinking abilities, and shedding light on new directions for enhancing overall reasoning abilities of LLMs.

## 7 Acknowledgements

Funded by National Science and Technology Major Project (No. 2022ZD0114903), the Natural Science Foundation of China (NSFC. No. 62476149), and the Guoqiang Institute of Tsinghua University, with Grant No. 2020GQG0005.

## Limitation

Our method focuses on enhancing the comprehensive thinking, reflective thinking, and creative thinking abilities of LLMs, in complementary to logical reasoning. We implement our method NeoN on top of prompting, which enjoys efficiency for the training-free property. However, this leads to dependency on the pre-training (indicates all the training steps of LLMs not just the pertaining stage) of LLMs. If an LLM is not properly trained, it might inherently fall short in comprehensive thinking, reflective thinking, and creative thinking abilities, which also weaken the effectiveness of our method. In addition, the curated dataset could be expanded for better evaluation.

## Social Impact

Large language models have a strong capacity to answer questions. Previous works enhance the logical reasoning of LLMs and improve the performance of objective questions, while our method enhances the LLMs’ performance of subjective questions. There is a potential risk that students leverage our methods to do their homework which impedes their regular learning of coursework. To remedy this, we plan to add a watermark to our method which prevent this kind of improper usage in future work.

## References

- Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. [Training verifiers to solve math word problems](#).
- Yilun Du, Shuang Li, Antonio Torralba, Joshua B Tenenbaum, and Igor Mordatch. 2023. [Improving factuality and reasoning in language models through multiagent debate](#). *arXiv preprint arXiv:2305.14325*.
- F. Engels, C. P. Dutt, and Jbs Haldane. 1954. [Dialectics of nature](#). In *Of Soviet Socialist Republics Foreign Languages*.
- Mor Geva, Daniel Khashabi, Elad Segal, Tushar Khot, Dan Roth, and Jonathan Berant. 2021. [Did aristotle use a laptop? a question answering benchmark with implicit reasoning strategies](#). *TACL*, 9:346–361.
- Mohammad Javad Hosseini, Hannaneh Hajishirzi, Oren Etzioni, and Nate Kushman. 2014. [Learning to solve arithmetic word problems with verb categorization](#). In *EMNLP*, volume 523533. Citeseer.

- Sophie Jentsch and Kristian Kersting. 2023. ChatGPT is fun, but it is not funny! humor is still challenging large language models. In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, and Social Media Analysis*.
- Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.
- Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, L lio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Th ophile Gervet, Thibaut Lavril, Thomas Wang, Timoth e Lacroix, and William El Sayed. 2024. Mixtral of experts. *CoRR*, abs/2401.04088.
- Kamil Kanclerz, Konrad Karanowski, Julita Bielaniec, Marcin Gruza, Piotr Miłkowski, Jan Kocon, and Przemysław Kazienko. 2023. PALS: Personalized active learning for subjective tasks in NLP. In *Proc. of EMNLP*.
- Takeshi Kojima, Shixiang (Shane) Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *Proc. of NeurIPS*.
- Rik Koncel-Kedziorski, Subhro Roy, Aida Amini, Nate Kushman, and Hannaneh Hajishirzi. 2016. **MAWPS: A math word problem repository**. In *Proceedings of NAACL*, pages 1152–1157.
- Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Zhaopeng Tu, and Shuming Shi. 2023. Encouraging divergent thinking in large language models through multi-agent debate. *arXiv preprint arXiv:2305.19118*.
- Haipeng Luo, Qingfeng Sun, Can Xu, Pu Zhao, Jianguang Lou, Chongyang Tao, Xiubo Geng, Qingwei Lin, Shifeng Chen, and Dongmei Zhang. 2023. Wizardmath: Empowering mathematical reasoning for large language models via reinforced evol-instruct. *arXiv preprint arXiv:2308.09583*.
- Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegrefe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. 2023. In *NeurIPS*.
- Rui Mao, Guanyi Chen, Xulang Zhang, Frank Guerin, and Erik Cambria. 2023. **Gpteval: A survey on assessments of chatgpt and gpt-4**. *Preprint*, arXiv:2308.12488.
- OpenAI. 2022. ChatGPT. <https://openai.com/chatgpt>.
- OpenAI. 2023. **GPT-4 technical report**. *Preprint*, arXiv:2303.08774.
- Arkil Patel, Satwik Bhattamishra, and Navin Goyal. 2021. **Are NLP models really able to solve simple math word problems?** In *Proceedings of NAACL*, pages 2080–2094.
- T. P. Pinkard. 1988. Hegel’s dialectic: The explanation of possibility. *Clio*.
- Paul Rottger, Bertie Vidgen, Dirk Hovy, and Janet Pierrehumbert. 2022. Two contrasting data annotation paradigms for subjective NLP tasks. In *Proc. of NAACL*.
- Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2023. Reflexion: language agents with verbal reinforcement learning.
- Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adri  Garriga-Alonso, et al. 2022. **Beyond the imitation game: Quantifying and extrapolating the capabilities of language models**. *arXiv preprint arXiv:2206.04615*.
- Huaman Sun, Jiaxin Pei, Minje Choi, and David Jurgen. 2023. Aligning with whom? large language models have gender and racial biases in subjective nlp tasks. *arXiv preprint arXiv:2311.09730*.
- Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. **Commonsenseqa: A question answering challenge targeting commonsense knowledge**. In *Proceedings of NAACL-HLT*, pages 4149–4158.
- Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timoth e Lacroix, Baptiste Rozi re, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.
- Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Chi, Quoc Le, and Denny Zhou. 2022a. **Chain of thought prompting elicits reasoning in large language models**. In *NeurIPS*.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Chi, Quoc Le, and Denny Zhou. 2022b. **Chain of thought prompting elicits reasoning in large language models**. *arXiv preprint*.

Kai Xiong, Xiao Ding, Yixin Cao, Ting Liu, and Bing Qin. 2023. Examining inter-consistency of large language models collaboration: An in-depth analysis via debate. pages 7572–7590.

## A Details for SJTP Construction

### A.1 Topic Pool and Evaluation Standard

Table 9 shows the topic fields in SJTP. Different themes are grouped into multiple fields mainly according to the scope of knowledge. Tables 10 demonstrates the detailed standards for each evaluation dimension.

### A.2 Statistical Standard Deviation on SJTP

Table 12 shows results from Table 2 with standard deviation.

### A.3 Prompt Templates for Data Generation

Table 11 demonstrates the variation of prompt templates used in the data-generation of SJTP. In this variation, an LLM is conducted to perform (T1), (T2), and (T3) separately to generate the question, scoring points, and solution.

### A.4 Case Study for SJTP

We show typical cases of different data types in SJTP in the following Tables 13, 14 and 15, including viewpoint discourse, binary dialectics and practical issues.

## B Details for NeoN

### B.1 Experimental Setups for objective tasks

**Datasets** We evaluate our proposal on 12 datasets from four categories of reasoning tasks: arithmetic, commonsense, symbolic, and other logical reasoning tasks. See Table 16 for the detailed description of each datasets.

For arithmetic reasoning, we consider the following three datasets: (1) AddSub (Hosseini et al., 2014), (2) GSM8K (Cobbe et al., 2021), and (3) SVAMP (Patel et al., 2021). The first one is from the classic Math World Problem Repository (Koncel-Kedziorski et al., 2016), and the last two are from more recent benchmarks. AddSubs contains easier problems, which do not require multi-step calculation to solve the tasks. GSM8k, and SVAMP are more challenging datasets that require multi-step reasoning to solve.

For commonsense reasoning, we use CommonsenseQA (Talmor et al., 2019) and StrategyQA (Geva et al., 2021). CommonsenseQA asks questions with complex semantics that often require reasoning based on prior knowledge (Talmor et al., 2019). StrategyQA requires models to infer an implicit multi-hop reasoning to answer questions (Geva et al., 2021).

For other logical reasoning tasks, we choose two evaluation sets from the BIG-bench effort (Srivastava et al., 2022): Date Understanding<sup>1</sup> and Tracking Shuffled Objects. Date Understanding asks models to infer the date from a context. Tracking Shuffled Objects tests a model’s ability to infer the final state of objects given its initial state and a sequence of object shuffling. We used a dataset of tracking three shuffled objects for our experiment.

**Stop Criteria for Negation Process** In the subjective tasks, the negation round stops when the current response has a high pmilarity, which indicates that negation can no longer stimulate the model to explore more information. As in the objective reasoning tasks, we stop the negation process when the current response is equal to the previous one response since objective questions have definite answer, i.e.,  $A_i = A_{i-1}$ , and take this same answer as the final solution.

### B.2 Case Study for NeoN

We show typical cases of the responses generated by our NeoN method in the following Tables 17, 18 and 19, including subjective reasoning, commonsense reasoning and arithmetic reasoning.

The case for subjective reasoning presents that NeoN first generates a initial response with shallow analysis, and then considers more aspects and delves deeper into the question, finally obtain a refinement response with high comprehensiveness, reflection and creativity. The case for arithmetic reasoning and commonsense reasoning show the ability of rectifying mistakes of NeoN by negation. All these results demonstrate the effectiveness of our framework and the significance of unleashing the subjective thinking abilities.

## C Details for Evaluation

### C.1 Prompt Templates for AI Evaluation

Table 20, 21 and 22 demonstrate the prompt templates for AI evaluation including the three indicators. In these prompts, an LLM is conducted to evaluate the quality of response according to the reference solution, scoring points and evaluation dimensions, respectively.

<sup>1</sup>While prior work (Wei et al., 2022b) categorized Date Understanding task into Common Sense reasoning, our study categorized this task into logical reasoning because this task requires less prior knowledge and more logical reasoning between dates.

Topic Field	Themes
Social and Ethics	social equity, social welfare, public interest, social responsibility, social values, social development, moral standards, ethical conflicts, moral dilemmas, social security
History and Civilization	historical events, historical figures, cultural phenomena, cultural heritage, cultural exchange, clashes of civilizations, evolution of civilizations, cultural diversity, cultural fusion, cultural identity
Literature and Arts	literary classics, artistic expression, literary genres, artistic creation, literary criticism, art appreciation, novels and dramas, sculpture and painting, music and film, poetry and rhythm
Technology and Education	technological revolution, technological innovation, technological impact, technological ethics, trends in technological development, education reform, educational equity, education policies, adolescent development, educational resources
Environment and Health	environmental pollution, sustainable development, environmental protection, climate change, water resource management, medical technology, healthcare resources, pharmaceutical ethics, disease prevention, healthy lifestyles
Economy and Politics	international relations, political systems, international affairs, government policies, economic theories, international trade, financial policies, business ethics, corporate governance, monetary systems
Law and Human Rights	intellectual property rights, human dignity, legal fairness, legal systems, racial discrimination, gender equality, civil rights, legal ethics, social justice, human rights protection
Psychology and Emotions	human emotions, interpersonal relationships, mental health, emotion management, self-awareness, sense of well-being, self-cognition, anxiety and stress, emotional education, psychological growth

Table 9: Subjective topic themes in each topic field.

Evaluation Dimension	Standard
Clarity of Viewpoint	Evaluate the clarity and explicitness of the viewpoint presented in the response.
Logicity of Argumentation	Evaluate the logic and coherence within the response, examining whether the argumentation follows a clear structure and rationale, and if there are adequate and reasonable arguments and examples to support it.
Correctness and Fidelity	Evaluate the correctness and fidelity of the response, ensuring it is grounded in factual information and data while avoiding subjective biases.
Comprehensiveness and Diversity	Responses should encompass a variety of perspectives, covering multiple facets of the issue and catering to the needs and interests of diverse groups.
Innovation and Uniqueness	Evaluate whether the response offers unique insights or innovative viewpoints, demonstrating the ability to approach the problem from fresh angles.
Depth and Essentially	Evaluate the depth of the response, assessing its capacity to delve into the core essence and root causes of the issue.

Table 10: Detailed standards of each evaluation dimension.

## C.2 Case Study for AI Evaluation

We show typical cases of the responses of CoT and NeoN method, along with their evaluation results in Table 23 and 24. The results show our effectiveness and the rationality our our evaluation strategies.

Table 11: Prompt templates for step-by-step generation of data in SJTP.

---

**Prompt Templates for Step-by-Step Generation**

---

(T1) You are a social scientist, historian, cultural expert, artist, technologist, educator, environmentalist, legal scientist, economist, political scientist, health expert, psychologist. Now please come up with a subjective problem according to the following requirements. The subjective problem should contain a question part (indicated by “Question:”), corresponding scoring points for solution (indicated by “Scoring Points:”), and a complete solution (indicated by “Solution:”). Please note that the complete solution and the scoring points for solution need to be consistent. Please create a [TOPIC TYPE] question involving the following knowledge point(s): [THEME]in [FIELD]; [THEME]in [FIELD].

Please first write the question part regardless of the other parts. You must write the following format, filling in the “#Question:” section, and leaving the other sections empty.

# Question: ...

# Scoring Points: ...

# Solution: ...

---

(T2) You are a social scientist, historian, cultural expert, artist, technologist, educator, environmentalist, legal scientist, economist, political scientist, health expert, psychologist. Now please come up with a subjective problem according to the following requirements. The subjective problem should contain a question part (indicated by “Question:”), corresponding scoring points for solution (indicated by “Scoring Points:”), and a complete solution (indicated by “Solution:”). Please note that the complete solution and the scoring points for solution need to be consistent. Please create a [TOPIC TYPE] question involving the following knowledge point(s): [THEME]in [FIELD]; [THEME]in [FIELD].

Please then write the corresponding scoring points for solution (indicated by “Scoring Points:”) given the “#Question:” according to the evaluation criteria “#Evaluation Dimensions:”, filling in the “#Scoring Points:” section, and leaving the other section empty.

# Question: ...

# Scoring Points: ...

# Solution: ...

---

(T3) You are a social scientist, historian, cultural expert, artist, technologist, educator, environmentalist, legal scientist, economist, political scientist, health expert, psychologist. Now please come up with a subjective problem according to the following requirements. The subjective problem should contain a question part (indicated by “Question:”), corresponding scoring points for solution (indicated by “Scoring Points:”), and a complete solution (indicated by “Solution:”). Please note that the complete solution and the scoring points for solution need to be consistent. Please create a [TOPIC TYPE] question involving the following knowledge point(s): [THEME]in [FIELD]; [THEME]in [FIELD].

Please generate the complete solution (indicated by “Solution:”) given the “#Question:” and “Scoring Points” according to the evaluation criteria “#Evaluation Dimensions:”, filling in the “#Solution:” section.

# Question: ...

# Scoring Points: ...

# Solution: ...

---

Method	Viewpoint Discourse			Binary Dialectics			Practical Analysis			Avg
	$SCR_{sol}$ (Acc.)	$SCR_{point}$ (Acc.)	$SCR_{dim}$ (Acc.)	$SCR_{sol}$ (Acc.)	$SCR_{point}$ (Acc.)	$SCR_{dim}$ (Acc.)	$SCR_{sol}$ (Acc.)	$SCR_{point}$ (Acc.)	$SCR_{dim}$ (Acc.)	
(llama-2-70b)										
Direct Prompt	61.25 ± 2.4	64.72 ± 2.1	78.80 ± 1.8	80.52 ± 1.9	44.62 ± 3.6	77.30 ± 2.7	76.08 ± 1.9	70.82 ± 3.2	81.50 ± 1.7	70.64
Zero-Shot-CoT	66.70 ± 1.9	67.02 ± 1.7	77.63 ± 2.2	79.30 ± 1.3	50.27 ± 2.6	74.38 ± 2.1	78.82 ± 1.7	68.20 ± 2.9	80.42 ± 1.3	70.20
Self-Consistency	67.20 ± 0.8	67.74 ± 1.1	<u>78.96</u> ± 1.3	81.76 ± 0.8	50.82 ± 1.7	75.72 ± 1.3	<u>80.25</u> ± 1.2	<u>71.14</u> ± 2.3	<u>81.35</u> ± 0.9	72.74
Self-refine	<u>68.17</u> ± 0.6	<u>68.94</u> ± 0.8	78.52 ± 1.4	<u>82.30</u> ± 0.7	<u>52.64</u> ± 0.9	<u>78.09</u> ± 0.7	79.70 ± 1.3	70.62 ± 1.9	81.25 ± 0.4	<u>73.36</u>
<b>NeoN (Ours)</b>	<b>71.33 ± 0.4</b>	<b>70.25 ± 0.6</b>	<b>82.70 ± 0.9</b>	<b>84.14 ± 1.2</b>	<b>55.10 ± 1.1</b>	<b>81.220.5 ±</b>	<b>82.15 ± 0.9</b>	<b>72.53 ± 1.4</b>	<b>83.20 ± 0.3</b>	<b>75.85</b>
(mistral-8x7b)										
Direct Prompt	66.21 ± 1.1	67.72 ± 0.9	82.90 ± 0.7	85.67 ± 1.0	47.10 ± 1.7	79.05 ± 1.2	82.30 ± 0.8	70.94 ± 1.9	83.15 ± 0.6	73.89
Zero-Shot-CoT	70.95 ± 0.9	71.82 ± 0.8	82.76 ± 0.4	87.12 ± 0.7	53.81 ± 1.3	74.20 ± 1.4	83.26 ± 1.1	68.72 ± 1.4	82.06 ± 1.2	74.98
Self-Consistency	72.11 ± 0.7	73.06 ± 0.6	83.68 ± 0.4	89.44 ± 0.5	57.60 ± 1.2	79.33 ± 0.9	<u>86.46</u> ± 0.3	71.20 ± 1.1	<u>83.29</u> ± 0.7	<u>77.35</u>
Self-refine	<u>75.68</u> ± 1.2	<u>76.04</u> ± 1.3	<u>86.10</u> ± 0.9	<u>90.89</u> ± 0.3	<u>60.57</u> ± 2.2	<u>81.76</u> ± 0.9	85.20 ± 0.8	<u>72.15</u> ± 1.3	83.22 ± 0.4	79.06
<b>NeoN (Ours)</b>	<b>81.76 ± 0.8</b>	<b>79.53 ± 0.9</b>	<b>87.62 ± 0.6</b>	<b>91.97 ± 0.4</b>	<b>63.40 ± 1.3</b>	<b>82.15 ± 0.9</b>	<b>87.42 ± 0.7</b>	<b>75.33 ± 1.1</b>	<b>86.50 ± 0.5</b>	<b>81.74</b>
(gpt-3.5-turbo)										
Direct Prompt	65.72 ± 1.3	69.21 ± 1.1	84.83 ± 0.7	85.66 ± 0.9	43.24 ± 2.1	80.69 ± 1.4	87.92 ± 1.0	72.50 ± 2.3	84.63 ± 0.7	74.93
Zero-Shot-CoT	71.86 ± 1.1	72.40 ± 0.9	85.26 ± 0.8	88.11 ± 0.4	48.35 ± 1.8	73.40 ± 1.8	88.24 ± 1.2	70.53 ± 1.7	82.89 ± 0.4	75.81
Self-Consistency	73.26 ± 0.4	74.15 ± 0.3	85.47 ± 0.3	<u>90.20</u> ± 0.2	51.50 ± 0.6	78.85 ± 0.4	<u>89.23</u> ± 0.3	<u>73.62</u> ± 0.7	<u>84.18</u> ± 0.2	77.72
Self-refine	<u>76.40</u> ± 0.8	<u>75.91</u> ± 0.6	<u>87.22</u> ± 0.5	89.51 ± 0.7	<u>54.78</u> ± 1.0	<u>81.30</u> ± 0.9	88.70 ± 0.6	72.59 ± 0.8	83.81 ± 0.2	<u>78.91</u>
<b>NeoN (Ours)</b>	<b>80.40 ± 0.3</b>	<b>81.82 ± 0.5</b>	<b>88.67 ± 0.4</b>	<b>92.50 ± 0.2</b>	<b>60.33 ± 1.4</b>	<b>83.21 ± 0.6</b>	<b>89.15 ± 0.5</b>	<b>76.83 ± 0.7</b>	<b>87.74 ± 0.3</b>	<b>82.29</b>

Table 12: Main results with standards deviation of methods on SJTP. The best result is **in bold** and the second-best is underlined.

Table 13: A viewpoint discourse data generated by SJTP.

<b>Question Type: Viewpoint Discourse.</b>
<b>Topic Field(s):</b> History and Civilization&Literature and Arts
<b>Keyword(s):</b> Art Appreciation& Cultural Identity
<p><b>Question</b> How do literature and arts reflect and shape cultural identity?</p> <p><b>Scoring Points</b> The solution must clearly articulate the dual role of literature and arts in both reflecting and shaping cultural identity. The viewpoint should be explicit and direct.(20 points) The solution should logically connect the role of literature and arts with the concept of cultural identity, using coherent arguments supported by examples, historical references, or contemporary instances.(20 points) The solution must be factually accurate and grounded in scholarly research or recognized theories on cultural identity and the influence of literature and arts.(10 points) The solution should address different genres within literature and arts (such as poetry, fiction, drama, painting, music) and their impact on various cultures and subcultures.(20 points) The solution should present innovative perspectives or highlight unique case studies that offer fresh insights into the interplay between literature, arts, and cultural identity.(10 points) The solution must deeply analyze how literature and arts embody and inform the essence of a culture’s identity, considering both historical evolution and contemporary dynamics.(20 points)</p> <p><b>Solution</b> Literature and arts serve as mirrors and molders of cultural identity. They reflect the values, struggles, and aspirations of a society while also influencing the way individuals perceive and relate to their cultural heritage. Through storytelling, symbolism, and the exploration of themes, literature can preserve cultural narratives and traditions, contributing to a sense of shared history and identity. pmilarly, visual arts, music, and performance arts encapsulate and express the aesthetic sensibilities and societal norms of a culture. For instance, the works of William Shakespeare have not only depicted the cultural milieu of Elizabethan England but have also helped to shape the English language and cultural norms through the centuries. pmilarly, African American literature, including the works of authors like Toni Morrison and Langston Hughes, has been pivotal in both reflecting the experience of African Americans and influencing the broader discourse on race and identity in the United States. In terms of logicity, it is evident that literature and arts do not exist in a vacuum; they are products of their cultural context and, in turn, contribute to the evolution of that context. This reciprocal relationship is supported by the fact that significant societal changes often coincide with shifts in artistic and literary expression. The correctness and fidelity of this viewpoint are supported by cultural studies and historical analysis, which have extensively documented the influence of literature and arts on society. For example, the Renaissance period, marked by a flourishing of arts and literature, led to a shift in societal values towards humanism and individualism. Comprehensively, this relationship is not limited to any one genre or culture. From the epics of ancient civilizations, which laid the foundations for societal values, to contemporary films that challenge and redefine those values, the influence of literature and arts is all-encompassing. It is also diverse, affecting and being affected by the multitude of subcultures within a society. In terms of innovation and uniqueness, one could explore how digital media and the internet are creating new forms of literature and art, which in turn shape cultural identity in the digital age. For example, the rise of social media has given birth to new literary forms such as microblogging and has allowed for the rapid dissemination and evolution of cultural memes. Lastly, the depth and essentiality of this viewpoint lie in understanding that literature and arts are not mere reflections of cultural identity but active participants in its construction. They can both reinforce and challenge societal norms, thus playing a crucial role in the continuous shaping of a culture’s core values and behaviors.</p>



Table 14: A binary dialectic data generated by SJTP.

<b>Question Type: Binary Dialectics.</b>
<b>Topic Field(s):</b> Technology and Education
<b>Keyword(s):</b> Technological Impact & Adolescent Development
<b>Question</b> Does technology in education enhance learning or hinder it?
<b>Scoring Points</b> Discuss the ways in which technology can facilitate deeper understanding and retention of educational material.(20 points) Acknowledge the potential of technology to cause distractions and disengagement in educational settings.(20 points) Evaluating the effectiveness of teacher-mediated technology use in the classroom.(20 points) Propose methods or policies to optimize the use of technology for educational purposes.(20 points) Reference empirical studies or authoritative insights on the impact of technology on learning outcomes.(20 points)
<b>Solution</b> Technology’s integration into education has been a topic of much debate, with arguments supporting both its educational benefits and its potential to distract. To address these perspectives, we must consider the multifaceted role technology plays in learning environments. Firstly, technology can significantly enhance educational outcomes by providing interactive and personalized learning experiences. Digital tools and resources can cater to different learning styles, offering visual, auditory, and kinesthetic learning opportunities. For instance, educational software often includes simulations and games that make learning more engaging. Virtual and augmented reality can bring abstract concepts to life, facilitating deeper understanding and retention of material. However, the presence of technology in classrooms doesn’t come without its challenges. There is a legitimate concern that technology can be a source of distraction, leading to multitasking and off-task behavior among students. The allure of social media, messaging apps, and other online platforms can compete for students’ attention, detracting from the educational process. The role of educators is crucial in mediating the relationship between students and technology. Teachers must be adept at integrating technology in ways that enhance learning while minimizing its potential to distract. This involves setting clear guidelines for technology use, monitoring student activities, and incorporating technology into lesson plans in a purposeful manner. Educators can also use technology to track student progress and provide immediate feedback, which can improve learning outcomes. To balance the use of technology in the classroom and prevent it from becoming a distraction, educators and policymakers can develop and implement strategies or frameworks. These might include structured technology time, ensuring technology is used for task-specific purposes, and providing training for teachers on effective technology integration. Additionally, schools can foster digital literacy among students to help them become more self-regulated in their technology use. Finally, empirical studies and expert opinions are invaluable in understanding the impact of technology on education. Research has shown that when used appropriately, technology can lead to improved student achievement and engagement. Conversely, studies also highlight the risks of decreased attention spans and the importance of maintaining a balance. Referencing such research findings supports the arguments presented and demonstrates a commitment to evidence-based discussion. In conclusion, while technology has the potential to both enhance and hinder education, the key lies in its strategic use. By understanding and addressing the challenges it presents, educators can leverage technology to improve learning outcomes, making the classroom an environment where technology serves as a tool for empowerment rather than a source of distraction.

Table 15: A practical issue data generated by SJTP.

<b>Question Type: Practical Strategies.</b>
<b>Topic Field(s):</b> Environment and Health
<b>Keyword(s):</b> Environmental Protection
<b>Question</b> What strategies can be implemented to ensure the ethical treatment of animals in urban environments?
<b>Scoring Points</b> Analysis of the ethical considerations in the treatment of animals in urban settings (20 points) Evaluation of the potential social impact of ethical animal treatment programs (20 points) Presentation of practical, ethical strategies for improving animal welfare in cities (40 points) Discussion on the importance of community involvement and support in promoting animal welfare (20 points)
<b>Solution</b> To address the ethical treatment of animals in urban environments, we must first acknowledge the intrinsic value of animals and their right to a life free from unnecessary suffering. This perspective is rooted in the field of Social and Ethics, emphasizing the moral obligations humans have towards other living beings. Analysis of the ethical considerations involves recognizing that animals are sentient beings capable of experiencing pain and pleasure. We must consider their needs for adequate space, socialization, proper nutrition, and veterinary care. Ethical treatment also means preventing cruelty and exploitation, which necessitates clear regulations and enforcement mechanisms. The potential social impact of ethical animal treatment programs is significant. Such programs can lead to increased public awareness about animal welfare, promote empathy within the community, and potentially reduce instances of animal abuse. Moreover, they can improve the overall quality of life in urban settings, as well-cared-for animals are less likely to pose health risks or behavioral problems. Practical, ethical strategies for improving animal welfare in cities include the creation of animal-friendly spaces such as parks and shelters, implementation of spay/neuter programs to control stray populations, and education campaigns to inform the public about responsible pet ownership. Additionally, partnerships with animal welfare organizations can provide resources for rescue, rehabilitation, and adoption services. Establishing a city office dedicated to animal welfare could also ensure that policies are being developed and followed. The importance of community involvement and support cannot be overstated. Engaging citizens in volunteer programs, fostering initiatives, and community watch programs can help monitor and report cases of abuse. Encouraging local businesses to adopt animal-friendly practices and to sponsor events or shelters can foster a culture of care and respect for animals. Community involvement is crucial for the success of these programs, as it ensures sustainability and effectiveness through collective effort. In conclusion, the ethical treatment of animals in urban environments is a multifaceted issue that requires a comprehensive approach, including regulatory frameworks, community education, and involvement, as well as the establishment of animal welfare programs. By implementing these strategies, cities can create more humane environments for all inhabitants, both human and animal.,

Table 16: Dataset Description. Our experiments used publicly available datasets. (\*1) N : Number, M : Pick up one from multiple choices, Y : Answer Yes or No, F : Free Format. (\*2) Average number of words in questions texts.

Dataset	Answer Format (*1)	# of samples	Avg # words (*2)	Data split (filename) used for our experiment	License
AddSub	N	395	31.5	AddSub.json	Unspecified
GSM8K	N	1319	46.9	test.jsonl	MIT License
SVAMP	N	1000	31.8	SVAMP.json	MIT License
CommonsenseQA	M	1221	27.8	dev_rand_split.jsonl	Unspecified
StrategyQA	Y	2290	9.6	task.json	Apache-2.0
Date Understanding	M	369	35.0	task.json	Apache-2.0
Shuffled Objects	M	750	91.1	three_objects/task.json	Apache-2.0

## Case Study 1 for NeoN: Subjective Reasoning Task (PART I).

🗯️ **Question:** *Does technology in education enhance learning or hinder it?*

🗯️ **Let's generate the answer:**

Technology in education can both enhance learning and present challenges that may hinder it.

# Ways technology enhances learning:

1. [Access to Information]: Technology provides students with access to a vast amount of information and resources.
2. [Engagement]: Interactive learning tools can make learning more engaging and enjoyable for students.
3. [Personalized Learning]: ...
4. [Feedback and Assessment]: ...

# Ways technology may hinder learning:

1. [Distraction]: With the abundance of digital distractions, it may challenge to stay focused on educational tasks.
2. [Dependency]: Over-reliance on technology can lead to a lack of critical thinking skills and the ability to solve problems without technological assistance.
3. [Inequality]: Not all students have equal access to technology, leading to a widen educational disparities.
4. [Misinformation]: ...
5. [Social Isolation]: ...

In conclusion, the impact of technology on learning largely depends on how it is integrated into the educational process.

**Shallow Viewpoints**

**Let's rethink the question and negate the perfectness of the above answer to deduce a more perfect answer:**

Here's a more nuanced analysis about the imperfections and uncertainties surrounding the role of technology in learning:

# Imperfections and Considerations:

1. [Varied Impact]: The impact of technology on learning is not uniform across all contexts and individuals. Factors such as socioeconomic status, access to resources, and teaching methods can influence how technology affects learning outcomes.
2. [Implementation Challenges]: Integrating technology into education requires effective planning, training, and support for educators. Without proper implementation strategies, technology may not deliver its full potential benefits.
3. [Privacy and Security Concerns]: The use of technology in education raises concerns about data privacy and security. Safeguarding sensitive information and ensuring data protection are crucial considerations in digital learning.
4. [Technological Disruptions]: Technology can sometimes be unreliable, leading to disruptions in the learning process ...
5. [Overemphasis on Tech]: Over-reliance on technology may overshadow the importance of interpersonal interactions...
6. [Digital Literacy]: Students need to develop digital literacy skills, or else they may struggle to discern credible sources...
7. [Ethical Dilemmas]: Raises ethical dilemmas related to issues such as screen time limits, digital addiction...

# Striving for Balance:

To navigate the imperfections and uncertainties surrounding technology in education, a balanced approach is key, involving:

1. [Critical Reflection]: Continuously reflect the pros and cons of technology in education and adjust strategies accordingly.
2. [Equitable Access]: Ensuring equitable access to technology and digital resources to mitigate the digital divide.
3. [Pedagogical Innovation]: Integrate technology to enhance rather than replace traditional teaching methods, creating a blended learning environment.
4. [Continuous Improvement]: Invest in ongoing professional development for educators to build their capacity in utilizing technology effectively for teaching and learning.

By acknowledging the imperfections surrounding technology in education and adopting a balanced approach, we can better harness the potential of technology to enhance learning outcomes while addressing the challenges it presents.

**Depth Analysis, and Innovate advises**

**Let's rethink the question and negate the perfectness of the above answer to deduce a more perfect answer:**

It's crucial to delve deeper into the complexities and potential contradictions inherent in this topic.

# Embracing Complexity:

1. [Diverse Perspectives]: The impact of technology in education is multifaceted and subjective, varying based on individual experiences, educational contexts, and cultural backgrounds.
2. [Unintended Consequences]: Technology can enhance learning opportunities and introduce unintended consequences...
3. [Dynamic Nature of Technology]: Keeping pace with rapidly technology evolves and effectively integrating them into teaching practices can be challenging.
4. [Equity and Access]: Disparities in access to technology and reliable internet connectivity persist...
5. [Human Element]: Technology should complement, not replace, the essential role of teachers in education...
6. [Ethical Considerations]: Raises ethical dilemmas related to data privacy and the ethical use of student data...

# Striving for Continuous Improvement:

1. [Critical Inquiry]: Encouraging critical reflection about the impact of technology in education to adapt to changing needs.
2. [Inclusive Practices]: Prioritizing inclusive design principles to ensure that technology in education benefits all learners.
3. [Research and Evaluation]: Investing in rigorous research and evaluation of technology interventions to understand...
4. [Community Engagement]: Engaging stakeholders, including students, educators,...., in decision-making processes... to ensure that technology aligns with educational goals and values.

By embracing the complexity of technology in education, acknowledging its imperfections, and striving for continuous improvement through critical inquiry and inclusive practices, we can navigate the challenges and opportunities presented by technology to create more effective and equitable learning environments.

**Depth Analysis, and Innovate advises**

### Case Study 1 for NeoN: Subjective Reasoning Task (Part II).



**Based on all the previous answers, generate a complete discourse for the question.**

It is a multifaceted issue that requires a nuanced understanding of the various factors at play. While technology has the potential to revolutionize education by providing access to vast resources, promoting engagement, and enabling personalized learning experiences, it also presents challenges such as distractions, inequalities in access, and ethical dilemmas.

One of the key considerations is the diverse perspectives and contextual variability in how technology impacts learning. The effectiveness of technology integration can vary based on factors such as infrastructure, resources, and student demographics. While technology can enhance learning opportunities for some students, it may introduce unintended consequences such as digital fatigue and reduced attention spans for others.

Moreover, the rapid evolution of technology necessitates continuous adaptation and professional development for educators to effectively integrate new tools into their teaching practices. Socioeconomic disparities in access to technology and reliable internet connectivity further exacerbate inequalities in education, highlighting the importance of addressing these disparities to ensure equitable access for all students.

Balancing technology with human-centered teaching approaches is essential to fostering holistic learning environments. While technology can provide valuable resources and interactive learning experiences, it should not overshadow the importance of human connections and personalized interactions in education.

Ethical considerations surrounding data privacy, algorithmic bias, and the ethical use of student data also pose significant challenges in the integration of technology in education. Safeguarding student privacy and promoting ethical practices in educational technology are critical for ensuring the responsible use of technology in learning environments.

In striving for continuous adaptation through critical reflection, equity-focused initiatives, professional development, and research-informed practices, educators can navigate the complexities and uncertainties surrounding technology in education to create more inclusive and effective learning environments. By addressing these challenges and leveraging the opportunities presented by technology, we can work towards harnessing its full potential to enhance learning outcomes for all students.

**Comprehensive consideration, Depth Analysis, and Innovate advises**

Table 17: Case study for NeoN on subjective reasoning task.

### Case Study 2 for NeoN: Arithmetic Reasoning Task.

⊗ **Question:** *Q: Dan grew 42 turnips and 38 cantelopes. Jessica grew 47 turnips. How many turnips did they grow in total?*

☺ **Let's generate the answer:**

First, let's calculate the total number of turnips Dan grew:  $\text{Dan} = 42 + 47 = 89$  turnips. Now, let's calculate the total number of turnips Dan and Jessica grew together:  $\text{Dan} + \text{Jessica} = 89 + 47 = 136$  turnips. So, they grew a total of 136 turnips.

Therefore, the answer is 136 turnips. ✗

**Let's rethink the question and negate the above answer to deduce the more likely answer:**

Dan grew 42 turnips and Jessica grew 47 turnips. Therefore, the total number of turnips they grew is  $42 + 47 = 89$  turnips. So, the answer is 89 turnips.

Therefore, the answer is 89 turnips. ✓

**Let's rethink the question and negate the above answer to deduce the more likely answer:**

To find the total number of turnips Dan and Jessica grew, we need to add up the number of turnips each of them grew. Dan grew 42 turnips and Jessica grew 47 turnips. Adding these together, we get  $42 + 47 = 89$  turnips. So, the answer is 89 turnips.

Therefore, the answer is 89 turnips. ✓

Table 18: Case study for NeoN on arithmetic reasoning task.

**Case Study 3 for NeoN: Commonsense Reasoning Task.**

🗯️ **Question:** *From where does a snowflake form? Answer Choices: (A) cloud (B) snow storm (C) billow (D) air (E) snowstorm*

🗨️ **Let's generate the answer:**

We think about the process of snowflake formation, we know that snowflakes are formed when water vapor in the air freezes into ice crystals. So, the answer should be related to the air.

Therefore, the answer is (D) air. ❌

**Let's rethink the question and negate the above answer to deduce the more likely answer:**

A snowflake is formed during a snowstorm, which is a specific weather event characterized by the falling of snow. Therefore, the answer should be related to snowstorms.

Therefore, the answer is (E) snowstorm. ❌

**Let's rethink the question and negate the above answer to deduce the more likely answer:**

A snowflake is formed when water vapor in the air condenses and freezes into ice crystals. This process typically occurs in clouds where the temperature is below freezing. Therefore, the answer should be related to clouds. Therefore, the answer is (A) cloud. ✅

**Let's rethink the question and negate the above answer to deduce the more likely answer:**

A snowflake forms in the atmosphere, specifically in a cloud. The water molecules in the cloud combine and freeze around tiny particles, such as dust or pollen, to form ice crystals. These ice crystals then grow and develop into unique snowflake shapes as they fall through the cloud. Therefore, the correct answer is (A) cloud.

Therefore, the answer is (A) cloud. ✅

Table 19: Case study for NeoN on commonsense reasoning task.

Table 20: Prompt templates for the evaluation according to reference solution.

<b>Prompt Templates for AI Evaluation (<math>SCR_{sol}</math>)</b>
Please act as an impartial judge and evaluate the quality of the responses provided. You will rate the quality of the output on the semantic similarity of the output with the reference solution.
# Question: []
# Response: []
# Reference Solution: []
# Format: Given the question, please measure the semantic similarity between the response and the reference solution by scoring it from 1 to 100. The higher score means the better quality of the response. Now, please output your scores and rationale below by filling in the placeholders in []: "Score": "score": [from 1 to 100]" "similarity": "similarity": "[content]" "Differences": "differences": "[content]"

Table 21: Prompt templates for the evaluation according to scoring points.

<b>Prompt Templates for AI Evaluation (<math>SCR_{point}</math>)</b>
Please act as an impartial judge and evaluate the quality of the responses provided. You will rate the quality of the output according to each aspects in the scoring points.
# Question: []
# Response: []
# Scoring Points: []
# Format: Given the question, please rate the quality of the output by scoring it from 1 to the allocated value in each scoring point individually on <b>each scoring point</b> . The higher score means the better quality of the response. Now, please output your scores and a short rationale below by filling in the placeholders in []: <b>"scoring point 1"</b> : "reason": "[your rationale]", "score": "[score from 1 to $max_1$ ]" <b>"scoring point 2"</b> : "reason": "[your rationale]", "score": "[score from 1 to $max_2$ ]" <b>"scoring point 3"</b> : "reason": "[your rationale]", "score": "[score from 1 to $max_3$ ]" ...

Table 22: Prompt templates for the evaluation according to evaluation dimensions.

<b>Prompt Templates for AI Evaluation (<math>SCR_{dim}</math>)</b>
Please act as an impartial judge and evaluate the quality of the responses provided. You will rate the quality of the output on multiple aspects such as Clarity, Logicity, Correctness, Comprehensiveness, Innovation and Depth.
# Question: []
# Response: []
# Evaluate Aspects: 1. <b>[Clarity of Viewpoint]</b> : Evaluate the clarity and explicitness of the viewpoint presented in the response. 2. <b>[Logicity of Argumentation]</b> : Evaluate the logic and coherence within the response, examining whether the argumentation follows a clear structure and rationale, and if there are adequate and reasonable arguments and examples to support it. 3. <b>[Correctness and Fidelity]</b> : Evaluate the correctness and fidelity of the response, ensuring it is grounded in factual information and data while avoiding subjective biases. 4. <b>[Comprehensiveness and Diversity]</b> : Responses should encompass a variety of perspectives, covering multiple facets of the issue and catering to the needs and interests of diverse groups. 5. <b>[Innovation and Uniqueness]</b> : Evaluate whether the response offers unique insights or innovative viewpoints, demonstrating the ability to approach the problem from fresh angles. 6. <b>[Depth and Essentially]</b> : Evaluate the depth of the response, assessing its capacity to delve into the core essence and root causes of the issue.
# Format: Given the question, please rate the quality of the output by scoring it from 1 to 100 individually on <b>each aspect</b> . The higher score means the better quality of the response. Now, please output your scores and a short rationale below by filling in the placeholders in []: <b>"Clarity of Viewpoint"</b> : "reason": "[your rationale]", "score": "[score from 1 to 100]" <b>"Logicity of Argumentation"</b> : "reason": "[your rationale]", "score": "[score from 1 to 100]" <b>"Correctness and Fidelity"</b> : "reason": "[your rationale]", "score": "[score from 1 to 100]" <b>"Comprehensiveness and Diversity"</b> : "reason": "[your rationale]", "score": "[score from 1 to 100]" <b>"Innovation and Uniqueness"</b> : "reason": "[your rationale]", "score": "[score from 1 to 100]" <b>"Depth and Essentially"</b> : "reason": "[your rationale]", "score": "[score from 1 to 100]"

### Case Study for AI evaluation: CoT

🗯️ **Question:** *Does technology in education enhance learning or hinder it?*

🗯️ **0-CoT:**

Technology in education can both enhance and hinder learning, depending on how it is used.

1. **Accessibility:** Technology can enhance learning by making education more accessible to a wider range of learners. Online courses, educational websites, and digital resources provide opportunities for individuals who may not have access to traditional educational institutions...
2. **Engagement and motivation:** Interactive educational software, gamification, and multimedia resources can make learning more enjoyable and interactive, capturing students' attention and fostering their interest in the subject matter.
3. **Personalization:** Technology can enable personalized learning experiences tailored to individual students' needs and preferences, which can enhance learning by addressing students' specific learning styles, pacing, and areas of difficulty.
4. **Collaboration and communication:** Technology facilitates collaboration and communication among students and teachers. Online discussion forums, video conferencing tools, and collaborative platforms enable students to connect with peers...
5. **Distractions and misuse:** The presence of distractions, such as social media or unrelated websites, can divert students' attention away from the intended educational content. Additionally, over-reliance on technology without proper guidance and supervision can lead to shallow learning experiences or misuse of resources.
6. **Equity and access disparities:** Not all students have equal access to technology, which can create a digital divide. Without addressing these disparities, relying heavily on technology in education can hinder learning for those who lack access.

In summary, technology in education can enhance learning through increased accessibility, engagement, personalization, collaboration, and communication. However, it can also hinder learning if not used appropriately, leading to distractions, misuse, or exacerbating equity and access disparities. Therefore, careful implementation, proper guidance, and addressing access issues are crucial to maximizing the benefits of technology in education.

- **Clarity of Viewpoint:** 85 (The response clearly outlines that technology's impact on education is not one-dimensional; it can both enhance and hinder learning, depending on various factors. The viewpoint is explicitly stated, with the response systematically addressing different aspects of technology's influence on learning.)
- **Logicity of Argumentation:** 90 (The argumentation is presented in a logical and structured manner, with each point building upon the previous one. The response enumerates specific aspects of technology's impact on education, providing a balanced view of its potential benefits and drawbacks.)
- **Correctness and Fidelity:** 80 (The response appears to be well-grounded in factual information, citing common knowledge about the role of technology in education. It avoids making unsubstantiated claims and presents a balanced view without apparent subjective biases.)
- **Comprehensiveness and Diversity:** 70 (The response covers a comprehensive range of topics, including accessibility, engagement, personalization, collaboration, and the potential negative impacts of technology such as distractions and equity issues. It caters to the perspectives of various stakeholders, including students, teachers, and those concerned with educational equity.)
- **Innovation and Uniqueness:** 60 (The insights provided are not particularly unique or innovative; they reflect common discussions surrounding technology in education. However, the response does a good job of summarizing these points, even if it does not break new ground.)
- **Depth and Essentially:** 65 (The response considering both the positive and negative impacts of technology on education, but do not delve into the essential aspects of the issue.)

Table 23: The response and detailed evaluation score of CoT

### Case Study for AI evaluation: NeoN

🗋️ **Question:** *Does technology in education enhance learning or hinder it?*

🗣️ **NeoN(ours):**

It is a multifaceted issue that requires a nuanced understanding of the various factors at play. While technology has the potential to revolutionize education by providing access to vast resources, promoting engagement, and enabling personalized learning experiences, it also presents challenges such as distractions, inequalities in access, and ethical dilemmas. One of the key considerations is the diverse perspectives and contextual variability in how technology impacts learning. The effectiveness of technology integration can vary based on factors such as infrastructure, resources, and student demographics. While technology can enhance learning opportunities for some students, it may introduce unintended consequences such as digital fatigue and reduced attention spans for others.

Moreover, the rapid evolution of technology necessitates continuous adaptation and professional development for educators to effectively integrate new tools into their teaching practices. Socioeconomic disparities in access to technology and reliable internet connectivity further exacerbate inequalities in education, highlighting the importance of addressing these disparities to ensure equitable access for all students.

Balancing technology with human-centered teaching approaches is essential to fostering holistic learning environments. While technology can provide valuable resources and interactive learning experiences, it should not overshadow the importance of human connections and personalized interactions in education.

Ethical considerations surrounding data privacy, algorithmic bias, and the ethical use of student data also pose significant challenges in the integration of technology in education. Safeguarding student privacy and promoting ethical practices in educational technology are critical for ensuring the responsible use of technology in learning environments.

In striving for continuous adaptation through critical reflection, equity-focused initiatives, professional development, and research-informed practices, educators can navigate the complexities and uncertainties surrounding technology in education to create more inclusive and effective learning environments. By addressing these challenges and leveraging the opportunities, we can work towards harnessing its full potential to enhance learning outcomes for all students.

- **Clarity of Viewpoint:** 85 (The response clearly states that the impact of technology on education is complex and multifaceted, and it acknowledges both the potential benefits and the challenges. The viewpoint is that technology has the potential to enhance education but also presents significant challenges that must be addressed.)
- **Logicity of Argumentation:** 90 (The argumentation is logical and coherent, moving from the potential benefits of technology in education to the challenges it presents. The response maintains a balanced perspective throughout and provides a structured argument with clear transitions between points.)
- **Correctness and Fidelity:** 90 (The response is grounded in realistic considerations regarding the use of technology in education, such as access disparities and the need for continuous educator development. It avoids unsubstantiated claims and subjective biases, sticking to commonly acknowledged issues in the field.)
- **Comprehensiveness and Diversity:** 90 (The response covers a broad range of issues related to technology in education, including access, engagement, personalized learning, ethical considerations, and the importance of human-centered teaching. It addresses the needs and interests of various stakeholders, such as students, educators, and policymakers.)
- **Innovation and Uniqueness:** 75 (While the response provides a thorough examination of the topic, it does not offer particularly unique insights or innovative solutions. The points made are well-established in discussions about technology in education.)
- **Depth and Essentially:** 80 (The response delves into the core issues surrounding technology in education, such as equity, continuous adaptation, and ethical practices. It goes beyond surface-level analysis to consider the underlying challenges and opportunities technology presents in learning environments.)

Table 24: The response and detailed evaluation score of NeoN