EoT: Evolution of Thoughts for Complex Reasoning Tasks

Qin Hua[1,2], Jiaqi Sun[1,2], Shiyou Qian[1,*], Dingyu Yang[3], Jian Cao[1], Guangtao Xue[1]
[1] Shanghai Jiao Tong University
[2] Alibaba Group
[3] The State Key Laboratory of Blockchain and Data Security, Zhejiang University
huaqin@sjtu.edu.cn, jotaro@sjtu.edu.cn, qshiyou@sjtu.edu.cn,
yangdingyu@zju.edu.cn, cao-jian@sjtu.edu.cn, gt_xue@sjtu.edu.cn

Abstract

Knowledge-based complex reasoning remains a significant challenge for large language models (LLMs) with in-context learning. To tackle this issue, previous studies focus on ensuring behavior fidelity, factuality, or reliability in generated reasoning processes that guide LLMs to produce solutions. However, these studies often neglect the simultaneous optimization on all these three aspects for each thought. The main challenges are the lack of comprehensive assessment mechanisms and the difficulty of efficient thought-level optimization. This paper introduces the Evolution of Thoughts (EoT) framework, which enhances the factuality, fidelity, and reliability of each thought in the reasoning process through a few LLM inferences. We propose a thought assessment method that is sensitive to knowledge and LLM behaviors, using three scorers to evaluate each thought by considering domain context, semantic alignment, and behavior impact. Additionally, we establish a self-reflective evolution mechanism to facilitate each reasoning process generation in a single-forward inference. Extensive experiments demonstrate that, for knowledge-based complex tasks, EoT improves the factuality and fidelity of reasoning processes by approximately 16.5% and 48.8%, respectively, while enhancing LLM reasoning capability by about 6.2%, outperforming advanced approaches. [1]

1 Introduction

Large language models (LLMs), exemplified by the GPT (Achiam et al., 2024) and DeepSeek (DeepSeek-AI, 2025) series, have achieved remarkable success in various natural language processing (NLP) tasks. These models often employ in-context learning (ICL) schemes, enabling them to learn from contextual examples without updating billions of parameters (Brown et al., 2020). However, complex reasoning tasks requiring the comprehension of long-context knowledge and the generation of intricate solutions remain challenging for LLMs with ICL prompting (Chen et al., 2024).

Many studies have shown that guiding LLMs through step-by-step thought prompts significantly enhances their reasoning capabilities (Chen et al., 2024; Lyu et al., 2023). These prompts inspire LLMs to create reasoning processes that explain their behaviors and aid in task resolution. Each reasoning process, such as the chain of thoughts (CoT) and its variants (Wang et al., 2023), comprises a sequence of coherent text units known as thoughts, which serve as intermediate reasoning steps. In these mechanisms, the reasoning capability of LLMs is often reflected in the reliability of reasoning processes, which assesses the confidence or correctness of the solutions produced by LLMs under the guidance of reasoning processes (Zhang et al., 2024; Madaan et al., 2023). Therefore, many efforts have been dedicated to exploring logical reasoning processes with high reliability (Radhakrishnan et al., 2023; Besta et al., 2024).

Existing studies on optimizing reasoning processes can be categorized into three main areas. Firstly, some studies focus on enhancing behavior fidelity (Chuang et al., 2024; Lyu et al., 2023). Previous research (Liang et al., 2024) indicates that LLMs often provide reasoning processes that differ significantly from their actual reasoning behaviors, which compromises their reasoning capabilities. This discrepancy arises from the lack of awareness regarding the internal knowledge state in black-box LLMs. To address this issue, xLLM (Chuang et al., 2024) revises reasoning processes through evolutionary iterations to improve their behavior fidelity. Secondly, some studies address non-factual errors in reasoning processes to mitigate the hallucination phenomenon in LLMs. For example, Ye et al. (Ye

*Corresponding author

[1]The code of EoT and the case of experiment data can be found at https://github.com/citsjtu2020/EoT.git and https://github.com/citsjtu2020/EoT_data.git respectively.

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 23864–23886
November 4-9, 2025

Table 1: Comparison of existing studies

Framework	Reasoning Process		Optimization & Assessment			
	Components	Generation	Factors			Level
			Reli.	Fact.	Fide.	
BoT (Chen et al., 2024)	ToT	ITE	✓	✗	✗	Thought
GoT (Besta et al., 2024)	GoT	ITE	✓	✗	✗	Thought
CoT-dec (Radhakrishnan et al., 2023)	$SQAT$	SFI	✓	✗	✗	CRP
Ye et al. (Ye and Durrett, 2022)	CoT	SFI	✗	✓	✓	CRP
xLLM (Chuang et al., 2024)	CoT	SFI	✗	✗	✓	CRP
EoT (ours)	CoT	SFI	✓	✓	✓	Thought

[1] CoT, ToT, GoT and $SQAT$ stand for chain of thoughts, tree of thoughts, graph of thoughts, and sub question-answer as thoughts, respectively. Reli., Fact. and Fide. stand for the factors of Reliability, Factuality and Fidelity respectively.
[2] ITE, SFI and CRP stand for iterations of each thought exploration, single forward inference, and complete reasoning process, respectively.

and Durrett, 2022) extract the most factual reasoning process in each generation. Thirdly, many studies directly explore reasoning processes with high reliability (Madaan et al., 2023; Besta et al., 2024). For instance, BoT (Chen et al., 2024) uses an iterative method to explore many trees of thoughts, and accumulates trial-and-error experiences to derive a reasoning process yielding a reliable answer.

Nevertheless, enhancing the reasoning process for complex tasks presents three major challenges. multi-objective optimization regarding factuality, fidelity, and reliability has been a long-standing issue, as highlighted by previous research (Ye and Durrett, 2022; Liang et al., 2024). Secondly, the complexity of long-textual tasks exacerbates the difficulty in optimizing reasoning thoughts. These tasks require detailed knowledge across domains and expect complex solutions, it necessitates a reasoning process with multiple thought steps, involving intricate logic and precise extraction of relevant facts. This complexity hinders the assessment and improvement of the effectiveness of each thought. Thirdly, achieving thought-level optimization for reasoning processes while maintaining efficiency is challenging. Some studies (Chen et al., 2024; Radhakrishnan et al., 2023) improve thought quality by generating individual reasoning thoughts through extensive explorations, which incurs significant overhead. Conversely, studies like (Chuang et al., 2024; Ye and Durrett, 2022) strive to produce an improved reasoning process using a single LLM inference, it enhances time efficiency but compromises the fine-grained optimization for each thought step.

To address these challenges, we propose EoT, a framework that evolves reasoning processes to achieve multi-objective optimization in factuality, fidelity, and reliability. Firstly, EoT includes an assessment mechanism designed for complex reasoning tasks, which evaluates each thought in reasoning processes from all three perspectives. Secondly, we introduce a prompting mechanism to facilitate the creation of evolved reasoning processes with a single-forward LLM inference. This enables LLMs to comprehend thought assessment outcomes and ensure collaborative optimization for each thought within a few rounds of self-reflective evolution.

As highlighted in Table 1, EoT distinguishes itself from existing studies with two key innovations. First, EoT evaluates reasoning processes more comprehensively in three critical dimensions, achieving collaborative optimization across them. Second, EoT effectively manages both time efficiency and thought-level optimization, producing a complete reasoning process in each iteration through single forward inference and ensuring optimization at each thought within the reasoning processes.

We conducted extensive experiments on two datasets, including one with 40 production operational maintenance tasks. Compared to five advanced frameworks, EoT improves reliability, factuality, and fidelity of reasoning processes by about 6.2%, 16.5%, and 48.8% respectively. The contributions of this work are summarized as follows:

- We consider reliability, factuality, and fidelity of reasoning processes to enhance LLMs' reasoning capabilities, and assess each thought in reasoning processes based on these factors.
- We propose an evolving mechanism that efficiently facilitates multi-objective optimization at the thought level via single-forward generation of the reasoning process in each iteration.
- We evaluate the effectiveness of EoT on a real production dataset and verify its generality on the LongBench dataset including questions obtained from diverse fields, such as MultifieldQA and HotpotQA. Each question involves a context with thousands of tokens.

2 Problem Setup

This section formulates reasoning processes, introduces three key factors that impact LLMs' reason-

ing capabilities, and defines the evolution problem.

2.1 Formalization of Reasoning Process

We first formalize LLM-generated reasoning processes used for solving complex tasks.

Let $Q = (q_1, q_2, \cdots)$, $X = (x_1, x_2, \cdots)$ and $A = (a_1, a_2, \cdots)$ represent a question description, knowledge context and a reference answer for any complex task. Each component consists of sequential statements, denoted by q_i, x_i, and a_i respectively. When an LLM p_θ receives a task (X, Q), it generates a reasoning process R to explain its reasoning behaviors. Using R, the LLM guides itself to produce a solution $\hat{Y}$ for the task, formalized as:

$$\hat{Y} = p_\theta(X, Q|R) \tag{1}$$

A generated reasoning process R comprises multiple thoughts, denoted as $R = \{T_1, T_2, \cdots T_n\}$. Each thought includes several statements, i.e., $T_i = (t_{i,1}, t_{i,2}, \cdots)$. This study explores an evolved reasoning process R^* for each task (X, Q), to enhance the reasoning ability of p_θ. The guidance of R to produce $\hat{Y}$, as depicted in Eq.(1), is realized through instructions to prompt LLM to reason based on thoughts in R, as shown in Figure 4 in Appendix B. Figures 5 and 7 in Appendix B presents examples of question-solving with thoughts.

2.2 Reliability, Factuality and Fidelity of Thoughts

Next, we outline and define three key factors of reasoning processes that impact LLM's reasoning capabilities. To aid comprehension, we illustrate them with examples in Appendix C.

Previous studies (Chen et al., 2024; Zheng et al., 2023; Paul et al., 2023) have revealed that the reliability of reasoning processes is essential in evaluating LLMs' question-solving abilities. To elucidate, we define the reliability of a reasoning process R:

Definition 1: Reliability of R is measured by the similarity between the answer from LLMs guided by R and the reference answer. Greater similarity indicates higher reliability. Given a task, a generated answer $\hat{Y}$ guided by R, and reference answer A, the reliability of R is defined as:

$$Reliability(R) = Similarity(\hat{Y}, A) \tag{2}$$

The similarity metric can be computed using various methods, such as token overlaps (Lin, 2004), learning-based distance (Sellam et al., 2020), and entailment extent in natural language inference (NLI) (Gao et al., 2023). Intuitively, Figure 6 and Figure 8 in Appendix C illustrate reasoning processes with different levels of reliability.

As outlined in Section 1, enhancing behavior fidelity and reducing hallucinations of reasoning processes can improve LLMs' reasoning capabilities. Inspired by prior studies (Ye and Durrett, 2022; Chuang et al., 2024), we define two key factors of reasoning processes, factuality and fidelity:

Definition 2: Factuality pertains to how well thoughts in reasoning processes are grounded in the relevant knowledge context. As shown in Figure 6 in Appendix C, a fully factual reasoning process excludes hallucinations that contradict the context. Conversely, a non-factual reasoning process with hallucinations can lead to erroneous solutions. Let $T_i = (t_{i,1}, t_{i,2}, \cdots, t_{i,J})$ represent a thought step consisting of J statements in R for question Q with context X. The factuality of T_i is expressed as:

$$Factual(T_i) = \frac{\sum_{t_{i,j} \in T_i} Ground(X, t_{i,j})}{J}$$
$$Ground(X, t_{i,j}) = \begin{cases} 1 & t_{i,j} \text{ is grounded in } X \\ 0 & Otherwise \end{cases} \tag{3}$$

Definition 3: Fidelity evaluates the faithfulness of thoughts in the reasoning process to explain the actual behaviors of the LLM when generating answers. Based on previous studies (Lopardo et al., 2023; Chuang et al., 2024), fidelity is defined by the extent to which an explained reasoning thought influences the LLM-generated answers, assuming ground-truth reasoning behaviors are typically available. A greater degree indicates higher fidelity. Specifically, for a reasoning process R, the fidelity of each thought $T_i \in R$ is assessed by comparing the answers generated by the LLM guided by R with and without T_i. This is expressed as:

$$Fidelity(T_i) = Diff(p_\theta((X, Q)|R), p_\theta((X, Q)|(R \backslash T_i))) \tag{4}$$

where $(R \backslash T_i)$ represents the reasoning process without thought T_i, and $Diff(\cdot, \cdot)$ denotes the difference estimation. Appendix C.2 illustrates thoughts with high and low fidelity.

In summary, EoT guides LLMs to find a refined n-step reasoning process $R^* = (T_1^*, T_2^*, ..., T_n^*)$ via iterative self-reflecting evolution. In each iteration, the LLM p_θ generates an enhanced reasoning process to address multiple-objective optimization

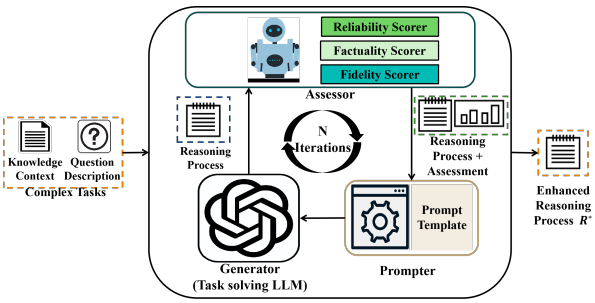


Figure 1: The framework of EoT.

in reliability, factuality and fidelity for each reasoning thought, as defined in Eq. (5):

$$\begin{aligned} &R^* \leftarrow \arg\max_{R} \ Reliability(R) \\ &maxmize\ Factual(T_i^*) \text{ for } \forall T_i^* \in R^* \\ &maxmize\ Fidelity(T_i^*) \text{ for } \forall T_i^* \in R^* \end{aligned} \tag{5}$$

3 Evolution of Thoughts

3.1 Overview

This section introduces the Evolution of Thoughts (EoT) framework to enhance reasoning processes. EoT refines thoughts evolutionarily to simultaneously improve their reliability, factuality, and behavior fidelity, thereby boosting LLMs' reasoning capabilities. Each iteration of evolution enables LLMs to comprehend the assessment of the current reasoning process and provides feedback to refine a group of individual thoughts in the reasoning process via a single-forward inference. In this way, EoT effectively and efficiently guides reasoning process optimization toward multiple objectives.

Figure 1 illustrates the framework of EoT, comprising three modules: an assessor, a prompter and a generator. These modules collaborate to achieve multi-objective optimization of reasoning processes over iterations of evolution. Initially, the assessor uses three scorers to evaluate the reliability, fidelity and factuality of thoughts in the reasoning process during each iteration. The prompter then guides LLMs to thoroughly comprehend the performance of the current reasoning process across the three aspects, prompting self-reflection in LLMs to ensure thought-level optimization. Lastly, in each iteration, the generator served by task-solving LLMs, uses our crafted prompts to generate a refined reasoning process through a single-forward inference. After N iterations, EoT produces an improved reasoning process, denoted as R^*. Further details of the three modules are provided below.

3.2 Assessor

EoT focuses on optimizing the factuality, reliability and fidelity of reasoning processes. Therefore, we design three scorers in the assessor to quantify the performance of thoughts across these three aspects.

3.2.1 Factuality Scorer

As defined in Eq.(3), the factuality of reasoning process for complex tasks is expressed by how well the thoughts in it can be supported by relevant domain-knowledge facts. For tasks with long-textual context, EoT leverages the impressive capability of LLMs in natural language inference (NLI) to tackle the scoring of factuality. We employ an NLI model ϕ to estimate $Ground(\cdot,\cdot)$ in the factuality definition in Eq. (3). Specifically, given a pair of premise and hypothesis statements, represented as x^{pre} and t^{hyp} respectively, ϕ infers their relationship through a triple-label classification task:

$$\phi(x^{pre}, t^{hyp}) = \begin{cases} 0 & t^{hyp} \textbf{ contradicts } \text{with } x^{pre} \\ 1 & x^{pre} \textbf{ entails } t^{hyp} \\ 2 & x^{pre} \text{ is } \textbf{neutral } \text{with } t^{hyp} \end{cases} \tag{6}$$

Definitions of "**Entailment**", "**Contradiction**" and "**Neutral**" are provided in Appendix E.1. Inspired by prior work (Gao et al., 2023), consider a reasoning process $R = (T_1, T_2, \cdots)$ for task (X, Q), where each thought $T_i = (t_{i,1}, t_{i,2}, \cdots, t_{i,J})$ contains J statements. For $\forall t_{i,j} \in T_i$, we define $Ground(X, t_{i,j}) = 1$ iff $\phi(X, t_{i,j}) = 1$; otherwise, it is 0. We use a GPT-4 model with few-shot learning for ϕ. The efficacy of this NLI scheme is presented in Appendix E.2. Then, the factuality score $S_{fac}(T_i)$ for $\forall T_i \in R$ is computed as:

$$S_{fac}(T_i) = Factual(T_i) \tag{7}$$

3.2.2 Reliability Scorer

According to Eq.(2), assessing the reliability of the reasoning process R largely relies on measuring the similarity between the answer generated under R and the reference answer. However, existing similarity metrics struggle to ensure both sufficient variation sensitivity and robust human-level alignment simultaneously. To address these issues, EoT proposes a novel similarity metric.

Recent studies propose learning-based methods, such as SimCSE (Gao et al., 2021) and BLEURT (Sellam et al., 2020), to improve sensitivity to semantic and syntactic variations beyond handcrafted metrics like ROUGE (Lin, 2004) and BLEU

(Papineni et al., 2002). These methods utilize language models such as BERT (Devlin et al., 2019) to encode statement pairs, producing scalar similarity scores based on the encoded representations.

However, learning-based metrics struggle to robustly align with human judgment in domain-specific scenarios. These metrics typically rely on end-to-end predictions from models trained on synthetic or commonsense datasets (Zhang et al., 2019). Discrepancies between distributions of training data and domain knowledge can lead to misalignment due to domain drift (Honovich et al., 2022). Moreover, while syntactic alignment strategies are prevalent, achieving semantic alignment akin to human judgment remains challenging.

To enhance robustness while maintaining high sensitivity, we introduce $Hssim$, a **h**ybrid metric to **s**core **sim**ilarity between statements, which combines the learned metric BLEURT with the NLI judgement of model ϕ, as detailed in Eq. (6). Specifically, given a reasoning process R for task (X, Q), the LLM generates an answer of n_1 statements $\hat{Y} = (\hat{y}_1, \hat{y}_2, \cdots, \hat{y}_{n_1})$ guided by R. Let $A = (a_1, a_2, \cdots a_{n_2})$ represent a reference answer n_2 statements. The similarity between $\hat{Y}$ and A is evaluated using $Hssim(\hat{Y}, A)$ calculated as:

$$Hssim(\hat{Y}, A) = \frac{\sum_{\hat{y}_i \in \hat{Y}} sim(\hat{y}_i, A)}{n_1}$$
$$sim(\hat{y}_i, A) = \begin{cases} 0 & \phi(A, \hat{y}_i) = 0 \\ \mu + (1-\mu)\beta_w(A, \hat{y}_i) & \phi(A, \hat{y}_i) = 1 \\ (1-\mu)\beta_w(A, \hat{y}_i) & Otherwise \end{cases} \tag{8}$$

where $\mu \in (0, 1)$ is the weight coefficient for measuring similarity between the NLI judgement and the learning-based scalar metric, and $\beta_w(\cdot, \cdot)$ denotes an approximation of the BLEURT score at the statement window level, calculated as:

$$\beta_w(A, \hat{y}_i) = \max(\{BLEURT(A[i_1 : i_2], \hat{y}_i) | \forall 0 < i_1 < i_2 \leq n_2\}) \tag{9}$$

Leveraging the exceptional in-context learning capabilities of LLMs, their NLI judgment can be robustly aligned with human judgement even facing domain drift, as noted in previous work (Ye and Durrett, 2022). Furthermore, the window-based estimation of BLEURT scores ensures sensitivity and alignment at the syntactic level. This allows for a rational and effective assessment of the reliability of any reasoning process R for question (X, Q), using $Hssim$, which balances variation sensitivity and semantic alignment robustness:

$$S_{rel}(R) = Hssim(p_\theta(X, Q|R), A) \tag{10}$$

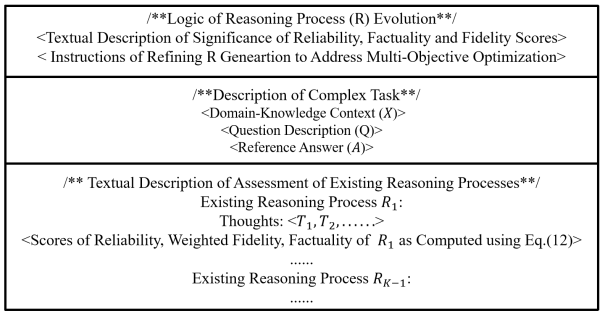


Figure 2: Prompt template to evolve reasoning process.

3.2.3 Fidelity Scorer

Based on the fidelity definition in Eq.(4), scoring the fidelity of thoughts involves two key aspects: 1) efficiently excluding a thought from LLMs' reasoning behaviors, and 2) accurately measuring the difference between answers produced with and without the thought. We use a thought masking scheme and metric $Hssim$ to achieve these goals.

Given a reasoning process R, each thought $T_i \in R$ is removed from R to create a masked reasoning process $(R \backslash T_i)$, as shown in Figure 8 in Appendix C. According to Eq.(1), a prompt directs the LLM to generate answers $\hat{Y}$ and $\hat{Y}_{\backslash T_i}$ for task (X, Q), using instructions from R and $(R \backslash T_i)$ respectively. An example is provided in Figure 4 in Appendix B. Finally, with $\hat{Y}$ as the reference, the fidelity score of T_i is assessed using the $Hssim$ metric:

$$S_{fid}(T_i) = 1 - Hssim(\hat{Y}_{\backslash T_i}, \hat{Y}) \tag{11}$$

Moreover, enhancing LLMs' reasoning capabilities requires fidelity optimization to focus on thoughts that faithfully represent the reasoning behaviors essential for reliable answers. Thus, we propose a weighted fidelity score for thought T_i, accounting for the reliability of the reasoning process:

$$S^{rel}_{fid}(T_i) = S_{fid}(T_i) \times S_{rel}(R) \tag{12}$$

Ultimately, we achieve a comprehensive assessment of the reasoning process and its thoughts:

$$\Big(S_{rel}(R), \{S_{fac}(T_i), S^{rel}_{fid}(T_i) | \forall T_i \in R\}\Big) \tag{13}$$

3.3 Prompter

The prompting mechanism of EoT is designed to enhance the reasoning process by enabling LLMs to comprehend the scoring outcomes from previous

iterations. This prompt LLMs to self-reflect and refine reasoning processes, optimizing reliability, fidelity, and factuality at the thought level.

The prompt template for the K-th iteration ($1 \leq K \leq N$) is presented in Figure 2. Each evolution iteration includes a prompt with three components: (1) the question description Q, knowledge context X, and reference answer A for the task; (2) the evolution logic, encompassing: a) the significance of three score types and b) instructions for multi-objective optimization; and (3) the assessing results of reasoning processes produced in prior iterations, as defined in Eq.(13). An example of a complete prompt is shown in Figure 9 in Appendix D.

3.4 Self-reflective Generator

To align the LLM's reasoning capability with the fidelity of its behavioral patterns, EoT employs the LLM used for question answering as the generator to iteratively refine reasoning processes via self-reflection. Each iteration uses prompts to guide the generator to produce an enhanced reasoning process, leading to multi-objective improvements recognized by LLMs. After N iterations, the reasoning process with the highest reliability score is chosen as the final evolved process R^*.

4 Experiments

4.1 Setup

Datasets To evaluate the effectiveness of EoT, we conduct experiments on two datasets: 1) OpsQA: a question-answering dataset with 40 complex operational maintenance (OM) tasks covering cloud computing, code management, application upgrades, and more. 2) LongBench (Bai et al., 2023b): a benchmark comprising problems from various open-source datasets such as HotpotQA, MultifieldQA, WikimQA. We select 60 representative questions from LongBench. Since EoT is designed to improve the capability of LLMs on knowledge-intensive reasoning tasks, we carefully select complex reasoning tasks for evaluation. Each task includes a long domain knowledge context with thousands or tens of thousands of tokens, a question description, and a reference answer, necessitating intricate solutions. To highlight the complexity of these 100 representative reasoning tasks, we provide statistics on the context length of selected questions, including the mean, the 95th percentile (P95) and the maximum (Max) value.

As presented in Table 2, the average length of knowledge context reaches $6,052$ and $5,700$ tokens for the selected tasks in the OpsQA and LongBench datasets, respectively. Moreover, the maximum context length can reach up to $21,045$ and $14,640$ tokens for these tasks in the OpsQA and LongBench datasets respectively. This considerable complexity of domain knowledge context for tasks could demonstrate that our experiments accurately assess the effectiveness of EoT in addressing complex reasoning tasks that require understanding extensive domain knowledge and producing intricate solutions.

Table 2: Statistics of Context Length of Selected Tasks

Dataset	Context (token)		
	Mean	P95	Max
OpsQA	$6,052$	$11,640$	$21,045$
LongBench (selected)	$5,700$	$11,782$	$14,640$

Table 3: The parameter settings in evaluations

Parameter	Value	Description
N	10	The number of iterations
μ	0.5	The alignment weight in Eq.(8)

Testbed and Parameter Settings We utilize two widely used LLMs, Qwen2-72B (Bai et al., 2023a) and GPT-4 Turbo (Openai, 2023), each serving as the generator of both reasoning processes and solutions. For the assessor, EoT uses GPT-4 Turbo to implement the NLI model ϕ, due to its advanced NLI accuracy as evaluated in Appendix E.2, and compute window-based BLEURT scores in Eq. (9) using V100 GPUs. Table 3 outlines the hyper-parameter settings in our evaluations.

Metrics To assess LLMs' reasoning capabilities, we compute four metrics by comparing answers generated by the reasoning process $\hat{Y}$ with the reference answer A: 1) reliability defined in Eq. (10), 2) BLEURT, 3) ROUGE-L (Lin, 2004), and 4) NLI results from GPT-4 Turbo, denoted as $entail(\%)$, which measures the percentage of statements in $\hat{Y}$ entailed by A. Additionally, following prior studies (Ye and Durrett, 2022; Lyu et al., 2023), we use S_{fac} and S_{fid} from Eq.(7) and Eq.(11) to assess performance in factuality and behavior fidelity, respectively. Scores of a reasoning process are averaged from those of thoughts. Higher values indicate better performance for each metric.

Baselines We compare EoT with five advanced frameworks for evolving reasoning processes: 1) BoT (Chen et al., 2024), CoT-dec (Radhakrishnan et al., 2023), and its variant Factor-dec, which emphasize reliability optimization; 2) xLLM (Chuang

et al., 2024) designed for fidelity optimization; and 3) Calibrator (Ye and Durrett, 2022) targeting factuality and fidelity optimization.

According to the schema of reasoning process generation, these frameworks can be categorized into three groups. (1) Exploration of structured thoughts: BoT guides LLMs to explore ensemble of trees of thoughts (ToTs) and acquire trial-and-error reasoning experiences for each explored ToT. In BoT, a reasoning thought is generated by the LLM as a node of ToT in each inference, and the complete reasoning process is formed by a series of such LLM inferences. (2) Revision of the complete reasoning process: a) xLLM provides feedback on the fidelity assessment for the complete reasoning process and guides LLMs to optimize reasoning processes through iterative improvements; b) Calibrator focuses on revising reasoning processes to improve factuality and further enhance the reasoning capability of LLMs through iterations. In these studies, LLMs generate a complete reasoning process in textual paragraphs via a single-forward inference. (3) Question decomposition: a) CoT-dec prompts LLMs to decompose the reasoning process to a sequence of subquestion-subanswer pairs produced in single-forward inference in each iteration. It seeks decomposed results that achieve the best reliability performance as the final reasoning process after multiple iterations; b) Factor-dec is a variant of CoT-dec. In each evolution iteration, Factor-dec guides LLMs to decompose a reasoning process into a subqeustion sequence and prompts LLMs to answer these subquestions through step-by-step inferences. In other words, each iteration of Factor-dec involves multiple steps of question-answering in generating a reasoning process.

4.2 Overall Performance

We conduct experiments on the OpsQA and LongBench datasets to evaluate the performance of the six frameworks in terms of the reasoning capability, factuality and fidelity of reasoning processes and time efficiency. Moreover, the ablation study of EoT is detailed in Appendix G. To further examine EoT's generality, in Appendix H, we conduct evaluations on additional 40 OM tasks collected in production, with LLMs of diverse parameter sizes.

4.2.1 Reasoning Capability (Reliability)

Table 4 presents the average reasoning capability of two LLMs utilizing the six frameworks. EoT surpasses the five baselines in three areas. First, compared to the leading baseline, CoT-dec, EoT boosts reliability scores on the OpsQA and LongBench datasets by about 11.7% and 5.8% using Qwen2-72B, and by about 3.3% and 3.7% using GPT-4 Turbo. Second, in terms of sensitivity to token and semantic variation, EoT improves the ROUGE-L and BLEURT on LongBench dataset by about 7.2% and 16.2% using Qwne2-72B, and 2.6% and 6.8% using GPT-4 Turbo. Third, regarding semantic alignment robustness, on the OpsQA dataset, EoT improves $entail(\%)$ by around 14.4% and 6.7% using Qwen2-72B and GPT-4 Turbo, respectively. These findings indicate that EoT empowers diverse LLMs to enhance overall reasoning capability in complex tasks.

Figure 3a intuitively shows that, with GPT-4 Turbo, EoT achieves a reliability score > 0.75 for 52.5% of tasks in the OpsQA dataset, surpassing BoT, xLLM, Calibrator, CoT-dec, and Factor-dec by about 50%, 27.5%, 22.5%, 7.5% and 35%, respectively. Furthermore, for tasks in OpsQA and LongBench, the complete CDF results of reasoning capability using Qwen-72B and GPT-4 Turbo, equipped with the six frameworks, are detailed in Appendix F.1. These findings confirm EoT's significant generality in enhancing reasoning capability across diverse fields with various LLMs.

EoT's enhanced reasoning capability stems from two key aspects. Firstly, we introduce an effective similarity scoring metric $Hssim$ to measure the reliability of reasoning processes. By accounting for sensitivity to semantic variations and robustness in semantic alignment, $Hssim$ promote the reliability improvement of EoT, facilitating comprehensive optimization of LLMs' reasoning capabilities across various semantic aspects. For instance, on the OpsQA dataset, despite EoT's BLEURT performance being about 1.79% lower than CoT-dec using GPT-4 Turbo, EoT improves NLI results by about 6.7%, thereby raising the reliability score by 3.3%. Secondly, unlike baselines focusing on individual or partial factors, EoT efficiently achieves multi-objective optimization in reliability, factuality, and fidelity during reasoning process evolution, thereby enhancing LLMs' reasoning capabilities.

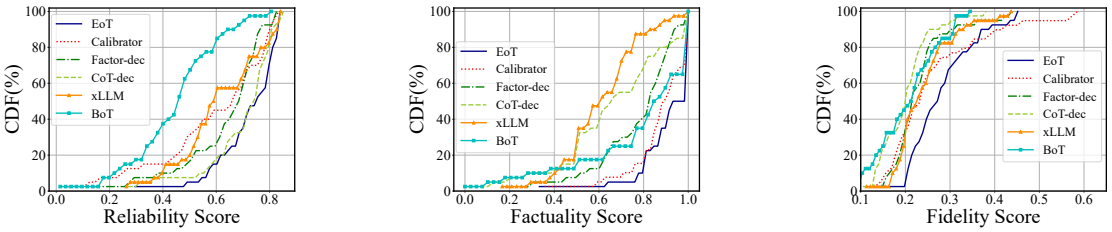
4.2.2 Factuality & Fidelity Performance.

The central concept of EoT is to improve the factuality and behavior fidelity of thoughts, in a direction of enhancing LLMs' reasoning capabilities. To assess these efforts, Table 5 presents the average performance in fidelity and factuality of reason-

Table 4: Reliability performance for reasoning processes evolved by six frameworks with two LLMs on two datasets.

LLMs	Models	OpsQA				LongBench			
		BLEURT	ROUGE-L	NLI results $entail(\%)$	Reliability Score	BLEURT	ROUGE-L	NLI results $entail(\%)$	Reliability Score
Qwen2	BoT	0.499	0.297	43.29	0.401	0.494	0.677	62.92	0.526
	xLLM	0.554	0.355	65.83	0.576	0.572	0.739	87.50	0.721
	Calibrator	0.559	0.365	70.27	0.597	0.575	0.744	80.88	0.675
	CoT-dec	0.579	0.361	73.35	0.632	0.586	0.769	**92.63**	0.757
	Factor-dec	0.545	0.324	62.55	0.569	0.570	0.745	84.47	0.729
	EoT (ours)	**0.595**	**0.395**	**83.90**	**0.706**	**0.681**	**0.824**	91.52	**0.801**
GPT-4	BoT	0.486	0.307	48.50	0.443	0.530	0.704	67.99	0.568
	xLLM	0.575	0.408	72.16	0.609	0.629	0.801	85.09	0.729
	Calibrator	0.554	0.380	68.50	0.589	0.592	0.775	83.98	0.709
	CoT-dec	**0.613**	**0.446**	83.02	0.694	0.617	0.805	91.65	0.767
	Factor-dec	0.541	0.358	74.48	0.632	0.587	0.767	89.44	0.739
	EoT (ours)	0.602	0.427	**88.58**	**0.717**	**0.659**	**0.826**	**92.47**	**0.795**

[1] Qwen2 and GPT-4 represent the LLMs of Qwen2-72B and GPT-4 Turbo, respectively.



(a) Reliability score (b) Factuality score (c) Fidelity score

Figure 3: The CDF results of the reliability, factuality and fidelity scores of reasoning processes evolved by the six frameworks using GPT-4 Turbo on 40 tasks in the OpsQA dataset.

Table 5: The performance on factuality and fidelity for reasoning processes evolved by six frameworks.

LLMs	Models	OpsQA		LongBench	
		Factuality S_{fac}	Fidelity S_{fid}	Factuality S_{fac}	Fidelity S_{fid}
Qwen2	BoT	0.752	0.234	0.879	0.208
	xLLM	0.720	0.216	0.864	0.152
	Calibrator	**0.860**	0.258	0.935	0.141
	CoT-dec	0.638	0.167	0.839	0.179
	Factor-dec	0.706	0.208	0.901	0.197
	EoT (ours)	0.823	**0.267**	**0.943**	**0.243**
GPT-4	BoT	0.775	0.203	0.896	0.237
	xLLM	0.607	0.238	0.848	0.169
	Calibrator	0.878	0.257	0.942	0.180
	CoT-dec	0.685	0.199	0.934	0.171
	Factor-dec	0.769	0.228	0.936	0.208
	EoT (ours)	**0.906**	**0.281**	**0.956**	**0.275**

ing processes evolved by six frameworks, using Qwen2-72B and GPT-4 Turbo on the two datasets. These results highlight two key aspects.

Firstly, EoT significantly reduces non-factual errors in thoughts while boosting the behavior fidelity of LLMs. Compared to the leading baseline Calibrator, using GPT-4 Turbo, EoT improves the factuality score by about 3.2% and 1.3% on the OpsQA and LongBench datasets respectively. With Qwen2-72B, EoT shows a factuality improvement of around 4.0% on the LongBench dataset. Additionally, on the OpsQA dataset, EoT surpasses Calibrator by enhancing the fidelity score by about 9.3% using GPT-4 Turbo. Similarly, on the LongBench dataset, compared to BoT, EoT improves the fidelity score by roughly 16.8% and 16.0% using Qwen-72B and GPT-4 Turbo respectively. These results confirm EoT's capability to effectively enhance both factuality and behavior fidelity.

Secondly, EoT's optimization on factuality and fidelity effectively boosts reasoning capabilities of LLMs. On the OpsQA dataset, EoT surpasses CoT-dec in factuality and fidelity by about 28.9% and 59.9% respectively, resulting in a reliability improvement of around 11.7%. This enhanced reasoning capability is largely due to substantial advancements in factuality and fidelity of reasoning processes. Moreover, although Calibrator improves factuality by about 4.4% over EoT on the OpsQA dataset, EoT improves fidelity and reliability by about 3.5% and 18.3% respectively.

To illustrate EoT's generality in optimizing factuality and fidelity, Figures 3b and 3c show the CDF of factuality and fidelity scores for reasoning processes evolved by six frameworks using GPT-4 Turbo on the OpsQA dataset, respectively. EoT achieves a factuality score of 1.0 for 50% of OpsQA questions, surpassing BoT, xLLM, Calibrator, CoT-dec, and Factor-dec by around 15%, 47.5%, 20%, 35%, and 42.5%, respectively. Moreover, EoT exceeds a fidelity score of 0.25 for 60% of questions, surpassing BoT, xLLM, Calibrator, CoT-dec, and Factor-dec by about 32.5%, 20%,

Table 6: The average performance on time efficiency of the six frameworks on the OpsQA dataset.

Models	Overhead(s)	Iterations
BoT	945.85	4.48
Factor-dec	114.77	5.21
xLLM	70.89	5.25
Calibrator	67.75	4.92
CoT-dec	70.22	5.34
EoT (ours)	75.46	4.62

25%, 50%, and 37.5%, respectively. Further CDF results for fidelity and factuality on OpsQA and LongBench datasets are provided in Appendix F.2 and F.3. These results confirm that EoT's improvements in fidelity and factuality are generalized to tasks across various domains, including operational maintenance and academic literature.

These improvements stem from two aspects. Firstly, EoT effectively assesses factuality and fidelity at a granular thought level, providing a foundation for effective optimization. Secondly, our prompter fully leverages ICL capabilities of LLMs, enabling them to rationally comprehend performance experiences and simultaneously optimize across three objectives, as defined in Eq.(5).

4.2.3 Time Efficiency

We assess the time efficiency of six frameworks using two criteria: 1) the average overhead per iteration, where lower values denote higher efficiency; and 2) the convergence rate, measured as the average number of iterations needed to produce the final enhanced reasoning process for various questions. Lower values signify quicker convergence.

Table 6 showcases the time efficiency of the six frameworks on the OpsQA dataset. EoT reduces iteration overhead by about 92.0% compared to BoT and 33.4% versus Factor-dec, both of which involve multiple thought explorations per iteration. In addition, compared to three other baselines employing single-forward inference, EoT uses fewer iterations to generate refined reasoning processes with similar iteration time. These results validate that EoT achieves comprehensive, fine-grained optimization of each thought with high time efficiency.

5 Conclusion

Through in-depth analysis, we outline three key factors in reasoning processes to enhance LLMs' capabilities, namely factuality, fidelity and reliability. We propose EoT to evolve LLM-generated reasoning processes across these dimensions. An assessor is designed to quantify performance on these aspects for each thought in reasoning processes. Additionally, we propose a prompting mechanism that efficiently guides LLMs to comprehend assessments and trigger self-reflection to achieve thought-level and multi-objective optimization of reasoning processes via single-forward inference. EoT offers two key advantages. First, it considers comprehensive factors to improve reasoning capability. Second, it ensures thought-level evolution with high time efficiency. In the future, we will address unsupervised evolution of reasoning processes to further enhance reasoning capabilities of LLMs for out-of-domain tasks without reference solutions.

Acknowledgments

This work was supported by Alibaba Group through Alibaba Innovative Research Program and Alibaba Research Intern Program, Hangzhou High-Tech Zone (Binjiang) Institute of Blockchain and Data Security, and the Artificial Intelligence Technology Support Project of the Science and Technology Commission of Shanghai Municipality (22DZ1100103).

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and et al. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, and et al. 2023a. Qwen technical report. *Preprint*, arXiv:2309.16609.

Yushi Bai, Xin Lv, Jiajie Zhang, Hongchang Lyu, Jiankai Tang, Zhidian Huang, Zhengxiao Du, Xiao Liu, Aohan Zeng, Lei Hou, Yuxiao Dong, Jie Tang, and Juanzi Li. 2023b. Longbench: A bilingual, multitask benchmark for long context understanding. *Preprint*, arXiv:2308.14508.

Maciej Besta, Nils Blach, Ales Kubicek, Robert Gerstenberger, Michal Podstawski, Lukas Gianinazzi, Joanna Gajda, Tomasz Lehmann, Hubert Niewiadomski, Piotr Nyczyk, and 1 others. 2024. Graph of thoughts: Solving elaborate problems with large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 17682–17690.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and et al. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901.

Sijia Chen, Baochun Li, and Di Niu. 2024. Boosting of thoughts: Trial-and-error problem solving with large language models. *arXiv preprint arXiv:2402.11140*.

Yu-Neng Chuang, Guanchu Wang, Chia-Yuan Chang, Ruixiang Tang, Fan Yang, Mengnan Du, Xuanting Cai, and Xia Hu. 2024. Large language models as faithful explainers. *arXiv preprint arXiv:2402.04678*.

DeepSeek-AI. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple contrastive learning of sentence embeddings. In *Empirical Methods in Natural Language Processing (EMNLP)*.

Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. 2023. Enabling large language models to generate text with citations. In *Empirical Methods in Natural Language Processing (EMNLP)*.

Or Honovich, Roee Aharoni, Jonathan Herzig, Hagai Taitelbaum, Doron Kukliansy, Vered Cohen, Thomas Scialom, Idan Szpektor, Avinatan Hassidim, and Yossi Matias. 2022. TRUE: Re-evaluating factual consistency evaluation. In *Proceedings of the 2nd DialDoc Workshop on Document-grounded Dialogue and Conversational Question Answering*, pages 161–175, Dublin, Ireland. Association for Computational Linguistics.

Yuxin Liang, Zhuoyang Song, Hao Wang, and Jiaxing Zhang. 2024. Learning to trust your feelings: Leveraging self-awareness in llms for hallucination mitigation. *arXiv preprint arXiv:2401.15449*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Gianluigi Lopardo, Frederic Precioso, and Damien Garreau. 2023. Faithful and robust local interpretability for textual predictions. *arXiv preprint arXiv:2311.01605*.

Qing Lyu, Shreya Havaldar, Adam Stein, Li Zhang, Delip Rao, Eric Wong, Marianna Apidianaki, and Chris Callison-Burch. 2023. Faithful chain-of-thought reasoning. *Preprint*, arXiv:2301.13379.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, and et al. 2023. Self-refine: Iterative refinement with self-feedback. In *Advances in Neural Information Processing Systems*, volume 36, pages 46534–46594.

Openai. 2023. Gpt-4 turbo in the openai api.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, page 311–318, USA. Association for Computational Linguistics.

Debjit Paul, Mete Ismayilzada, Maxime Peyrard, Beatriz Borges, Antoine Bosselut, Robert West, and Boi Faltings. 2023. Refiner: Reasoning feedback on intermediate representations. *arXiv preprint arXiv:2304.01904*.

Ansh Radhakrishnan, Karina Nguyen, Anna Chen, Carol Chen, Carson Denison, Danny Hernandez, Esin Durmus, Evan Hubinger, Jackson Kernion, Kamilė Lukošiūtė, and 1 others. 2023. Question decomposition improves the faithfulness of model-generated reasoning. *arXiv preprint arXiv:2307.11768*.

Thibault Sellam, Dipanjan Das, and Ankur P Parikh. 2020. Bleurt: Learning robust metrics for text generation. *arXiv preprint arXiv:2004.04696*.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models. *Preprint*, arXiv:2203.11171.

Adina Williams, Nikita Nangia, and Samuel R. Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. *Preprint*, arXiv:1704.05426.

Xi Ye and Greg Durrett. 2022. The unreliability of explanations in few-shot prompting for textual reasoning. In *Advances in Neural Information Processing Systems*, volume 35, pages 30378–30392.

Chenrui Zhang, Lin Liu, Chuyuan Wang, Xiao Sun, Hongyu Wang, Jinpeng Wang, and Mingchen Cai. 2024. Prefer: Prompt ensemble learning via feedback-reflect-refine. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(17):19525–19532.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Chuanyang Zheng, Zhengying Liu, Enze Xie, Zhenguo Li, and Yu Li. 2023. Progressive-hint prompting improves reasoning in large language models. *arXiv preprint arXiv:2304.09797*.

A Limitations

The EoT framework exhibits two principal limitations that warrant discussion. First, its assessing mechanism inherently relies on reference solutions to assess reasoning reliability and subsequently refines the reasoning processes based on these supervised signals. This paradigm creates a dependency on domain-specific answer templates, potentially constraining the framework's capacity to generalize and enhance LLM reasoning performance for out-of-domain tasks where authoritative references are unavailable or undefined.

Second, the $Hssim$ metric implementation introduces an intrinsic dependency on the LLM's natural language inference (NLI) capabilities. The effectiveness of $Hssim$ measurements becomes contingent upon the model's cross-domain NLI accuracy, introducing potential error propagation across different knowledge domains. To address this limitation, our future work will focus on integrating state-of-the-art LLMs with enhanced NLI datasets while developing domain-agnostic verification mechanisms to strengthen metric robustness.

B Examples of Reasoning Process Formalization

As formulated in Section **Problem Setup**, for each complex task (X, Q), an LLM expresses its reasoning behaviors in the form of a reasoning process R. This process consists of multiple thought steps and helps guide the LLM in producing the task's solution, denoted as $\hat{Y}$ in Eq. (1) of our submitted manuscript. Specifically, LLMs that respond to questions by following a reasoning process operate by using specific prompting instructions. The prompt template used in these invocations is presented in Figure 4. This prompting strategy encourages LLMs to execute their genuine reasoning behaviors in a way that closely reflects self-explanatory thoughts, with the goal of enhancing the influence of these thoughts on the LLMs' solution generation.

On this basis, to intuitively illustrate the self-explained reasoning process, Figure 5 and 7 present two examples of reasoning processes generated by the widely used LLM Qwen2-72B. These examples are specifically aimed at guiding the LLM in addressing two distinct tasks in LongBench.

C Illustration of Factuality, Fidelity and Reliability

In Section **Problem Setup** of our submitted manuscript, we define three critical factors that influence the reasoning capability of LLMs: factuality, fidelity and reliability. To aid in the intuitive understanding of the measurement of these three factors, this appendix presents both positive and negative examples of reasoning thoughts for each of them. For each factor, the positive examples demonstrate strong performance, while the negative ones illustrate poor performance.

C.1 Factuality

Figure 6 presents a completely factual reasoning process R_{fac}, alongside a reasoning process that contains non-factual errors R_{nonfac}. Specifically, all four thought steps in R_{fac} are factually grounded in the knowledge context. In contrast, T_2 and T_3 in R_{nonfac} contain errors that contradict established domain knowledge, as highlighted in the underlined part, which demonstrates that T_2 and T_3 exhibit poor factual performance. On this basis, Qwen2-72B generates two answers to the question "Which film came out first", denoted as $\hat{Y}_{fac}$ and $\hat{Y}_{nonfac}$, under the guideline of R_{fac} and R_{nonfac}, respectively.

Then, we observe that, when Qwen2-72B responds to the question following the guideline of reasoning process R_{nonfac} which contains non-factual date errors, non-factual hallucination may emerge in its generated solution $\hat{Y}_{nonfac}$. In contrast, a correct answer $\hat{Y}_{fac}$ can be produced by LLMs under the guideline of R_{fac}, where the factuality of the explanatory reasoning thoughts is ensured. These clear examples highlight the importance of ensuring the factuality of reasoning processes during optimization, as this is crucial to prevent the reasoning capability of LLMs being compromised by non-factual hallucinations.

C.2 Fidelity

We review the reasoning process R that guides Qwen2-72B in solving the question, "What are some fields in which the inverse problem is encountered?", as presented in Figure 7. This reasoning process consists of four steps of thought. To investigate whether each step faithfully reflects the actual reasoning behaviors of LLMs, we demonstrate the measurement of behavior fidelity for each reasoning step as defined in Eq.(4). Specifically, we mask

/**Instruction to emphasize LLMs to sovle questions following self-explained thoughts**/
Hello, you are the intelligent robot of the smart Q&A project. Now, your task is to observe the currently explained reasoning process, Context, Question, guide yourself to refer to the valid content in Context based on the thinking and answering methodologies provided by the provided reasoning process, and achieve an answer to the Question (i.e., generating the Answer) subsequently.
Please pay attention, you need to ensure the following requirements:
(1) In the process of generating the answer, please ensure as much as possible to think and deduce answers step-by-step according to the provided reasoning process's STEP-BY-STEP Thought;
(2) Please ensure that you do not introduce new reasoning process or step of thoughts to the process of reasoning the answer, and ensure that you do not revise the provided reasoning process. On this basis, please ensure as much as possible that the thinking method or STEP-BY-STEP Thought used in reasoning is consistent with the provided reasoning process. Specifically, please ensure that the way you refer to/quote Context and the way of question answering is consistent with the provided reasoning process.
(3) Please note that your task is to refer to or quote the text according to the thoughts of the reasoning process to achieve question answering, and ultimately your output is the generated answer namely Answer. Therefore, please note, whether the generated Answer is right or wrong, you can not offer additional modification suggestions for the answer generated under the guideline of the provided reasoning process, nor modify the generated answer."

/**Description of Complex Task**/
Given Context: <Domain Knowledge Context (X)>
Given Question: <Question Description (Q)>

/** Given a Reasoning Process to guide LLMs**/
Your explained Reasoning Process: <Explained Reasoning Process (R)>

/**Instruction to Trigger the Question Answering**/
Following the thoughts and answering method of this reasoning process to answer the above Question. Your Answer to the Question is:

Figure 4: An example prompt template that guides LLMs to generate task solutions under the guideline of specified self-explanatory reasoning processes.

thoughts T_1 and T_3 in R to obtain two masked reasoning processes $R\backslash T_1$ and $R\backslash T_3$ respectively, as discussed in subsection **Fidelity Scorer**. Figure 8 shows that Qwen2-72B generates two answers to the question under the guidelines of $RP\backslash T_1$ and $RP\backslash T_3$, denoted as $\hat{Y}_{\backslash T_1}$ and $\hat{Y}_{\backslash T_3}$, respectively. A comparison reveals that, relative to the answer guided by R, represented as $\hat{Y}$, a significant change is observed in $\hat{Y}_{\backslash T_1}$, while only a minor modification is noted in $\hat{Y}_{\backslash T_3}$.

In light of this, we can derive two insights. Firstly, the reasoning behavior exhibited in T_1, namely “Extract Key Concepts", faithfully reflects a critical step in the actual reasoning actions of Qwen2-72B. This indicates that LLMs delve into the significance of each key entity, including “inverse problem" and its related fields, at the beginning of inference. In other words, T_1 significantly affects the reasoning results of LLMs, which indicates that T_1 exhibits strong behavior fidelity for Qwen2-72B. Secondly, the reasoning claim in T_3, termed “Contextual Understanding", inadequately uncovers the actual behaviors that Qwen2-72B executes, as omitting these claims has only a minor impact on the LLM’s reasoning results. Thus, reasoning thoughts such as T_3 can be regarded as unfaithful thoughts in the reasoning processes explained by LLMs. In other words, thoughts like T_3 demonstrate poor behavior fidelity for the LLM.

C.3 Reliability

As demonstrated in Section **Problem Setup**, the reliability of a reasoning process for LLMs in addressing any task can be measured by the similarity between the generated solution and the reference answer. Since solutions often consist of long-textual statements, the assessment of similarity should occur at both syntactic and semantic levels. In addition, all measurements should aim to align with the understanding of experts in the relevant domain.

Intuitively, the reasoning processes presented in Figure 5 and Figure 7 perform satisfied reliability as they effectively guide Qwen2-72B in producing solutions that adequately address the respective questions. Consequently, we should evaluate the high similarities between the produced and reference solutions for these two reasoning processes accordingly. Conversely, the assessed similarity should be significantly low for unreliable reasoning processes. For example, as shown in Figure 6, the non-factual reasoning process R_{nonfac} results in

Question (Q): Which film came out first, "Donne con le gonne" or "The Bag Man"?

Knowledge Context (X):
Passage 1:
"Bagman" (video game), a 1982 French platform arcade game.....
"Begman" (film) , a 2010 film about Jack Abramoff.....
.....
Passage 3.
.....
Donne con le gonne
Women in Skirts (Italian: Donne con le gonne) is a 1991 Italian romantic comedy film directed by Francesco Nuti.....
.....
Passage 7:
The Bag Man (also known as Motel or The Carrier) is a 2014 neo-noir crime thriller film directed by.....
.....

Reference Answer (A): Donne con le gonne.

Factual Reasoning Process (R):
Thought T_1: Task Identification: The task necessitates establishing which movie, "Donne con le gonne" or "The Bag Man", was released first by examining their respective release dates.
Thought T_2: Date Extraction: From Passage 3, I extract that "Donne con le gonne" was released in 1991. Meanwhile, Passage 7 indicates "The Bag Man" as a 2014 film.
Thought T_3: Temporal Sequence: Comparing the two years, it's evident that 1991 comes before 2014, confirming "Donne con le gonne" was released prior to "The Bag Man."
Thought T_4: Conclusion Formation: Based on the chronological order of their release dates, "Donne con le gonne" was released first.

Answer Guided by R ($\widehat{Y}$ in Eq.(1)): "Donne con le gonne" was released first.

Figure 5: A factual reasoning process produced by Qwen2-72B for problem "Which film came out first?". The highlighted part represents the critical information and claims in the context and reasoning behaviors.

<table>
<tr><td colspan="2">Question (Q): Which film came out first, "Donne con le gonne" or "The Bag Man"?</td></tr>
<tr><td colspan="2">Knowledge Context (X):
Passage 1:
"Bagman" (video game), a 1982 French platform arcade game....."Begman" (film) , a 2010 film about Jack Abramoff.....
.....
Passage 3.
.....
Donne con le gonne: Women in Skirts (Italian: Donne con le gonne) is a 1991 Italian romantic comedy film directed by Francesco Nuti.....
.....
Passage 7:
The Bag Man (also known as Motel or The Carrier) is a 2014 neo-noir crime thriller film .....
.....</td></tr>
<tr><td>Factual Reasoning Process (R_{fac}):
Thought T_1: Task Identification: The task necessitates establishing which movie, "Donne con le gonne" or "The Bag Man", was released first by examining their respective release dates within the given texts.
Thought T_2: Date Extraction: From Passage 3, I extract that "Donne con le gonne" was released in 1991. Meanwhile, Passage 7 indicates "The Bag Man" as a 2014 film.
Thought T_3: Temporal Sequence: Comparing the two years, it's evident that 1991 comes before 2014, confirming "Donne con le gonne" was released prior to "The Bag Man".
Thought T_4: Conclusion Formation: Based on the chronological order of their release dates, "Donne con le gonne" was released first.</td><td>Non-Factual Reasoning Process (R_{nonfac}):
Thought T_1: Task Identification: The task necessitates establishing which movie, "Donne con le gonne" or "The Bag Man", was released first by examining their respective release dates within the given texts.
Thought T_2: Date Extraction: From Passage 3, I extract that "Donne con le gonne" was released in 1991. Meanwhile, Passage 1 indicates "The Bag Man" as a 1982 film.
Thought T_3: Temporal Sequence: Comparing the two years, it's evident that 1982 comes before 1991, confirming "The Bag Man" was released prior to "Donne con le gonne" .
Thought T_4: Conclusion Formation: Based on the chronological order of their release dates, "The Bag Man" was released first.</td></tr>
<tr><td>Correct Answer: "Donne con le gonne" was released first.</td><td>Hallucination: "The Bag Man" was released first.</td></tr>
</table>

Figure 6: Examples of total factual and non-factual reasoning processes produced by Qwen2-72B for problem "Which film came out first?". The highlighted part in factual reasoning process represents the key factual claims, while the underlined part is the claims with non-factual errors which cause the hallucination in answers.

Question (Q): What are some fields in which the inverse problem is encountered?

Knowledge Context (X):

.....

Given a data set and a model with some unknown parameters, the inverse problem aims to find the values of the model parameters that best fit the data.

.....

Versions of this problem are encountered in physics, biology [1,2], social sciences and finance [3,4], neuroscience [5], and are becoming more and more important due to .....

.....

In particular, the inverse problem on the pairwise XY model analyzed here might be of help in recovering images.....

.....

We introduce the PLM with L_2 regularization and with decimation, two variants of the PLM respectively introduced in [7,8] for the inverse Ising problem.

.....

Reference Answer (A): Physics, biology, social sciences, finance, and neuroscience are some fields in which the inverse problem is encountered.

Reasoning Process (R):

Thought T_1: Extract Key Concepts: We learn about the inverse problem, which involves statistical inference for systems like the XY model applicable across various fields. The document highlights physics, biology, social sciences, finance, and neuroscience as domains where inverse problems occur.

Thought T_2: Relate Inverse Problem to Fields: In physics, inverse problems are crucial for understanding complex systems. Biology utilizes these methods for studying synchronization phenomena. Social sciences employ them for network analysis, finance for modeling financial systems, and neuroscience for mapping brain networks.

Thought T_3: Contextual Understanding: The inverse problem's relevance across these fields underscores its importance in extracting meaningful information from data. By applying statistical inference techniques like PLM, researchers can uncover hidden structures and interactions in complex systems.

Thought T_4: Formulate Answer: Based on the the text, the inverse problem is encountered in diverse fields such as physics, biology, social sciences, finance, and neuroscience, demonstrating its wide applicability in understanding intricate systems.

Answer Guided by R ($\widehat{Y}$ in Eq.(1)): The inverse problem is encountered in diverse fields such as physics, biology, social sciences, finance, and neuroscience, demonstrating its wide applicability in understanding intricate systems.

Figure 7: A reasoning process of Qwen2-72B for the problem “What are some fields in which the inverse problem is encountered?". The highlighted part represents the critical information and claims in the context and reasoning behaviors.

an incorrect answer to the question. Therefore, we should accurately measure a low similarity between these hallucinations and the correct reference. In addition, as presented in Figure 8, a reasoning process excluding faithful thoughts, denoted as $R\backslash T_1$, will diminish the reasoning capabilities of LLMs, and results in the emergence of redundant statements in generated answers, which are unrelated to problem-solving. Thus, a low similarity needs to be measured to represent the decline of reliability in terms of the precision of extracting useful knowledge.

D Prompting Mechanism of EoT

To illustrate the design of prompting scheme in EoT that drives the evolution of the reasoning process, Figure 9 provides an example prompt in an evolution iteration. This prompt is intended to guide a specific optimization of the reasoning process explained by Qwen2-72B for the question “What are some fields in which the inverse problem is encountered?". Specifically, this prompt follows the template illustrated in Figure 2 of our submitted manuscript, and includes a textual description of the assessment results from an existing reasoning process, as shown in Figure 7.

E NLI mechanism in Assessment of Reasoning process

In EoT, we prompt LLMs to perform natural language inference (NLI) judgements to evaluate the effectiveness of reasoning processes. Overall, lever-

Question (Q): What are some fields in which the inverse problem is encountered?

Knowledge Context (X):
.....
Given a data set and a model with some unknown parameters, the inverse problem aims to find the values of the model parameters that best fit the data.
.....
Versions of this problem are encountered in physics, biology [1,2], social sciences and finance [3,4], neuroscience [5], and are becoming more and more important due to .....
.....
In particular, the inverse problem on the pairwise XY model analyzed here might be of help in recovering images.....
.....
We introduce the PLM with L_2 regularization and with decimation, two variants of the PLM respectively introduced in [7,8] for the inverse Ising problem.
.....

Reference Answer (A): Physics, biology, social sciences, finance, and neuroscience are some fields in which the inverse problem is encountered.

Answer Guided by complete reasoning process R ($\widehat{Y}$ in Eq.(1)): The inverse problem is encountered in diverse fields such as physics, biology, social sciences, finance, and neuroscience, demonstrating its wide applicability in understanding intricate systems.

Masked Reasoning Process excluding T_1 ($R \setminus T_1$):
Thought T_1: <MASK>
Thought T_2: Relate Inverse Problem to Fields: In physics, inverse problems are crucial for understanding complex systems. Biology utilizes these methods for studying synchronization phenomena. Social sciences employ them for network analysis, finance for modeling financial systems, and neuroscience for mapping brain networks.
Thought T_3: Contextual Understanding: The inverse problem's relevance across these fields underscores its importance in extracting meaningful information from data. By applying statistical inference techniques like PLM, researchers can uncover hidden structures and interactions in complex systems.
Thought T_4: Formulate Answer: Based on the the text, the inverse problem is encountered in diverse fields such as physics, biology, social sciences, finance, and neuroscience, demonstrating its wide applicability in understanding intricate systems.

Masked Reasoning Process excluding T_3 ($R \setminus T_3$):
Thought T_1: Extract Key Concepts: We learn about the inverse problem, which involves statistical inference for systems like the XY model applicable across various fields. The document highlights physics, biology, social sciences, finance, and neuroscience as domains where inverse problems occur.
Thought T_2: Relate Inverse Problem to Fields: In physics, inverse problems are crucial for understanding complex systems. Biology utilizes these methods for studying synchronization phenomena. Social sciences employ them for network analysis, finance for modeling financial systems, and neuroscience for mapping brain networks.
Thought T_3: <MASK>
Thought T_4: Formulate Answer: Based on the the text, the inverse problem is encountered in diverse fields such as physics, biology, social sciences, finance, and neuroscience, demonstrating its wide applicability in understanding intricate systems.

Answer Guided by $R \setminus T_1$ ($\widehat{Y}_{\setminus T_1}$): The inverse problem is encountered in various fields such as physics, biology, social sciences, finance, and neuroscience. In physics, it plays a significant role in understanding systems with interacting elements, particularly in disordered media and laser dynamics. Biology employs inverse problems to study synchronization phenomena, like those observed in starling flocks. Social sciences utilize these methods for network analysis, while finance relies on them for modeling complex financial systems. Additionally, neuroscience benefits from inverse problem-solving techniques in mapping intricate brain networks.

Answer Guided by $R \setminus T_3$ ($\widehat{Y}_{\setminus T_3}$): The inverse problem is encountered in various fields including physics (e.g., understanding complex systems and laser dynamics), biology (e.g., studying synchronization phenomena), social sciences (e.g., network analysis), finance (e.g., modeling financial systems), and neuroscience (e.g., mapping brain networks)..

Figure 8: Examples of masked reasoning processes produced by Qwen2-72B for problem "What are some fields in which the inverse problem is encountered?". The highlighted part represents the critical information and claims in the context and reasoning behaviors. We exclude the claims of thoughts T_1 and T_3 respectively, and find that T_1 has a higher fidelity performance compared with T_3.

aging the NLI of LLMs is intended to improve semantic alignment with the human-level perception, particularly when estimating the extent of estimation of knowledge entailment and assessing similarities. This approach further enhances the robustness of the scoring scheme in EoT. In this appendix, we first define the triple labels of the statement-pair relationships. Then, we evaluate the effectiveness of NLI mechanism when using various types of LLMs and discuss the threats to the EoT framework given the use of the GPT-4 model.

E.1 Definition of Triple Labels in NLI Mechanism

In Section 3.2 in our submitted manuscript, EoT regards the NLI judgements as triple-label classification tasks as shown in Eq.(6). Specifically, let x^{pre} and t^{hyp} denote a premise and hypothesis statement, respectively. In a canonical NLI mechanism, there are three category labels to represent the relationship between any statement pair (x^{pre}, t^{hyp}), which is defined as follows.

Definition 4 Entailment: If hypothesis statement t^{hyp} is necessarily true or appropriate whenever the premise statement x^{pre} is true, we label the relationship between the statement pair (x^{pre}, t^{hyp}) as "Entailment". In other words, x^{pre} entails t^{hyp}.

Definition 5 Contradiction: If hypothesis statement t^{hyp} is necessarily false or inappropriate whenever the premise statement x^{pre} is true, we label the relationship between the statement pair (x^{pre}, t^{hyp}) as "Contradiction". In other words, t^{hyp} contradicts with x^{pre}.

Definition 6 Neutral: When neither "Entailment" nor "Contradiction" applies to relationship between the statement pair (x^{pre}, t^{hyp}), we label this relationship between (x^{pre}, t^{hyp}) as "Neutral". In other words, x^{pre} is neutral with t^{hyp}.

/**Logic of Reasoning Process Evolution**/

Hello, you are the intelligent robot of the smart Q&A project. Now, your task is to observe the given Context, Question, and Reference Answer, and refer to the existing reasoning process in the provided History and the quantitative scores of these reasoning processes to generate an optimized reasoning process that you believe can achieve higher scores. Specifically, each generated reasoning process is what you think you need to use or follow in terms of thoughts or logical reasoning process (i.e., Reasoning Process) when you approach the reliable Answer better as you answer the Question based on the Context.

Please note the following requirements when generating the optimized reasoning process:

1. Please think step by step to generate the reasoning process. Reflect your thinking method or thinking logic during this process.
2. Please note that the reasoning process you output should be an optimized result of the existing reasoning process previously provided in History, and should not repeat the reasoning process in History. Specifically, we have quantitatively scored each existing reasoning process provided in History in three dimensions:
   (1) Reliablity score: ranging from 0 to 1, this score represents the similarity between the answer generated by guiding you with the corresponding reasoning process when answering the Question based on the provided Contexts and the provided Reference Answer, the higher the better. The higher the score, the more reliably the corresponding reasoning process can guide the large model to generate answers similar to the target Answer. Specifically, the higher the score, the higher the degree to which the text meaning or semantics of the generated answer guided by reasoning process is at the target Answer.
   (2) Weighted Fidelity score: ranging from 0 to 1, a fidelity score is given for each step of thoughts in the reasoning process, the higher the better. This score for each step of thoughts represents the degree of fidelity of this thought when you reliably generate an answer similar to the Reference Answer. The higher the score, the greater the contribution of the corresponding thought to the goal of solving Question reliably, that is, the higher the score represents the higher degree of fidelity in employing the reasoning logic of the thought when answering the question, and the greater the benefit of approaching the Reference Answer due to using the logic of that thought.
   (3) Factuality score: ranging from 0 to 1, a factuality score is given for each step of thoughts in the reasoning process, the higher the better. This score for each thoght represents the degree of factuality of the content expressed in this thought. Specifically, the higher the score, the more the reasoning basis of the thought comes from the reference or understanding of the provided Context or Question, accordingly, the higher the score, the lower the probability that the reasoning basis of the thought comes from your own guess and the lower the probability of the reasoning basis deviating from the provided Context and Question.
3. Please ensure that the optimized reasoning process as a whole can achieve a higher overall reliablity score. Under this premise, please try to improve the weighted fidelity score and factuality score of each step of thought in your optimized reasoning process. Also, please note that the number and titles of thoughts you generate do not need to match those in the existing reasoning processes in History.
4. Please note that the purpose of generating a reasoning process is to guide the large model to provide an answer that reliablity solve the Question and ensure the factuality and fidelity of the reasoning behaviors in the reasoning process.
5.Please ensure not to answer the Question, nor to provide any optimization suggestions for the target Answer.
6.When generating the optimized reasoning process, please note that you need to output both the optimized reasoning process. The specific text content of each reasoning process is a sequence of multiple steps of thoughts, and each step of thoughts needs to provide a title.

/**Description of Complex Task**/

..... (Omitting. Please Reiew Figure 6)

/** Textual Description of Assessment of Existing Reasoning Processes**/

History:

Existing Reasoning Process R_1:

Thought T_1: Extract Key Concepts: We learn about the inverse problem, which involves statistical inference for systems like the XY model .....
Factuality Score of T_1: 1.0; Weighted Fidelity Score of T_1: 0.5265
Thought T_2: Relate Inverse Problem to Fields: In physics, inverse problems are crucial for understanding complex systems. Biology utilizes these .....
Factuality Score of T_2: 1.0; Weighted Fidelity Score of T_2: 0.1659
Thought T_3: Contextual Understanding: The inverse problem's relevance across these fields underscores its importance in extracting meaningful information .....
Factuality Score of T_3: 1.0; Weighted Fidelity Score of T_3: 0.1427
Thought T_4: Formulate Answer: Based on the the text, the inverse problem is encountered in diverse fields such as physics, biology, social sciences, finance, ...
Factuality Score of T_4: 1.0; Weighted Fidelity Score of T_4: 0.1724
Reliablity Score of R_1: 0.8843

..... (Please Reiew the Omitted Part of Thoughts in Figure 6)

Figure 9: Instance of a specific complete prompt to evolve the reasoning process in Figure 7, involving the instructions of evolution and the textual description of assessment results of reasoning process shown in Figure 7

Table 7: The Accuracy of NLI using LLMs

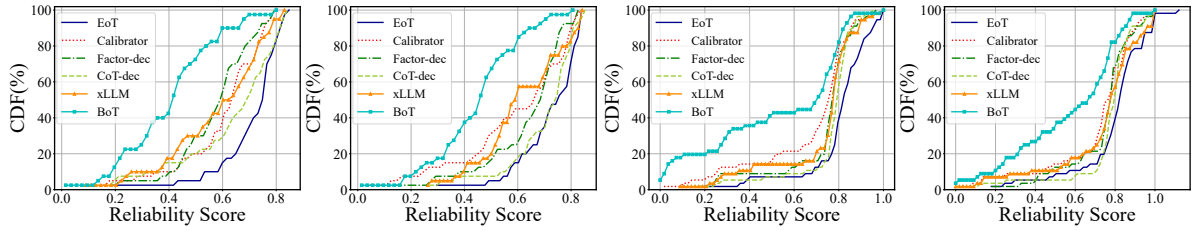
LLM Model	Accuracy(%)
Qwen2-72B	76.14
GPT-3.5 Turbo	67.64
GPT-4 Turbo	85.80

E.2 Performance of NLI using Various Types of LLMs

To evaluate the effectiveness of NLI in LLMs using few-shot prompts, we conduct experiments on the MNLI dataset (Williams et al., 2018). This dataset is well-known and comprises $40,000+$ samples of NLI tasks collected from dozens of different domains, including transcribed speech, fiction, and reports. For our evaluation, we select $10,000$ representative samples from the MNLI dataset, considering the scale and overhead of experiments. On this basis, we evaluate the NLI performance of three widely used LLMs: Qwen2-72B, GPT-3.5 Turbo and GPT-4 Turbo, using the selected samples.

Table 7 presents the accuracy performance of NLI achieved by the three LLMs. It is evident that GPT-4 Turbo outperforms the others in NLI tasks that require context from various domains. Moreover, based on previous studies (Gao et al., 2023) and our experiment results, we observe that GPT-4 Turbo achieves state-of-the-art accuracy in NLI. This suggests that using GPT-4 Turbo for NLI can lead to superior alignment between the judgements of LLMs and human assessments. Consequently, EoT employs GPT-4 Turbo with tailored few-shot learning to perform NLI by default, aiming to better align with human judgement.

Finally, we discuss the limitations of the NLI mechanism based on the GPT-4 Turbo. As we can see, the classification accuracy of NLI in EoT still needs to be improved, which could result in threats to the precise alignment of the assessment of EoT with that of humans. To further enhance EoT's capabilities, we intend to improve NLI performance in EoT by integrating novel LLMs with advanced NLI capabilities or optimizing the few-shot prompting mechanism.



(a) OpsQA + Qwen2-72B (b) OpsQA + GPT-4 Turbo (c) LongBench + Qwen2-72B (d) LongBench + GPT-4 Turbo

Figure 10: The CDF results of the reliability scores of evolved reasoning processes for various complex questions

F Experiment Results of Diverse Questions

In this appendix, we intuitively present the performance of the six evolution frameworks applied to various tasks within the OpsQA and LongBench datasets, utilizing two LLMs Qwen2-72B and GPT-4 Turbo. The results demonstrate that EoT significantly enhances the reasoning capabilities of diverse LLMs, as well as behavior fidelity and factuality of the explained reasoning processes across a broad spectrum of tasks.

F.1 Reasoning Capability (Reliability)

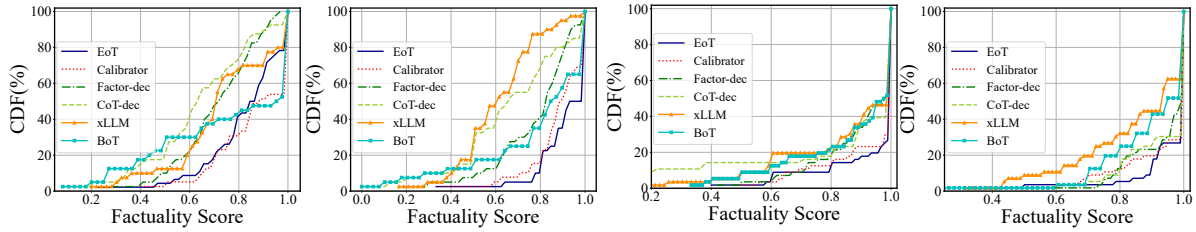
Figure 10a $\sim$ Figure 10d present the CDF results of reliability scores of reasoning processes evolved by the six frameworks on the two LLMs for all questions in the two datasets. Notably, EoT demonstrates superior reasoning capabilities compared to the five baselines for most tasks across diverse LLMs. For instance, in the case of the 40 questions in the OpsQA dataset, Figure 10a shows that EoT achieve a reliability score exceeding 0.6 for 85.0% of the questions. This performance surpasses that of BoT, xLLM, Calibrator, CoT-dec and Factor-dec by about 75.0%, 35.0%, 22.5%, 15.0% and 40.0%, respectively, when using Qwen2-72B. As for the 60 questions in the LongBench dataset, when using Qwen2-72B, Figure 10c illustrates that EoT achieves a reliability score exceeding 0.8 for about 61.67% of the questions. This performance surpasses that of BoT, xLLM, Calibrator, CoT-dec and Factor-dec by about 43.43%, 33.34%, 43.34%, 21.67% and 38.34%, respectively. In addition, when employing GPT-4 Turbo, Figure 10d indicates that EoT attains a reliability score higher than 0.8 for roughly 63.3% of these 60 questions, which exceeds the performance of BoT, xLLM, Calibrator, CoT-dec and Factor-dec by roughly 45.0%, 20.0%, 34.67%, 16.67% and 31.67%, respectively.

In summary, these results confirm that EoT achieves exceptional reliability across a wide range of tasks. This suggests that the improvement of reasoning capabilities in LLMs, facilitated by EoT, is well-suited for problem-solving in various domains with leading generality.

F.2 Factuality

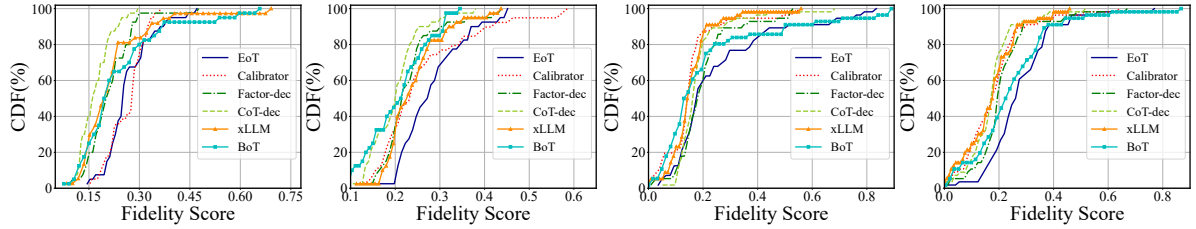
Figure 11a $\sim$ Figure 11d present the CDF results of factuality scores of reasoning processes evolved by the six frameworks on the two LLMs for all the questions in the two datasets. We find that, compared with the five baselines, EoT remarkably optimizes the factuality of reasoning processes for a wider range of questions on diverse LLMs. Specifically, as for the 40 questions in OpsQA, Figure 11b illustrates that EoT achieves the upper limit of factuality score 1.0 for about 50.0% of the questions, which exceeds the performance of BoT, xLLM, Calibrator, CoT-dec and Factor-dec by about 15.0%, 47.5%, 20.0%, 35.0% and 42.5%, respectively when using GPT-4 Turbo. As for the 60 questions in the LongBench dataset, when using Qwen-72B, Figure 11c illustrates that EoT achieves factuality scores of 1.0 for about 73.3% of the questions, which exceeds the performance of BoT, xLLM, Calibrator, CoT-dec and Factor-dec by about 25.0%, 20.0%, 3.3%, 13.3% and 26.7%, respectively. Moreover, when using GPT-4 Turbo, Figure 11d shows that EoT obtains factuality scores of 1.0 for 75% of these 60 questions, which surpasses the performance of BoT, xLLM, Calibrator, CoT-dec and Factor-dec by about 26.7%, 36.7%, 3.3%, 5.0% and 25.0%, respectively.

These results suggest that EoT demonstrates remarkable generality in ensuring the factuality of reasoning processes, proving to be highly robust across various knowledge domains. This assurance stems from the detailed evaluation of thoughts and the effective evolving framework that enables LLMs to eliminate non-factual errors during the explanation of these thoughts.



(a) OpsQA + Qwen2-72B (b) OpsQA + GPT-4 Turbo (c) LongBench + Qwen2-72B (d) LongBench + GPT-4 Turbo

Figure 11: The CDF results of the factuality scores of evolved reasoning processes for various complex questions



(a) OpsQA + Qwen2-72B (b) OpsQA + GPT-4 Turbo (c) LongBench + Qwen2-72B (d) LongBench + GPT-4 Turbo

Figure 12: The CDF results of the fidelity scores of evolved reasoning processes for various complex questions

F.3 Fidelity

Figure 12a $\sim$ Figure 12d illustrate the behavior fidelity performance of reasoning processes evolved by the six frameworks on the two LLMs for all the questions in the two datasets. Compared to the five baselines, EoT implements an evolution mechanism that guides diverse LLMs in generating reasoning processes, resulting in state-of-the-art performance in behavior fidelity across most tasks from various knowledge domains. These enhancements demonstrate the exceptional generality of EoT in improving behavior fidelity. Specifically, as for the 40 questions in the OpsQA dataset, when using Qwen2-72B, Figure 12a illustrates that EoT obtains fidelity scores higher than 0.225 for 71.7% of the questions, which exceeds the performance of BoT, xLLM, Calibrator, CoT-dec and Factor-dec by about 36.7%, 53.3%, 5.0%, 58.3% and 43.3%, respectively. Moreover, when using GPT-4 Turbo, Figure 12b presents that EoT achieves fidelity scores higher than 0.25 for 60.0% of these 40 questions, which surpasses the performance of BoT, xLLM, Calibrator, CoT-dec and Factor-dec by roughly 32.5%, 20.0%, 25.0%, 50.0% and 37.5%, respectively. As for the 60 questions in the LongBench dataset, when using Qwen2-72B, Figure 12c presents that EoT achieves fidelity scores higher than 0.2 for 40% of the questions, which surpasses the performance of BoT, xLLM, Calibrator, CoT-dec and Factor-dec by roughly 10%, 26.7%, 26.7%%, 21.7% and 6.7%, respectively. In addition, when using GPT-4 Turbo, Figure 12d illustrates that EoT obtains fidelity scores higher than 0.2 for 75.0% of the questions, which exceeds the performance of BoT, xLLM, Calibrator, CoT-dec and Factor-dec by about 16.7%, 41.7%, 36.7%, 51.7% and 26.7%, respectively.

These results demonstrate that EoT significantly improves the fidelity of reasoning processes as explained by various LLMs, and this improvement is broadly applicable across diverse tasks rooted in different domains of knowledge. In summary, this enhancement in behavior fidelity can be attributed to two key factors. Firstly, building on prior studies, EoT implements an effective mechanism for assessing the behavior fidelity of explained reasoning thoughts in a detailed manner. Secondly, the evolution mechanism of EoT effectively encourages LLMs to clarify their reasoning with enhanced faithfulness.

The optimization of reasoning processes regarding factuality and fidelity, as shown in Appendices F.2 and F.3, is beneficial for further enhancing the reasoning capabilities of LLMs, as illustrated in Appendix F.1.

G Ablation Study

G.1 Setup

We conduct ablation studies for EoT on the OPSQA dataset. To evaluate the effectiveness of the optimization achieved by EoT concerning the three factors, namely reliability, factuality and behavior fidelity of reasoning processes, we design three variants of EoT: EoT_w/o_fact, EoT_w/o_fide, and

Table 8: The performance on reliability for reasoning processes evolved by EoT and its variants using two LLMs on the OpsQA dataset.

LLMs	Models	OpsQA			
		BLEURT	ROUGE-L	NLI results $entail(\%)$	Reliability Score
Qwen2	EoT_w/o_fact	0.593	0.362	70.16	0.607
	EoT_w/o_fide	0.602	0.377	76.32	0.651
	EoT_w/o_reli	**0.609**	0.386	77.33	0.655
	EoT	0.595	**0.395**	**83.90**	**0.706**
GPT-4	EoT_w/o_fact	0.573	0.403	78.75	0.658
	EoT_w/o_fide	0.597	0.417	80.03	0.675
	EoT_w/o_reli	0.593	0.402	75.84	0.634
	EoT	**0.602**	**0.427**	**88.58**	**0.717**

[1] Qwen2 and GPT-4 represent the LLMs of Qwen2-72B and GPT-4 Turbo, respectively.

Table 9: The performance on factuality and fidelity for reasoning processes evolved by EoT and its variants using two LLMs on the OpsQA dataset.

LLMs	Models	OpsQA	
		Factuality	Fidelity
		S_{fac}	S_{fid}
Qwen2	EoT_w/o_fact	0.737	0.231
	EoT_w/o_fide	**0.840**	0.226
	EoT_w/o_reli	0.723	**0.272**
	EoT	0.823	0.267
GPT-4	EoT_w/o_fact	0.722	0.277
	EoT_w/o_fide	0.869	0.262
	EoT_w/o_reli	0.882	**0.296**
	EoT	**0.906**	0.281

EoT_w/o_reli. Each variant removes the optimization related to factuality, behavior fidelity, and reliability respectively from the evolution of thoughts. We then compare the performance of the canonical EoT with that of these three variants.

Table 8 presents the reasoning capability of Qwen2-72B and GPT-4 Turbo equipped with the canonical EoT and its three variants. Additionally, Table 9 shows the fidelity and factuality performance of the reasoning processes evolved by the EoT and its variants.

G.2 Effectiveness of Factuality Optimization

We compare the performance of EoT with that of EoT_w/o_fact to assess the effectiveness of evolution in terms of factuality. First of all, it can be observed that EoT significantly enhances the factuality of reasoning processes through the fine-grained evolution explicitly optimizing factuality. For instance, Table 9 shows that, when using GPT-4 Turbo, EoT increases the factuality score by about 25.5% compared to EoT_w/o_fact. Furthermore, the experiment results validate that the optimization on the factuality of reasoning processes in EoT further enhances the reasoning capabilities of diverse LLMs. For instance, Table 8 demonstrates that compared to EoT_w/o_fact, EoT improves BLEURT, ROUGE-L and NLI result $entail(\%)$ by about 5.1%, 6.0%, and 12.5% respectively, and finally obtains the improvement of reliability score by about 9.0% when using GPT-4 Turbo.

G.3 Effectiveness of Fidelity Optimization

We compare the performance of EoT with that of EoT_w/o_fide, and then there are two insights can be observed, which indicates the effectiveness of behavior fidelity evolution in EoT. Firstly, compared to EoT_w/o_fide, EoT produces reasoning processes that achieve a significantly improved behavior fidelity. Specifically, as shown in Table 9 , when using Qwen2-72B, EoT increases the fidelity score by about 18.1%. This enhancement is attributed to the fine-grained assessment of fidelity of reasoning thoughts and the effective prompting scheme aimed at faithfully explaining the behaviors of LLMs.

Secondly, the mechanism that enhances behavior fidelity in EoT facilitates the positive emergence of reasoning capabilities in LLMs. For instance, compared to EoT_w/o_fide, EoT increases the reliability score by about 8.5% when using Qwen2-72B. In detail, this improvement is mainly attributed to that the fidelity enhancement achieved by EoT simultaneously enhances the reasoning capabilities of LLMs from aspects of token overlapping and semantic alignment with human judgement. Specifically, EoT improves ROUGE-L and NLI result $entail(\%)$ by about 4.8% and 9.9% respectively when using Qwen2-72B.

G.4 Effectiveness of Reliability Optimization

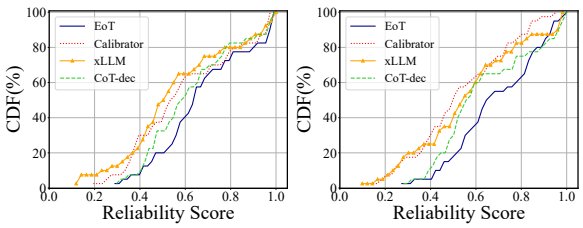
We assess the performance of EoT in comparison to EoT_w/o_reli to determine the effectiveness of the evolution incorporated in the canonical EoT, which aims to directly improve the overall reliability of reasoning processes. As shown in Table 8, EoT increases the reliability score by about 13.1% when using GPT-4 Turbo, compared to EoT_w/o_reli. For further details, EoT improves the BLEURT, ROUGE-L and $entail(\%)$ by about 1.5%, 6.2% and 16.8% respectively. These findings suggest that explicitly addressing the reliability factor can lead to a significant improvement in the overall reasoning capabilities of LLMs, in contrast to the unsupervised evolution that does not consider the solving reliability under the guideline of previously produced reasoning processes.

G.5 Effectiveness of Multi-objective Optimization on Three Factors

According to the experiment results of ablation studies as mentioned above, it can be observed that EoT achieves the best collaborative optimization among the three factors compared with the three variants. In other words, when considering the three factors simultaneously, EoT produces the evolved reasoning processes that achieves the best trade-off among the three factors under the promise of reasoning capability improvement on LLMs.

Firstly, excluding a factor from EoT can make evolved reasoning processes achieve a better performance on some other factors. However, this improvement of performance on other factors is always along with the degradation of the performance on the excluded factor, and would further compromise reasoning capabilities of LLMs. For instance, as shown in Table 8 and Table 9, when using Qwen2-72B, EoT_w/o_fide improves the factuality performance by about 2.1% compared to EoT. However, EoT attains the improvement on fidelity performance and the overall reasoning capabilities by about 18.1% and 8.5% respectively.

Secondly, evolving reasoning process in terms of fidelity and factuality in a unsupervised way, can effectively optimize the behavior fidelity of reasoning processes produced by diverse LLMs. Nevertheless, EoT which evolves reasoning process with supervised awareness of reliability performance further enhance the factuality of produced reasoning processes and the reasoning capability of LLMs simultaneously. Specifically, compared to EoT, EoT_w/o_reli improves the fidelity score by about 1.9% and 5.3% when using Qwen2-72B and GPT-4 Turbo, respectively. In contrast to that, EoT improves the factuality score and reliability score by about 13.8% and 7.8% when using Qwen2-72B, and by about 2.7% and 13.1% when using GPT-4 Turbo, respectively.



(a) Qwen3-32B (b) Qwen3-235B

Figure 13: The CDF results of the reliability scores of evolved reasoning processes for 40 complex tasks

H Additional Experiments to Validate the Generality of EoT

In this appendix, we conduct additional experiments for EoT using Qwen3-32B and Qwen3-253B on 40 new QA tasks sourced from real operational maintenance (O&M) systems of Alibaba's cloud clusters. These tasks cover application compilation errors, Docker containers deployment, management of access permission in workflows, and more. We compare the performance of EoT with that of three outstanding frameworks, namely xLLM, Calibrator and CoT-dec. Combining the results from Section 4.2 and this appendix, we examine the generality of EoT from three key aspects: 1) the effectiveness on diverse, knowledge-intensive reasoning tasks; 2) scalability across LLMs of varying sizes; and 3) adaptability to tasks with different knowledge scales.

Table 10: The reliability performance for reasoning processes evolved by EoT and baselines using LLMs of diverse parameter sizes on additional complex tasks.

LLMs	Models	Additional OM Tasks			
		BLEURT	ROUGE-L	NLI results *entail*(%)	Reliability Score
Qwen3-32B	xLLM	0.637	0.457	59.12	0.546
	Calibrator	0.679	0.528	64.31	0.585
	CoT-dec	0.681	0.494	69.94	0.616
	EoT (ours)	**0.699**	**0.543**	**73.87**	**0.659**
Qwen3-235B	xLLM	0.668	0.494	60.82	0.553
	Calibrator	0.627	0.452	58.34	0.531
	CoT-dec	0.715	0.530	67.56	0.630
	EoT	**0.736**	**0.577**	**77.83**	**0.686**

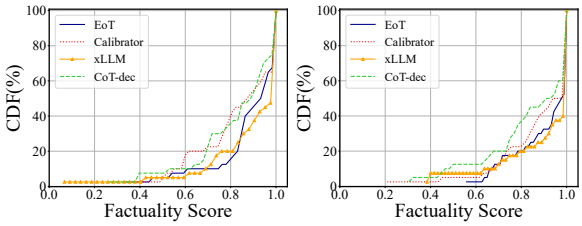
H.1 Effectiveness of EoT on Additional Knowledge-intensive Tasks

Reasoning Capability Table 10 presents the average reasoning performance of Qwen3-32B and Qwen3-235B, when applying the four evolution frameworks. The results indicate that the EoT outperforms its counterparts in three aspects. First, compared to the leading baseline, CoT-dec, EoT improves reliability scores on the 40 questions by about 7.0% and 8.9% using Qwen3-32B and Qwen3-235B respectively. Second, in terms of sensitivity to token and semantic variation, EoT increases ROUGE-L and BLEURT scores by about 5.9% and 2.8% on average with Qwen3 models. Third, for semantic alignment robustness, EoT raises *entail*(%) by around 5.6% and 15.2% using Qwen3-32B and Qwen3-235B respectively. These findings further demonstrate that EoT effectively enhance the reasoning capabilities of LLMs for diverse, knowledge-intensive reasoning tasks.

Moreover, Figure 13 visually illustrates the reasoning abilities of the two Qwen-3 models empowered by the four evolution frameworks on the 40 complex tasks. It can be observed that, when using Qwen3 models, EoT achieves a reliability score > 0.6 for 63.75% of tasks on average. This surpasses xLLM, Calibrator, and CoT-dec by approximately 27.5%, 27.5%, and 20%, respectively. These findings further confirm the strong generality of EoT in improving reasoning capabilities across different tasks.

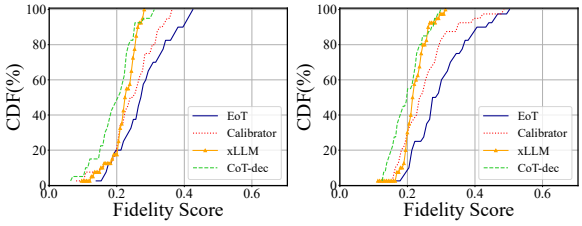
Table 11: The performance on factuality and fidelity for reasoning processes evolved by EoT and baselines using LLMs of varying scales on additional tasks.

LLMs	Models	Additional OM Tasks	
		Factuality	Fidelity
		S_{fac}	S_{fid}
Qwen3-32B	xLLM	**0.884**	0.220
	Calibrator	0.831	0.242
	CoT-dec	0.822	0.193
	EoT	0.879	**0.274**
Qwen3-235B	xLLM	0.900	0.221
	Calibrator	0.877	0.248
	CoT-dec	0.833	0.199
	EoT	**0.907**	**0.298**



(a) Qwen3-32B (b) Qwen3-235B

Figure 14: The CDF results of the factuality scores of evolved reasoning processes for 40 complex tasks



(a) Qwen3-32B (b) Qwen3-235B

Figure 15: The CDF results of the fidelity scores of evolved reasoning processes for 40 complex tasks

Factuality and Fidelity One of the key innovations of EoT is its ability to achieve collaborative optimization between the factuality and behavior fidelity of thoughts and the reasoning capabilities of LLMs. In this sub-appendix, we focus on evaluating the generality of EoT in enhancing the factuality and fidelity of thoughts. Table 11 reports the average fidelity and factuality performance of reasoning processes evolved by the four frameworks, using Qwen3-32B and Qwen3-235B on the 40 additional tasks. These results highlight the effectiveness of EoT in different knowledge-intensive tasks from two aspects.

First, EoT significantly enhances the behavior fidelity of LLMs, demonstrating a leading ability to reduce non-factual errors in thoughts. Compared to the leading baseline Calibrator, EoT increases the fidelity score by approximately 13.2% and 20.1% on the 40 tasks when using Qwen3-32B and Qwen3-235B respectively. Second, in terms of factuality, with Qwen3 models, EoT achieves an average factuality score that is extremely close to that of the leading baseline xLLM. These results confirm the strong adaptability of EoT in enhancing both factuality and behavior fidelity across various question-answering tasks. Furthermore, by combining the results from Table 10 and Table 11, it is evident that the collaborative optimization among the three key factors of reasoning demonstrates significant robustness on a wide range of tasks and different LLMs.

Additionally, Figure 14 and Figure 15 intuitively show the CDF results for the factuality and behavior fidelity of reasoning processes evolved by the four frameworks across the 40 complex tasks, respectively. We find that, when using Qwen3 models, EoT achieves a factuality score > 0.8 for an average of 82.5% of tasks. This surpasses xLLM, Calibrator and CoT-dec by approximately 1.25%, 13.75%, and 18.75%, respectively. Meanwhile, EoT achieves a fidelity score > 0.25 for an average of 63.75% of tasks, outperforming xLLM, Calibrator and CoT-dec by around 40.0%, 22.5% and 48.75% respectively. These findings further confirm that EoT can effectively improve both the factuality and fidelity of evolved thoughts across a wide range of questions.

H.2 Scalability Across LLMs of Varying Sizes

As shown in Table 10 and Table 4 in Section 4, compared to the leading baseline, EoT improves the average reliability score by approximately 7.0%, 8.8%, and 8.9% when using Qwen3-32B, Qwen2-72B, and Qwen3-235B, respectively. These results indicate that EoT effectively enhances the reasoning capabilities of LLMs across different parameter sizes.

Additionally, Table 11 and Table 5 show that, compared to leading counterparts, EoT improves

Table 12: The performance of six frameworks on the three task categories in terms of reliability, factuality and fidelity.

LLMs	Models	Short			Medium			Long		
		Reliability	Factuality	Fidelity	Reliability	Factuality	Fidelity	Reliability	Factuality	Fidelity
		S_{rel}	S_{fac}	S_{fid}	S_{rel}	S_{fac}	S_{fid}	S_{rel}	S_{fac}	S_{fid}
Qwen2	BoT	0.435	0.797	0.211	0.563	0.862	0.229	0.510	0.884	**0.246**
	xLLM	0.616	0.749	0.196	0.774	0.898	0.151	0.668	0.881	0.150
	Calibrator	0.608	0.861	0.217	0.729	**0.971**	0.152	0.657	0.926	0.149
	CoT-dec	**0.721**	0.740	0.170	0.743	0.797	0.175	0.744	0.716	0.178
	Factor-dec	0.619	0.792	0.204	0.741	0.900	0.198	0.625	0.770	0.196
	EoT (ours)	0.719	**0.864**	**0.265**	**0.829**	0.945	**0.251**	**0.851**	**0.929**	0.208
GPT-4	BoT	0.467	0.815	0.230	0.631	0.869	0.206	0.540	0.949	0.213
	xLLM	0.636	0.680	0.218	0.762	0.876	0.152	0.695	0.829	0.187
	Calibrator	0.623	0.899	0.229	0.744	0.952	0.187	0.673	0.926	0.182
	CoT-dec	0.701	0.766	0.176	0.818	0.940	0.167	0.764	0.928	0.180
	Factor-dec	0.653	0.827	0.215	0.752	0.950	0.208	0.788	0.893	0.220
	EoT (ours)	**0.732**	**0.917**	**0.287**	**0.825**	**0.966**	**0.278**	**0.794**	**0.963**	**0.227**

the behavior fidelity of evolved thoughts by about 13.2%, 10.8% and 20.1% when using the Qwen models with 32B, 72B and 235B parameters, respectively. Meanwhile, EoT achieves factuality performance that is close to the leading baselines across these LLMs of different scales. Overall, these results confirm that EoT demonstrates strong scalability when applied to LLMs with a wide range of parameter sizes.

H.3 Generality of EoT under Varying Context Length

As shown in Table 2, we have conducted experiments on hundreds of tasks which address complex reasoning within contexts ranging from thousands to tens of thousands of words. To further examine the generality of EoT for tasks having varying knowledge intensity, this sub-appendix investigates the effectiveness of EoT under the different conditions of context length. Specifically, our overall evaluation presented in Section 4 involves 100 different tasks, and we classify them into three categories: 1) **short** tasks, each of which contains context fewer than 5,000 words; 2) **medium** tasks, each of which has context more than 5,000 words and fewer than 10,000 words; and 3) **long** tasks, each of which involves context more than 10,000 words. The **short**, **medium** and **long** categories account for 64%, 24% and 12% of the 100 tasks respectively.

Table 12 presents the performance of EoT and five baselines on the three task categories in terms of reliability, factuality, and fidelity. It can be observed that, when using GPT-4, EoT outperforms all five baselines across all three categories of tasks with respect to these metrics. Additionally, when using Qwen2-72B to handle tasks with varying context lengths, EoT still achieves the best overall performance in collaborative optimization across reliability, factuality, and fidelity. These results indicate that EoT demonstrates strong adaptability to tasks with different knowledge scales.

I Significance of Performance Improvement or Decrease

In this appendix, we use the T-test, an appropriate method of statistical test, to evaluate the significance of performance improvement or decrease presented in Table 4 and Table 5 in our submitted manuscript. In general, for each of the five baselines, we compute the p-value in T-test between the distribution of each performance metric obtained by the baseline and EoT respectively when using each type of LLMs on each dataset. Specifically, the performance metric includes BLEURT, ROUGE-L, NLI score and reliability score for reasoning capability evaluation and the factuality score and fidelity score for the evaluation of factuality and behavior fidelity of reasoning processes. When computing the p-value of any metric for each baseline on a specific dataset, the performance distribution is estimated among the metric value for each question in the dataset. A lower p-value represents a higher significance for performance improvement and decrease, and p-value ≤ 0.05 often indicates that the evaluation results achieve the sufficient significance in statistical tests.

As for the performance metrics reflecting the reasoning capability of LLMs, namely BLEURT, ROUGE-L, NLI results and reliability score, As presented in Table 13, the p-value between performance achieved by the baselines and that obtained by EoT can be limited below 0.05 in most scenarios. These results indicate that the performance improvement or decrease achieved by EoT in terms of reasoning capability of LLMs which are presented in our submitted manuscript, have the outstanding significance. Thus, our conducted evaluations and the conclusion that EoT effectively achieves the

Table 13: The p-value of performance improvement or decrease in terms of reasoning capability of LLMs between the five baselines and EoT in T-test.

LLMs	Models	OpsQA				LongBench			
		BLEURT	ROUGE-L	NLI results $entail(\%)$	Reliability Score	BLEURT	ROUGE-L	NLI results $entail(\%)$	Reliability Score
Qwen2	BoT	8.602×10^{-5}	3.720×10^{-6}	6.449×10^{-10}	6.463×10^{-10}	6.007×10^{-9}	2.843×10^{-7}	3.143×10^{-6}	2.308×10^{-9}
	xLLM	0.011	0.003	1.362×10^{-5}	3.263×10^{-7}	1.561×10^{-9}	1.598×10^{-7}	0.026	0.001
	Calibrator	0.028	0.036	0.006	9.268×10^{-6}	1.032×10^{-10}	8.322×10^{-8}	0.003	3.122×10^{-7}
	CoT-dec	0.023	0.043	0.008	0.043	1.690×10^{-7}	6.397×10^{-5}	0.221	0.038
	Factor-dec	0.014	1.343×10^{-5}	1.657×10^{-6}	2.380×10^{-7}	5.173×10^{-9}	2.270×10^{-6}	0.030	0.003
GPT-4	BoT	9.132×10^{-7}	1.546×10^{-7}	2.110×10^{-9}	5.074×10^{-11}	5.971×10^{-9}	2.993×10^{-9}	1.892×10^{-6}	3.809×10^{-9}
	xLLM	0.036	0.045	5.120×10^{-8}	3.042×10^{-8}	0.047	0.034	0.013	9.875×10^{-4}
	Calibrator	0.034	0.037	3.330×10^{-4}	2.664×10^{-4}	4.896×10^{-5}	1.813×10^{-4}	7.135×10^{-4}	6.534×10^{-6}
	CoT-dec	0.014	0.446	0.039	0.040	0.005	0.041	0.235	0.043
	Factor-dec	0.001	1.073×10^{-8}	0.001	7.063×10^{-4}	5.572×10^{-6}	3.859×10^{-5}	3.864×10^{-4}	0.001

enhancement of reasoning capability have a reasonable robustness. Nevertheless, we find that p-value significantly exceeding 0.05 only occurs between the value of NLI results achieved by the CoT-dec and EoT respectively on LongBench dataset. This is mainly attributed to that CoT-dec and EoT obtains the similar performance of reasoning capability on the aspects of semantic alignment robustness on LongBench dataset. Since EoT achieves adequate significance of performance improvement for all the remaining metrics i.e., BLEURT, ROUGE-L and reliability score, it still can be regarded that EoT enhances the reasoning capability of LLMs compared with CoT-dec in a robust way on LongBench dataset. In future, we will evaluate the performance difference between NLI results of CoT-dec and EoT on a larger scale of datasets.

Table 14: The p-value of performance improvement or decrease in terms of factuality and fidelity for reasoning processes of reasoning processes between the five baselines and EoT in T-test.

LLMs	Models	OpsQA		LongBench	
		Factuality S_{fac}	Fidelity S_{fid}	Factuality S_{fac}	Fidelity S_{fid}
Qwen2	BoT	0.039	0.041	0.017	0.045
	xLLM	0.003	1.215×10^{-4}	1.590×10^{-4}	2.047×10^{-5}
	Calibrator	4.258×10^{-4}	0.046	0.072	2.550×10^{-4}
	CoT-dec	1.142×10^{-9}	3.172×10^{-13}	0.007	0.002
	Factor-dec	8.081×10^{-5}	1.488×10^{-6}	0.045	0.032
GPT-4	BoT	0.001	2.604×10^{-6}	0.009	0.027
	xLLM	3.667×10^{-16}	1.010×10^{-4}	2.988×10^{-6}	1.470×10^{-9}
	Calibrator	0.039	0.041	0.087	6.041×10^{-6}
	CoT-dec	1.644×10^{-7}	8.337×10^{-7}	0.074	1.372×10^{-6}
	Factor-dec	8.267×10^{-5}	4.683×10^{-5}	0.062	8.401×10^{-6}

As for the performance metrics mirroring factuality and behavior fidelity of evolved reasoning processes, It can be observed in Table 14 that, on OpsQA dataset, p-value ≤ 0.05 occurs for each metric achieved by the five baselines in our evaluation. Nevertheless, on LonBench dataset, we find that p-value for factuality score achieved by Calibrator, CoT-dec and Factor-dec slightly exceeds 0.05 when using GPT-4 Turbo. This is because that these three baselines and EoT all obtain the outstanding but close performance on factuality of reasoning processes. We will use more open-source dataset to further evaluate the performance difference on factuality among these four frameworks in the future.