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

Anonymous ACL submission

Abstract

Large language models (LLMs) exhibit powerful reasoning capacity, as evidenced by prior studies focusing on objective topics such as arithmetic, symbolic, and commonsense reasoning tasks. However, the reasoning of objective topics emphasizes more on *logical thinking*, but cannot effectively reflect the *comprehensive, reflective, and creative thinking* that is also critical for overall reasoning prowess of LLMs. In light of this, we build a dataset SJTP comprising **S**ub**J**ective **T**o**P**ics spanning diverse types and fields, as well as three evaluation indicators to fully explore the reasoning ability of LLMs. It has been observed 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 **Ne**gation **o**f **N**egation (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 objective topics, including arithmetic reasoning (Luo et al., 2023; Yang et al., 2023), symbolic reasoning (Wei et al., 2022a), and commonsense reasoning (Geva et al., 2021a; Talmor et al., 2019a), etc.

Despite the remarkable reasoning capabilities, their evaluation still lacks comprehensiveness. Previous research primarily investigates LLMs based on objective topics with clear-out answers and logical reasoning path (*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 of LLMs 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.

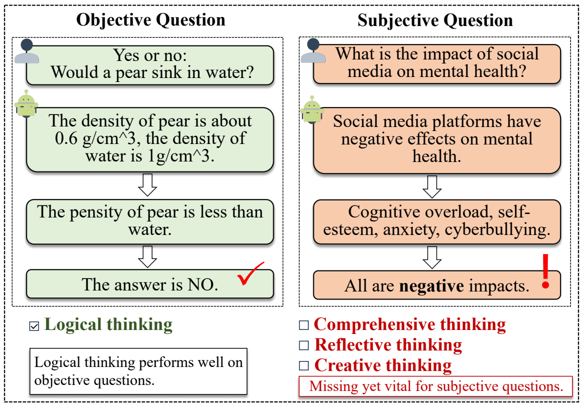

Figure 1: LLMs wich CoT prompting (Wei et al., 2022a) showcase strong logical thinking ability, which is adequate to solve objective questions. Meanwhile, it fails to solve subjective questions and shows insufficiency of comprehensive, reflective and creative thinking abilities.

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 to reflect these abilities, which consists of **S**ub**J**ective **T**o**P**ics of three types and eight fields as examples shown in Table 1. Furthermore, we design three evaluation indicators based on six evaluation dimensions to assess the responses of subjective questions. The evaluation di-

Topic Type	Example	SCR
Viewpoint Discourse	What do you think about the impact of social media on mental health?	76.5
Binary Dialectics	Should school wear uniforms?	69.9
Practical Analysis	How can the preservation of cultural heritage contribute to the promotion of social cohesion and identity ?	80.2
Topic Filed	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	
Evaluation Dimensions	① Clarity of Viewpoint, ② Logicality of Argumentation, ③ Correctness and Fidelity, ④ Comprehensiveness and Diversity, ⑤ Innovation and Uniqueness, ⑥ Depth and Essentially	

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

mension includes both objective aspects(e.g., clarity, logicality) and subjective aspects(e.g., reflection, creativity) as listed in Table 1 and 9. We then briefly evaluate gpt-3.5 armed with the promising CoT prompts, which showing limited performance on tacking subjective questions. 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 lack of depth and unbalanced analysis.

To elicit the language models' comprehensive, reflective and creative thinking abilities, we propose a structured framework inspired by the principle of the negation of negation, 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 is in line with the spirit of "abstraction-negation-concreteness" proposed by Hegel (Pinkard, 1988). Through constant negation, we can emphasize the multifaceted and complex nature of problems, and break established thinking patterns. Please not that if there are no defects or incompleteness in the initial topic, the "negation" stage is not necessary, which is equivalent to classical logical thinking.

Specifically, we proposed NeoN (Negation of negation), which casts an LLM as an agent of a negator instead of a logical reasoner. This agent will first determine whether to proceed with negation based on the similarity between the latest response and the existing ones. If so, the agent will be encouraged to generate unconsidered views, innovative ideas and deep insights by negating existing content. Note that the entire negation process is carried out upon maintaining correctness and fidelity. In the end, LLMs provides a refinement response by synthesizing all these answers. Furthermore, this method is built on top of the prompting technique, thus there is no trainable module. Technically, the proposed *NeoN* comprises three stages: straight answer, recurrent negation, and unification reasoning. To guarantee basic objective reasoning ability, we first allow LLMs to generate answers directly with strict logical thinking. Then, a negation link to the previous responses is constructed with a simple judgment under the correct premise, which has been shown helpful to supplement and surpass the original response in a spiral upward manner. Finally, we enable LLMs to engage in reasoning based on both the original question and the unification of the diversified responses.

We employ both API-based and open-source LLMs including GPT (OpenAI, 2023), InstructGPT (Ouyang et al., 2022), and LLaMA (Touvron et al., 2023), to validate the effectiveness of our framework. Experimental results show that *NeoN* leads to significant and consistent improvements under 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

In this section, we aim to explore the comprehensiveness, reflection, and creativity of LLMs by constructing a SJTP benchmark, which includes diverse and sophisticated subjective topics along with reasonable scoring points and solutions. Besides, we further elaborately develop three evaluation indicators to assess the quality of responses. The overall illustration is shown in Figure 2.

2.1 Question Generation

Topic Sampling. We first define and build a subjective topic pool that covers various types and

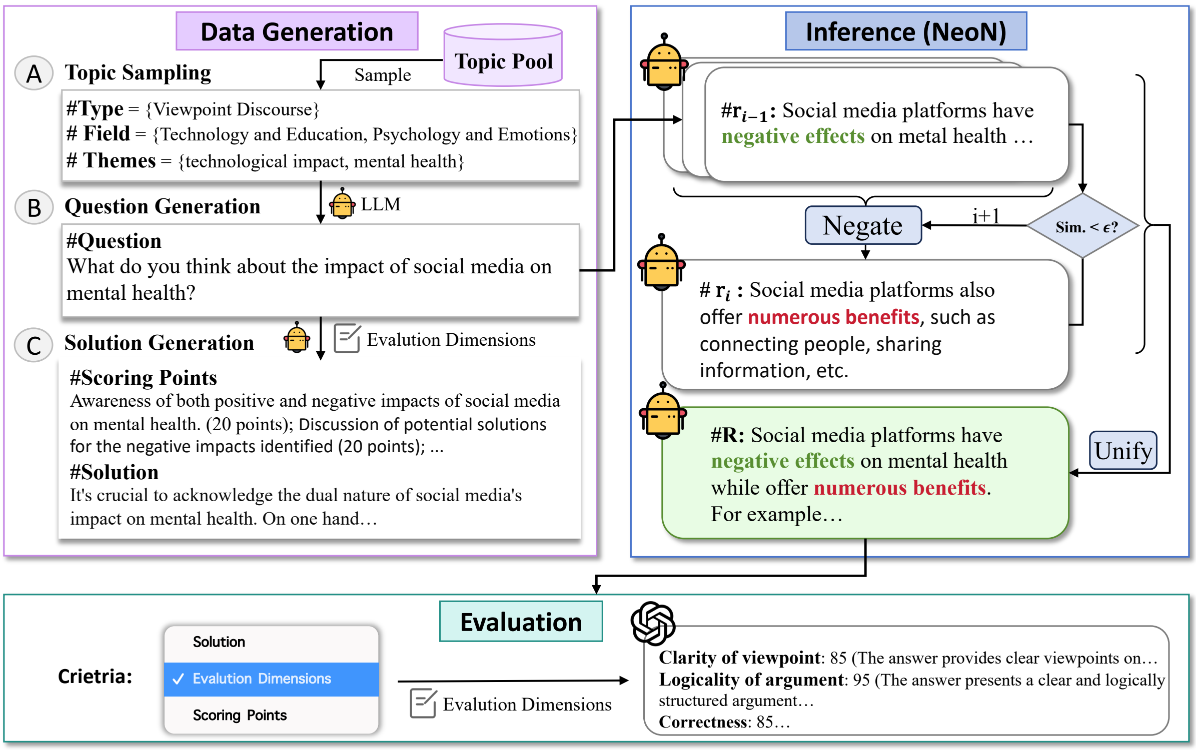

Figure 2: Illustration of the construction of SJTP dataset, the framework of NeoN, and the evaluation indicators.

fields that be widely discussed, as shown in Table 1. Specifically, we group the subjective topics into three representative types, i.e., *Viewpoint Discourse, Binary Dialectics, Practice Issues*, which can reflect the comprehensiveness, reflection and creativity of the respondents. Additionally, each type particularly emphasizes one of the thinking abilities, e.g., the *Practical Issues* is more reflective of creative thinking. Besides, we define eight topic fields covering a broad scope of knowledge such as social and technology. Each field contains subdivided related themes to improve the diversity of problem generation, see Appendix A.1 for details .

Given the topic pool, for each data, we randomly set the topic type and uniformly sample 1 or 2 themes from fields that with connection as seeds, and then generate corresponding type of subjective problems involving the themes by prompting LLM. This generation mechanism challenges the model to generate problems that join diverse topics while keeping the problems reasonable.

Question Generation. Specifically, we incorporate the sampled topic type, fields, and themes into the prompts and utilize LLM's excellent language comprehension ability to generate a specific question in a zero-shot manner. We do not include any exemplars or other manual interventions to avoid potential biases brought by the concepts within the exemplars and achieve a diverse generation.

2.2 Solution Generation

Evaluation dimensions. To generate qualitative solutions for the subjective topics, we first need to set standards for perfect answers. Considering both objective and subjective factors, we establish six evaluation dimensions and corresponding detailed requirements, as shown in Table 1 and Table 9.

Scoring Points & Solution Generation. Given the generated questions, we prompts an LLM to generate the corresponding scoring points according to the evaluation dimensions. Since LLM is enforced to consider the requirements of evaluation standards and the knowledge in different fields, it will generate clear, organized, and reasonable scoring points that more meet our expectations than directly answer the question. Furthermore, we generate the corresponding complete solutions given both the question and scoring points by prompting LLM, to serve as a more intuitive reference.

2.3 Evaluation Indicator

Given the inherent challenge in quantifying the quality of responses to subjective questions, we carefully craft three indicators to serve as robust measures: (1) SRC_{sol} takes the generated solution as standard, which assesses the semantic similarity between the response and the reference solution; (2) SRC_{point} takes the unique scoring points for each question as the standard, which measures to what extent the response meet the requirements of each scoring point; (2) SRC_{dim} takes the general evaluation dimensions as the standard, which assess the response from each dimension according to their detail standards respectively. The consistency presented by the three indicators in experiment results demonstrates their rationality.

As a result, the LLM needs to complete the following three tasks: (T1) Generating a subjective problem that relates to the given type and themes; (T2) Providing corresponding scoring points and reference solution of the subjective problem with the prefix evaluation requirements; (T3) Performing evaluation according to the three evaluation indicators. In this work, we use GPT-4 (OpenAI, 2023) as the LLM for generating and evaluating. See prompt templates in Appendix A.2 and C.1.

3 NeoN Framework

Our investigation into the reasoning performance of CoT prompting on the proposed SJTP dataset, as delineated in Table 1, reveals limitations of the current LLMs capacity for comprehensive, reflective, and creative thinking. To unleash these cognitive faculties in LLMs, we introduce the *NeoN* framework grounded on the principle of the negation of negation, allowing models to engage in a self-enhancing cycle of continuous improvement and transcendence. The overall framework of *NeoN* is schematically illustrated in Figure 2 and is characterized by two pivotal processes: iterative negation and integrative unification.

The *NeoN* framework first generates an initial solution through its inherent reasoning ability. Subsequently, it embarks on a process of negation, critiquing its earlier responses while maintaining a commitment to correctness and fidelity. This stage is designed to encourage the model to explore broader perspectives, uncover potential inconsistencies, engage in thorough analysis, and break away from established concepts. Finally, it obtains a refined response by methodically assimilating and integrating the preceding responses.

Formally, given a subjective question $\mathcal{Q}$, our goal is to let the LLM $\mathcal{M}$ solve the question $\mathcal{Q}$.

Step 1: Direct Response. We first let LLM $\mathcal{M}$ directly generate a reasonable response $\mathbf{r}_0$ according to the question $\mathcal{Q}$. Specifically, we have

$$\mathbf{r}_0 = \mathcal{M}(\mathcal{Q} \oplus \mathcal{P}_1), \quad (1)$$

where $\oplus$ denotes concatenation operation. $\mathcal{P}_1$ is a prompt serving as a trigger sentence, for example, we can set $\mathcal{P}_0$ as "*Let's generate the answer*".

Step 2: Negation of Negation. Then, building upon the principle of the negation of negation, i.e., things develop and progress in constant negation, we let LLM $\mathcal{M}$ constantly negate the previous responses, thereby facilitating the generation of novel and advanced content:

$$\mathbf{r}_n = \mathcal{M}(\mathcal{Q} \oplus \mathbf{r}_0 \oplus \cdots \oplus \mathbf{r}_{n-1} \oplus \mathcal{P}_2), \quad (2)$$

where n denotes the number of negation rounds. We terminate the negation process when the semantic similarity between the current response $\mathbf{r}_n$ and previous responses $\mathbf{r}_0 \oplus \cdots \oplus \mathbf{r}_{n-1}$ exceeds a threshold ϵ, which implies that $\mathbf{r}_n$ is approaching a state of refinement since $\mathcal{M}$'s diminishing capacity to yield additional novel insights. The value of n is usually between $2 \sim 3$ empirically. $\mathcal{P}_2$ is a prompt for making reasonable negation, e.g., we can set $\mathcal{P}_2$ as "*Negate the above responses to deduce a more perfect answer.*"

Step 3: Integration and Unification. Finally, we take question $\mathcal{Q}$ and all the responses as the input, letting LLM $\mathcal{M}$ give the final response $\mathcal{R}$:

$$\mathcal{R} = \mathcal{M}(\mathcal{Q} \oplus \mathbf{r}_0 \oplus \cdots \oplus \mathbf{r}_n \oplus \mathcal{P}_3), \quad (3)$$

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.*".

Remark In step 2, the negation process terminates when the current response presents a highly consistency with preceding responses. This design is predicated upon a claim by Hegel (Pinkard, 1988), which asserts 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 and consummate, further negation becomes futile, and reasoning is equivalent to logical reasoning."* In alignment with this philosophical foundation, by constantly negating old concepts, the model will be propelled to explore new perspectives, excavate deeper essence, and challenge established opinions, thereby fostering the evolution and refinement of its solution. Besides, we implement NeoN without any exemplars to avoid potential limitation, and without any restriction on data types.

4 Experiment

4.1 Setups

Datasets. We evaluate the performance of our framework on both subjective task and twelve ob-

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	78.96	81.76	50.82	75.72	80.25	71.14	81.35	72.74
Recite&Answer	66.01	68.15	76.70	82.50	51.27	76.19	78.30	69.34	80.76	71.89
NeoN (Ours)	**71.33**	**70.25**	**82.70**	**84.14**	**55.10**	**81.22**	**82.15**	**72.53**	**83.20**	**75.85**
(text-davinci-003)										
Direct Prompt	62.33	64.18	80.91	83.47	46.20	78.55	76.68	69.95	81.15	71.49
Zero-Shot-CoT	69.54	68.24	79.66	81.21	54.91	75.10	81.24	67.62	80.26	73.09
Self-Consistency	71.17	71.26	80.78	84.44	56.74	76.83	**84.63**	70.20	82.94	75.44
Recite&Answer	68.28	70.41	82.10	84.79	58.75	78.62	80.22	71.58	82.27	75.23
NeoN (Ours)	**77.66**	**75.43**	**86.30**	**89.87**	**63.70**	**82.05**	84.56	**73.43**	**85.35**	**79.81**
(gpt-3.5-turbo-1106)										
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	90.20	51.50	78.85	89.23	73.62	84.18	77.72
Recite&Answer	72.60	66.78	83.46	89.51	52.68	82.07	85.70	**77.86**	85.25	77.31
NeoN (Ours)	**80.40**	**81.82**	**88.67**	**92.50**	**60.33**	**83.21**	**89.15**	76.83	**87.74**	**82.29**

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

jective tasks, to fully validate the effectiveness of NeoN and the importance of subjective thinking for LLM's overall reasoning ability. For the subjective task, our proposed **SJTP** consists 500 data involving three types of subjective topics across eight fields. It aims to investigate LLM's subjective thinking abilities such as comprehensive, reflective and creative thinking. For objective tasks, we consider six **Arithmetic Reasoning** datasets, two **Commonsense Reasoning** datasets, two **Generic Reasoning** datasets and two **Symbolic Reasoning** datasets. See details in Appendix B.1.

Models. For LLMs, we evaluate on both API-based models including GPT-3.5 (OpenAI, 2023) and InstructGPT (Ouyang et al., 2022), and open-source model LLaMA-2-70B (Touvron et al., 2023). In particular, we use the released API version of gpt-3.5-turbo-1106 and text-davinci-003 by OpenAI. We set the decoding temperature as 0.5 to obtain the diversity of the responses generated by LLMs. For AI evaluation, we take the sota LLM gpt-4-1106-preview released by OpenAI as the evaluator.

Baselines. We compare NeoN mainly to zero-shot prompting methods to verify the effectiveness of its activation of subjective thinking abilities. Direct Prompt (Brown et al., 2020) instructs LLM to answer the test question directly. Zero-Shot-CoT (Kojima et al., 2022) appends the prompt "Let's think step by step" before reasoning. Self-consistency (Wang et al., 2022) first samples multiple solutions from the LLM with CoT technique and then take the majority vote as the final result.

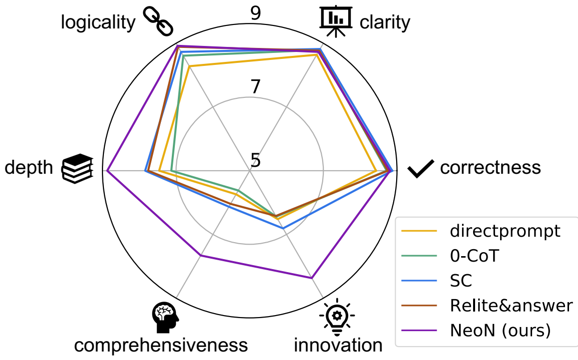

Figure 3: Categorized performance analysis for score in different evaluation dimensions and topic types.

Recite&Answer (Sun et al., 2023b) retrieves relevant passages first and then generates responses.

4.2 Main Results

Main Results on SJTP. In Table 2, we show the main results of the baselines and our *NeoN* method. Direct Prompt method directly respond to questions and gives the average score of 70.64∼74193 across different models, showing relatively limited performance for subjective topics. By explicitly prompting LLMs to "*Let's think step-by-step*" or "*recite relevant passages then give answers*", Zero-Shot-CoT and Recite&Answer surpass CoT in some cases thanks to their clear reasoning path or retrieved relevant information, but sometimes degrade possibly due to their limitation on diversity. Self-consistency generate multiple reasoning path via CoT prompt and make a synthesis that bene-

	Arithmetic					
	GSM8K	SVAMP	AQuA	MultiArith	SingleEq	AddSub
Direct Prompt	17.31	70.79	28.40	69.37	85.42	86.91
Zero-Shot-CoT	80.15	80.38	54.24	95.62	93.86	88.19
Self-Consistency	83.89	83.60	**63.39**	96.21	95.36	89.93
Recite&Answer	76.25	79.33	55.91	96.67	93.81	89.27
NeoN(Ours)	**84.17**	**85.67**	61.52	**96.94**	**95.67**	**93.09**

	Commonsense		Generic		Symbolic	
	CSQA	StrategyQA	Date Understand	Shuffled Objects	Last Letter	Coin Flip
Direct Prompt	72.25	62.33	46.83	32.98	2.46	53.54
Zero-Shot-CoT	70.80	62.08	64.44	69.66	70.52	89.59
Self-Consistency	73.16	63.76	73.27	72.08	86.30	97.52
Recite&Answer	71.81	61.34	70.34	72.76	71.34	87.93
NeoN(Ours)	**75.22**	**64.83**	**77.46**	**74.13**	**88.18**	**97.92**

Table 3: Main results of baseline methods and NeoN on twelve objective datasets (using gpt-3.5).

fits the comprehensiveness of response, which improves performance to a limited extent restricted by depth and innovation. Regarding our *NeoN* method which involves multi-turn of negation, it consistently surpasses all the baselines across models and topic types. Compared to the second-best ones, our method improves absolutely by $+3.11$, $+4.37$, and $+4.57$ by using LLaMa-2, InstructGPT, and GPT-3.5 model, respectively, which further demonstrates the effectiveness of negation in solving subjective tasks. Note that the consistency between the scores evaluated under the three evaluation indicators further demonstrates the rationality of our proposed evaluation indicators

To further verify that NeoN has unleashed the model's comprehensiveness, reflection and creativity, we visualize the SCR_{dim} score of methods in each evaluation dimension in Fig 3. It is clear that NeoN surpass all the baselines in the three subjective evaluation dimensions by a large margin, while maintaining the quality in terms of the three objective criteria, demonstrating the effectiveness of NeoN in activating the comprehensiveness, reflection and creativity of LLMs. Besides, NeoN achieve consistently improvement on all the three topic types, showing its universality.

Main Results on Objective Tasks. 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 twelve objective reasoning datasets. The results are shown in Table 3. It can be clearly observed that NeoN surpasses the baselines on 11 of 12 datasets, indicating that unleashing subjective thinking abilities also benefits objective reasoning process, which also need comprehensive

	AddSub		CSQA	
	F2T(%)	T2F(%)	F2T(%)	T2F(%)
Self-Consistency	18.42	0.79	13.33	2.91
NeoN (ours)	25.64	0.13	17.98	0.99

Table 4: The ratio to correct incorrect answers (F2T:True2False) and mislead models (T2F:True2False) of different methods on AddSub and CSQA dataset.

consideration, in-depth analysis, and constant reflection of the potential errors and omissions. We calculate the probability of NeoN correcting the initially incorrect answer and the proportion of misleading the original correct answer, as shown in Table 4, demonstrating that improve the subjective thinking abilities can alleviate the hallucination problem to a certain extent.

4.3 Analyses and Discussions

Distribution of Topic Fields in SJTP. We presents the proportion of data involving different topic fields in the proposed SJTP dataset, as shown in the Figure 4. We can see that SJTP covers 8 topic fields fairly evenly, in which 41% data involve two topic across two fileds. This shows the diversity of subjective topics and a wide range of knowledge covered in SJTP, which enables SJTP to be used as

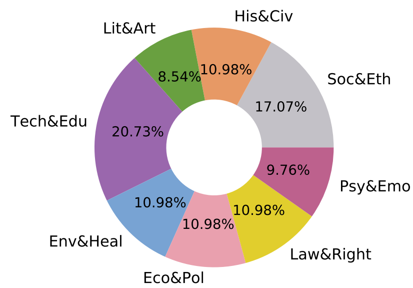

Figure 4: Different fields of subjective topics in SJTP.

Method	SCR_{Sol}	SCR_{point}	SCR_{dim}
NeoN (Ours)	**87.35**	**72.99**	**86.54**
w/o NG	↓ 2.33	↓ 2.67	↓ 2.69
w/o UF	↓ 1.62	↓ 1.78	↓ 1.01
w/o NG&UF	↓ 4.74	↓ 5.17	↓ 3.67

Table 5: Ablation study of *NeoN* method on SJTP with gpt-3.5. NG: negation process. UF: Unification process.

a benchmark dataset for measuring the subjective thinking abilities of LLMs.

Ablation study. We investigate the impact of negation process and unification process in our *NeoN* framework, as shown in Table 5. The full *NeoN* method, incorporating both steps, performs best on all indicators, highlighting the importance of both negation and unification. Removing negation process (*w/o* NG) (i.e., generate n response randomly without explicit negation prompts) generally leads to a performance drop by around 2 score, showing that it is crucial for constant negation to generate new insights. Removing unification process (*w/o* UF) (i.e., take the last turn of response as the final answer) decreases the score by around 1, indicating that the analysis and synthesis of all contents benefit the refinement of response. Excluding both steps (*w/o* NG&UF) leads to the worst performance. In summary, both the steps of negation and unification are important and the best performance is achieved when both of them are utilized.

Rounds of Negation. We do not fix the number of negation round and the termination depends on when the current response with a high semantic similarity with the previous responses. For subjective reasoning, we evaluate the similarity by gpt-4 and set the termination threshold $\epsilon = 90$. For objective tasks that have definite answers, the similarity means whether the current answer is equal to the previous one. The average rounds of negation on different reasoning tasks are shown in Table 6, which are all around 2∼3, indicating that complement contents can be quickly induced by negation.

	View. Dis.	Bi-Dial.	Prac. Iss.	Arith.
#Rounds	2.87	3.14	2.68	1.72
	ComSens.	**Generic**	**Symbol.**	**AVG.**
#Rounds	2.72	2.30	1.46	2.41

Table 6: The average numbers or negation rounds of *NeoN* in different tasks.

Then we specify the rounds of negation as 1∼5 to explore its effect on performance, the results are show in Figure 5a. We can see that as the rounds

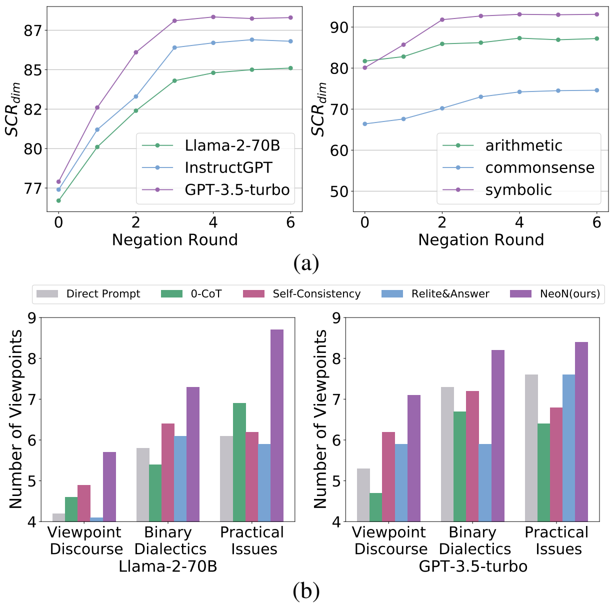

Figure 5: (a): Impact of the rounds of negation across models and reasoning types. (b): Comparison of the number of viewpoints in different methods.

of negation increases, the performance of model will presenting an overall improvement, but when the turn increases qround 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) our method only stimulates the potential performance that has not been released as much as possible by negation, rather than improving the essential reasoning ability. Therefore, when stimulated to a certain extent, the improvement will also reach a bottleneck.

Case Study. To better demonstrate the SJTP dataset we constructed, the effectiveness of our NeoN method, and the evaluation flow, we conduct multiple study cases involving different topic types and topic fields. We can see the examples of data in SJTP from Table 11 to Table 13 in Appendix A.3. These cases clearly presented the data format and content in SJTP. For the responses of NeoN, we also present three cases including subjective reasoning, commonsense reasoning and arithmetic reasoning from Table 15 to Table 17 in Appendix B.2. The responses of each step show that NeoN can indeed explore new perspectives and thorough insights by negation, and can also rectify the wrong reasoning path via reflection. We also provide some evaluation examples from Table 21 to Table 22 in Appendix C.2, to present the details of AI evaluation and be more convinced. All these study cases demonstrate our rationality and effectiveness.

Quality Comparison. To deeply analyze the

	AI Eval.	HU Eval.	Cohen's κ
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.

improvement of comprehensiveness by NeoN, we count the average number of perspectives involved in responses of different methods, as shown in Figure 5b. We can see that NeoN considers the most aspects, showing its comprehensiveness.

Human Evaluation. To further validate the reliability of AI evaluation indicators, we compared whether AI and human choose NeoN as the best on 100 samples. The cohen's κ shown in Table 7 indicates a relatively high consistency between our evaluators and human evaluation, validating the rationality of our evaluation indicators.

5 Related Work

5.1 Chain of thought Style Prompting

CoT prompting (Wei et al., 2022c) and its variants (Kojima et al., 2022; Zhang et al., 2023; Sun et al., 2023b) are widely used in augmenting the reasoning abilities of LLMs. These methods attempt to guide the models to think in a step-by-step manner by introducing reasoning path, which assists the model in logical thinking and has been shown effective for various objective tasks. Self-consistency (Wang et al., 2022) extend the single-reasoning path to multi-path by sampling multiple solutions with cot prompt independently and the take the majority as the final solution. Another line of work proposes reasoning leveraging relevant passages (Sun et al., 2023b) or expert modeling (Xu et al., 2023). However, we show in experiments that these methods have limited performance in subjective tasks since they rely more on the logical thinking. Different from them, this paper aims to explore and further unleash the comprehensive, reflective and creative thinking abilities of LLMs.

5.2 Subjective Tasks

Compared with objective tasks that have a clear solution or evaluation criteria that reflect logical thinking, subjective tasks have open and non-standard answers that can test more comprehensive thinking abilities (Kanclerz et al., 2023; Sun et al., 2023a). Traditional NLP studies have explored some subjective tasks including linguistic rhetoric, disambiguation, stance detection, etc. (Jentzsch and Kersting, 2023; Mao et al., 2023). Nevertheless, these tasks not only have been fewly explored in LLM researches, but also involve interpretation, judgment, and personal experiences (Rottger et al., 2022; Sun et al., 2023a) which emphasize on the ability to perceive context, language nuances, and emotions. Instead, the proposed SJTP consists of subjective topics across various types and fields, which can better evaluate the comprehensive, reflective and creative thinking abilities. Besides, the unsatisfactory performance of LLMs on SJTP suggests the challenge faced in subjective tasks and the significance of releasing the relevant thinking abilities.

5.3 Debate-based Reasoning

There have been recent studies concern the debate between LLMs. A line of work aim to explore the limitations of LLMs in simulating human interactions(Taubenfeld et al., 2024), the inter-consistency among multiple LLMs for collaboration(Wang et al., 2023), or the ability to resist misleading erroneous arguments and defend the truth (Xiong et al., 2023) by simulating debates. Furthermore, multi-agent debate framework is introduce for text evaluation (Chan et al., 2023) and mathematical and strategic reasoning(Liang et al., 2023; Du et al., 2023). In contrast, we aim to unleash the comprehensive, reflective and creative thinking abilities of LLM by conducting negation process on only one agent, which can be applied in both objective and subjective topics. Note that our framework needs no restriction on data types and preset debate positions, possessing more convenient and flexibility.

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.

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

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Proc. of NeurIPS*.

Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. 2023. Chateval: Towards better llm-based evaluators through multi-agent debate. *arXiv preprint arXiv:2308.07201*.

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. 2021a. Did aristotle use a laptop? a question answering benchmark with implicit reasoning strategies. *Transactions of the Association for Computational Linguistics*, 9:346–361.

Mor Geva, Daniel Khashabi, Elad Segal, Tushar Khot, Dan Roth, and Jonathan Berant. 2021b. 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 Jentzsch 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*.

Kamil Kanclerz, Konrad Karanowski, Julita Bielaniewicz, Marcin Gruza, Piotr Miłkowski, Jan Kocon, and Przemyslaw 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, Hannaneh Hajishirzi, Ashish Sabharwal, Oren Etzioni, and Siena Dumas Ang. 2015. Parsing algebraic word problems into equations. *TACL*, 3:585–597.

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*.

Wang Ling, Dani Yogatama, Chris Dyer, and Phil Blunsom. 2017. Program induction by rationale generation: Learning to solve and explain algebraic word problems. In *Proceedings of ACL*, pages 158–167.

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*.

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.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Proc. of NeurIPS*.

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*.

Subhro Roy and Dan Roth. 2015. Solving general arithmetic word problems. In *Proceedings of EMNLP*, pages 1743–1752.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià 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 Jurgens. 2023a. Aligning with whom? large language models have gender and racial biases in subjective nlp tasks. *arXiv preprint arXiv:2311.09730*.

Zhiqing Sun, Xuezhi Wang, Yi Tay, Yiming Yang, and Denny Zhou. 2023b. Recitation-augmented language models. In *Proc. of ICLR*.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019a. CommonsenseQA: A question answering challenge targeting commonsense knowledge. In *Proc. of NAACL*.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019b. Commonsenseqa: A question answering challenge targeting commonsense knowledge. In *Proceedings of NAACL-HLT*, pages 4149–4158.

Amir Taubenfeld, Yaniv Dover, Roi Reichart, and Ariel Goldstein. 2024. Systematic biases in llm simulations of debates. *arXiv preprint arXiv:2402.04049*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Boshi Wang, Xiang Yue, and Huan Sun. 2023. Can chatgpt defend its belief in truth? evaluating llm reasoning via debate. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 11865–11881.

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*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022c. Chain-of-thought prompting elicits reasoning in large language models. In *Proc. of NeurIPS*.

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.

Benfeng Xu, An Yang, Junyang Lin, Quan Wang, Chang Zhou, Yongdong Zhang, and Zhendong Mao. 2023. ExpertPrompting: Instructing large language models to be distinguished experts. *arXiv preprint arXiv:2305.14688*.

Zhen Yang, Ming Ding, Qingsong Lv, Zhihuan Jiang, Zehai He, Yuyi Guo, Jinfeng Bai, and Jie Tang. 2023. Gpt can solve mathematical problems without a calculator. *arXiv preprint arXiv:2309.03241*.

Zhuosheng Zhang, Aston Zhang, Mu Li, and Alex Smola. 2023. Automatic chain of thought prompting in large language models. In *Proc. of ICLR*.

A Details for SJTP Construction

A.1 Topic Pool and Evaluation Standard

Table 8 shows the topic fields in SJTP. Different themes are grouped into multiple fields mainly according to the scope of knowledge. Tables 9 demonstrates the detailed standards for each evaluation dimension.

A.2 Prompt Templates for Data Generation

Table 10 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.3 Case Study for SJTP

We show typical cases of different data types in SJTP in the following Tables 11, 12 and 13, 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 14 for the detailed description of each datasets.

For arithmetic reasoning, we consider the following six datasets: (1) SingleEq (Koncel-Kedziorski et al., 2015), (2) AddSub (Hosseini et al., 2014), (3) MultiArith (Roy and Roth, 2015), (4) AQUA-RAT (Ling et al., 2017), (5) GSM8K (Cobbe et al., 2021), and (6) SVAMP (Patel et al., 2021). The first three are from the classic Math World Problem Repository (Koncel-Kedziorski et al., 2016), and the last three are from more recent benchmarks. SingleEq and AddSub contain easier problems, which do not require multi-step calculation to solve the tasks. MultiArith, AQUA-RAT, GSM8k, and SVAMP are more challenging datasets that require multi-step reasoning to solve.

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

For symbolic reasoning, we use Last Letter Concatenation and Coin Flip (Wei et al., 2022b). Last letter Concatenation asks the model to concatenate the last letters of each word. We used randomly selected four names for each sample. Coin Flip asks the model to answer whether a coin is still heads up after people either flip or do not flip the coin. We created samples of four times flip or not flip trials. Although these tasks are easy for humans, LMs typically exhibit a flat scaling curve.

For other logical reasoning tasks, we choose two evaluation sets from the BIG-bench effort (Srivastava et al., 2022): Date Understanding [1] 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 similarity, 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 15, 16 and 17, 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.

[1] 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 8: 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.
Logicality 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 9: Detailed standards of each evaluation dimension.

C Details for Evaluation

C.1 Prompt Templates for AI Evaluation

Table 18, 19 and 20 demonstrate the prompt templates for AI evalution 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 evaluaition dimensions, respectively.

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 21 and 22. The results show our effectiveness and the rationality our our evaluation strategies.

Table 10: 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 “#Quesion:” 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 compelte 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: ...

Table 11: A viewpoint discourse data generated by SJTP.

Question Type: Viewpoint Discourse.

Topic Field(s): History and Civilization&Literature and Arts

Keyword(s): Art Appreciation& Cultural Identity

Question
How do literature and arts reflect and shape cultural identity?

Scoring Points
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)

Solution
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. Similarly, 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. Similarly, 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 logicality, 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.

Table 12: A binary dialectic data generated by SJTP.

Question Type: Binary Dialectics.

Topic Field(s): Technology and Education

Keyword(s): Technological Impact & Adolescent Development

Question

Does technology in education enhance learning or hinder it?

Scoring Points

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)

Solution

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 13: A practical issue data generated by SJTP.

Question Type: Practical Strategies.

Topic Field(s): Environment and Health

Keyword(s): Environmental Protection

Question
What strategies can be implemented to ensure the ethical treatment of animals in urban environments?

Scoring Points
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)

Solution
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 14: 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
SingleEq	N	508	27.4	questions.json	No License
AddSub	N	395	31.5	AddSub.json	Unspecified
MultiArith	N	600	31.8	MultiArith.json	Unspecified
GSM8K	N	1319	46.9	test.jsonl	MIT License
AQUA-RAT	M	254	51.9	test.jsonl	Apache-2.0
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
Last Letters	F	500	15.0	-	-
Coin Flip	Y	500	37.0	-	-

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

Question: *Does technology in education enhance learning or hindering 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 15: 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: Dan = 42 + 47 = 89 turnips. Now, let's calculate the total number of turnips Dan and Jessica grew together: Dan + 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 16: 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 17: Case study for NeoN on commonsense reasoning task.

Table 18: Prompt templates for the evaluation according to reference solution.

Prompt Templates for AI Evaluation (SRC_{sol})
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"**: "score": "[content]"

Table 19: Prompt templates for the evaluation according to scoring points.

Prompt Templates for AI Evaluation (SRC_{point})
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 **each scoring point**. 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 []: "**scoring point 1"**: "reason": "[your rationale]", "score": "[score from 1 to max_1]" "**scoring point 2"**: "reason": "[your rationale]", "score": "[score from 1 to max_2]" "**scoring point 3"**: "reason": "[your rationale]", "score": "[score from 1 to max_3]" …

Table 20: Prompt templates for the evaluation according to evaluation dimensions.

Prompt Templates for AI Evaluation (SRC_{dim})
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, Logicality, Correctness, Comprehensiveness, Innovation and Depth.
Question: []
Response:[]
Evaluate Aspects: 1.**[Clarity of Viewpoint]**: Evaluate the clarity and explicitness of the viewpoint presented in the response. 2.**[Logicality 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. 3.**[Correctness and Fidelity]**: Evaluate the correctness and fidelity of the response, ensuring it is grounded in factual information and data while avoiding subjective biases. 4.**[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. 5.**[Innovation and Uniqueness]**: Evaluate whether the response offers unique insights or innovative viewpoints, demonstrating the ability to approach the problem from fresh angles. 6.**[Depth and Essentially]**: 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 10 individually on **each aspect**. 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 []: "**Clarity of Viewpoint"**: "reason": "[your rationale]", "score": "[score from 1 to 100]" **"Logicality of Argumentation"**: "reason": "[your rationale]", "score": "[score from 1 to 100]" **"Correctness and Fidelity"**: "reason": "[your rationale]", "score": "[score from 1 to 100]" **"Comprehensiveness and Diversity"**: "reason": "[your rationale]", "score": "[score from 1 to 100]" **"Innovation and Uniqueness"**: "reason": "[your rationale]", "score": "[score from 1 to 100]" **"Depth and Essentially"**: "reason": "[your rationale]", "score": "[score from 1 to 100]"

Case Study for AI evaluation: CoT

Question: *Does technology in education enhance learning or hindering 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.)
- Logicality 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 21: The response and detailed evaluation score of CoT

Case Study for AI evaluation: NeoN

Question: *Does technology in education enhance learning or hindering 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: 8 (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.)
- Logicality of Argumentation: 9 (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: 9 (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: 9 (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: 7 (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: 8 (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 22: The response and detailed evaluation score of NeoN