Exclusion of Thought: Mitigating Cognitive Load in Large Language Models for Enhanced Reasoning in Multiple-Choice Tasks

Qihang Fu[1,2] , **Yongbin Qin**[1,2 ✉] , **Ruizhang Huang**[1,2 ✉] , **Yanping Chen**[1,2]
Yulin Zhou[1,2] , **Lintao Long**[1,2]
[1] State Key Laboratory of Public Big Data, Guizhou University,
[2] Text Computing & Cognitive Intelligence Engineering Research Center of National Education Ministry,
College of Computer Science and Technology, Guizhou University, Guiyang 550025, China
qihangfoo@gmail.com {ybqin,rzhuang}@gzu.edu.cn

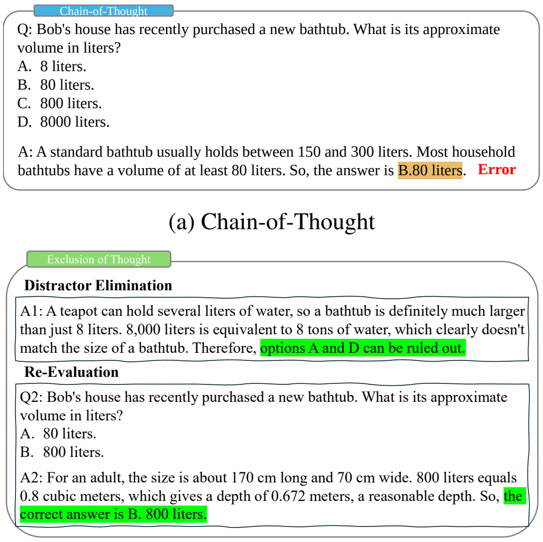
Abstract

Multiple-choice questions (MCQs) are a widely used and vital assessment format for evaluating large language models (LLMs). This study reveals that LLMs are susceptible to "cognitive load" caused by distractor options in MCQs, leading to excessive attention to distractors and consequent vacillation between correct and incorrect options. To mitigate this cognitive burden, we introduce a novel reasoning prompt strategy, called **EoT**, which effectively reduces cognitive load by steering the model's attention away from erroneous options. This enables the model to focus more effectively on reasonable answers. Additionally, by documenting the elimination process, EoT enhances the transparency and interpretability of the model's reasoning. Experimental results demonstrate that EoT, as a plug-and-play approach, significantly reduces cognitive load and improves performance, showcasing its potential to enhance both the accuracy and interpretability of LLMs[1].

1 Introduction

Multiple-choice questions (MCQs) represent a widely adopted task format in large language models (LLMs). These questions span diverse domains and exhibit varying levels of complexity, typically comprising a question and several candidate options from which the model must select the most appropriate answer. MCQs are extensively employed both for benchmarking LLM performance (Hendrycks et al., 2020; Zhong et al., 2023) and as the basis for automated evaluation frameworks (Zheng et al., 2023b). Consequently, it is critical for LLMs to reliably select correct answers in these tasks.

However, we have observed that LLMs exhibit excessive focus on distractor options in MCQs. As illustrated in Figure 1a, in zero-shot MCQ tasks, the presence of distractors and low-probability options often prevents the model from concentrating on the more challenging options, leading to predictions that appear correct but are ultimately erroneous. In Table 1, we demonstrate that removing incorrect options or simply adding obviously incorrect options can cause significant fluctuations in model performance. This suggests that LLMs are highly sensitive to the presence of distractors. For example, after adding distractor options to the AQuA dataset (Ling et al., 2017), the model's performance decreased by 5.34 percent (29.92 vs 24.58). In contrast, removing a distractor option from the GSM8K-MC dataset improved performance by 21.54 percent (35.61 vs 57.15) (Zhang



Chain-of-Thought

Q: Bob's house has recently purchased a new bathtub. What is its approximate volume in liters?
A. 8 liters.
B. 80 liters.
C. 800 liters.
D. 8000 liters.

A: A standard bathtub usually holds between 150 and 300 liters. Most household bathtubs have a volume of at least 80 liters. So, the answer is B.80 liters. Error

(a) Chain-of-Thought

Exclusion of Thought

Distractor Elimination

A1: A teapot can hold several liters of water, so a bathtub is definitely much larger than just 8 liters. 8,000 liters is equivalent to 8 tons of water, which clearly doesn't match the size of a bathtub. Therefore, options A and D can be ruled out.

Re-Evaluation

Q2: Bob's house has recently purchased a new bathtub. What is its approximate volume in liters?
A. 80 liters.
B. 800 liters.

A2: For an adult, the size is about 170 cm long and 70 cm wide. 800 liters equals 0.8 cubic meters, which gives a depth of 0.672 meters, a reasonable depth. So, the correct answer is B. 800 liters.

(b) Exclusion of Thought

Figure 1: The presence of distractor options increases the cognitive load of large language models, causing them to select seemingly reasonable but incorrect choices. The process of elimination reduces this cognitive load by removing distractors, allowing the model to infer the correct answer.

[1] This work is open sourced at: https://github.com/QihangFoo/EoT.
✉Corresponding author.

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 21673–21686
July 27 - August 1, 2025 ©2025 Association for Computational Linguistics

et al., 2024b). This issue may arise because distractors are often crafted with linguistic ambiguity, making incorrect options seem plausible or partially correct, thereby increasing the difficulty of distinguishing the correct answer. Additionally, biases in the training data may influence the model to develop an incorrect preference for linguistic patterns associated with distractors, further exacerbating their negative impact on reasoning processes.

Dataset	Orig	+1	+2	-1	-2
MMLU-PRO	37.43	35.24 (-2.19)	33.02 (-4.41)	39.21 (+1.78)	40.65 (+3.22)
ARC	82.40	81.58 (-0.82)	80.86 (-1.54)	85.75 (+3.35)	91.50 (+9.10)
CSQA	76.79	74.41 (-2.38)	73.12 (-3.61)	80.26 (+3.47)	83.22 (+6.43)
GSM8K-MC	35.61	33.81 (-1.80)	31.24 (-4.37)	44.67 (+9.06)	57.15 (+21.54)
AQuA	29.92	24.58 (-5.34)	23.66 (-6.26)	33.46 (+3.54)	46.45 (+16.53)

Table 1: Removing incorrect options or simply adding clearly incorrect options in multiple-choice questions can lead to significant fluctuations in model performance (5-shot Llama-3-8B-Instruct). This indicates that large language models are highly sensitive to the presence of distractor options.

This tendency to overemphasize distractors is pervasive in LLMs and cannot be effectively mitigated through straightforward prompting strategies. For instance, while Chain of Thought (CoT) (Wei et al., 2022b) has been successful in improving the model's reasoning ability, methods like PoE (Ma and Du, 2023) and its derivatives (Balepur et al., 2024), which use a two-step scoring approach to let the model score itself before proceeding with further reasoning, and integrated prompting (Zhang et al., 2024a; Tong et al., 2023), which guides LLMs through nonlinear thinking to eliminate incorrect options, offer some improvements in selection accuracy but remain vulnerable to distractors. These approaches still do not fully resolve the problem of overattention to distractor options.Traditional prompting strategies predominantly follow an "additive" approach by introducing more reasoning-related information into the prompt (Rai and Yao, 2024; Li et al., 2024). However, this additive strategy does not prevent the model from attending to distractors, as these options remain visible and continue to consume the model's attention.

To address this challenge, we draw inspiration from classical human strategies for solving MCQs and propose a novel prompting framework: Exclusion of Thought (EoT). As depicted in Figure 1b, EoT adopts a "subtractive" approach, fundamentally distinct from the additive nature of traditional strategies. The core idea behind EoT is to incrementally eliminate evidently incorrect options, thereby narrowing the focus to the remaining, more challenging options and reducing the influence of distractors on the reasoning process. Specifically, EoT introduces a confidence-based exclusion mechanism to help the model systematically discard incorrect options (Tian et al., 2023), optimizing its reasoning trajectory and minimizing the attention allocated to distractors. Experimental results demonstrate that EoT significantly outperforms existing methods across multiple MCQs benchmark datasets, particularly in high-distractor settings, where the model's resistance to interference is substantially enhanced. By incorporating the EoT framework, we not only improve the accuracy of LLMs on MCQ tasks but also provide new insights and technical support for tackling more complex reasoning challenges.

Our contributions are summarized as follows

- We propose a novel prompting strategy, EoT, inspired by human problem-solving techniques. By incrementally eliminating incorrect options, EoT effectively alleviates the issue of excessive attention to distractors in LLMs.

- EoT introduces a transparent decision-making process by recording the exclusion steps, offering new perspectives for understanding the reasoning logic of LLMs.

- We demonstrate that EoT achieves substantial performance improvements on multiple MCQs benchmark datasets, particularly enhancing the model's robustness in high-distractor tasks.

2 Related Works

2.1 Multiple Choice Questions

As a concise and widely utilized task format, MCQs play a pivotal role in evaluating the capabilities of LLMs. This format is prevalent in numerous benchmark datasets, which serve to assess model comprehension, reasoning, and domain-specific knowledge. Early efforts include AQuA (Ling et al., 2017), targeting algebraic problem-solving, and ARC (Clark et al., 2018), which focuses on scientific reasoning. CommonsenseQA

(Talmor et al., 2018) introduced commonsense reasoning tasks, while MMLU (Hendrycks et al., 2020) expanded evaluations to general and specialized knowledge across diverse subjects. Recent advancements include MMLU-Pro (Wang et al., 2024) for professional-level tasks and GSM8K-MC (Zhang et al., 2024b), emphasizing complex reasoning through a multi-choice adaptation of GSM8K. These datasets are designed with carefully crafted distractors to challenge model accuracy, making them a standard for assessing LLM performance across varying domains and complexities.

2.2 Large Language Models

Since the release of GPT-3, LLMs such as GPT-4 and ChatGPT have shown significant advancements in reasoning and understanding, using techniques like reinforcement learning from human feedback (RLHF) (Ouyang et al., 2022) and instruction-tuned fine-tuning (Wei et al., 2021). Meta's Llama series (Touvron et al., 2023) achieved strong performance with smaller parameter sizes, while the open-source o1 model demonstrated outstanding reasoning capabilities through multi-turn chain-of-thought reasoning (Zhao et al., 2024).

2.3 Prompting Strategies

Prompting strategies are crucial for enhancing the complex reasoning abilities of LLMs. Chain-of-Thought guides the model to generate a series of intermediate reasoning steps (Wang et al., 2023; Zhao et al., 2023; Taori et al., 2023), allowing it to analyze the problem incrementally and ultimately arrive at the correct answer (Zhou et al., 2023). This approach has significantly improved performance on complex reasoning tasks without requiring additional training(Huang and Chang, 2022; Min et al., 2022). This method leverages the emergent capabilities of LLMs, enabling the model to tackle complex problems by generating and utilizing reasoning chains (Qiao et al., 2022; Chu et al., 2024; Wei et al., 2022a; Kaplan et al., 2020).Zero-Shot CoT (Kojima et al., 2022) enables step-by-step reasoning based on instructions alone, without examples. Complex CoT (Fu et al., 2022) further optimizes reasoning by selecting prompts with more intermediate steps, improving performance on multi-step tasks. Other approaches, such as self-consistency (Wang et al., 2022; Li et al., 2023), progressive sampling (Zheng et al., 2023a), and meta-prompting (Suzgun and Kalai, 2024), focus on enhancing reliability and consistency. Maieutic Prompting enforces logical coherence by asking the model to verify its internal consistency, producing more dependable, self-consistent responses (Jung et al., 2022; Michael et al., 2023).

Despite these advances, challenges remain in applying prompting strategies to MCQ tasks. Traditional approaches typically adopt an "additive" strategy, increasing the amount of input information provided to the model. However, these methods do not effectively address the overattention to distractors, often leading models to select plausible but ultimately incorrect answers when faced with complex distractor options.

3 Preliminary

We begin by defining the fundamental concepts of standard prompting and CoT prompting, which serve as the foundation for the proposed EoT framework. Consider a scenario where a question is represented as q, a prompt as $\mathcal{T}$, and a LLM, denoted as $P_{\mathcal{M}}$.

Standard Prompting Under standard prompting, the LLM takes the question q, the set of candidate options $\mathcal{X}$, and the prompt $\mathcal{T}$ as input. The objective of the model is to identify the optimal answer o from a set of multiple possible options.In this approach, the model directly generates the final answer o based on the q, $\mathcal{X}$ and the $\mathcal{T}$, with the probability distribution defined as:

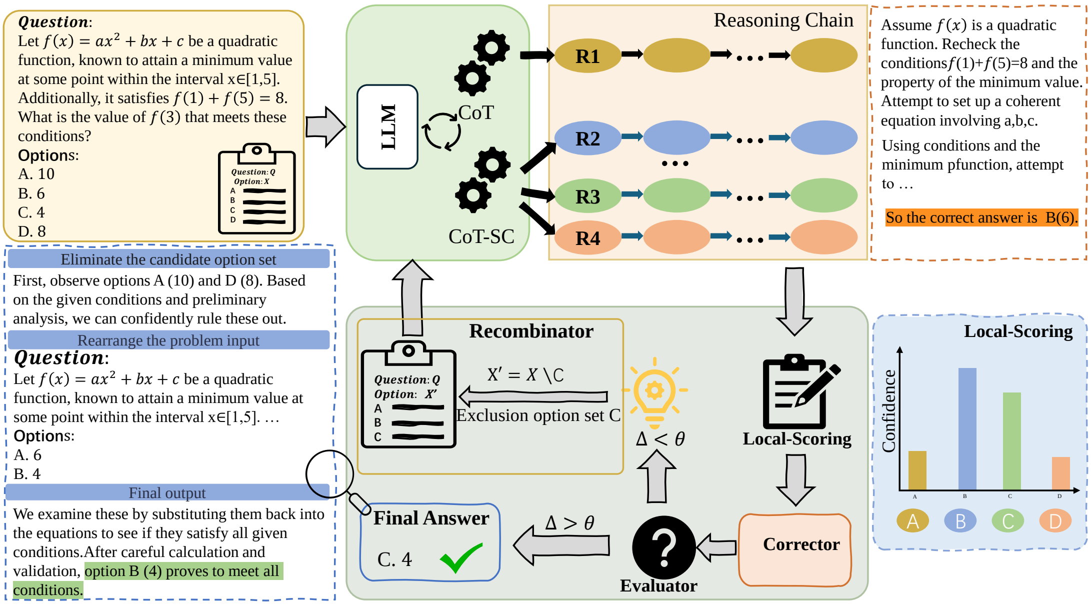
$$P(o \mid \mathcal{T}, q, \mathcal{X}) = \frac{\exp(P_{\mathcal{M}}(o \mid \mathcal{T}, q, \mathcal{X}))}{\sum_{d_k \in \mathcal{X}} \exp(P_{\mathcal{M}}(d_k \mid \mathcal{T}, q, \mathcal{X}))} \quad (1)$$

where $\mathcal{X} = \{d_1, d_2, \ldots, d_n\}$ represents the set of all possible candidate options.

Chain-of-Thought Prompting CoT prompting enhances the standard prompting paradigm by explicitly guiding the LLMs to generate intermediate reasoning steps prior to producing the final answer. Specifically, the prompt $\mathcal{T}$ is designed to encourage the LLMs to first generate a reasoning process r, followed by the final answer o. The combination of r and o is collectively referred to as a reasoning chain. The joint probability of generating the reasoning chain (r, o) conditioned on $\mathcal{T}$ and q is expressed as:

$$\begin{aligned} P(r, o \mid \mathcal{T}, q, \mathcal{X}) = & P(o \mid \mathcal{T}, q, \mathcal{X}, r) \cdot \\ & P(r \mid \mathcal{T}, q, \mathcal{X}) \end{aligned} \quad (2)$$

where $P(r \mid \mathcal{T}, q, \mathcal{X})$ and $P(o \mid \mathcal{T}, q, \mathcal{X}, r)$ are



Question:
Let $f(x) = ax^2 + bx + c$ be a quadratic function, known to attain a minimum value at some point within the interval x∈[1,5]. Additionally, it satisfies $f(1) + f(5) = 8$. What is the value of $f(3)$ that meets these conditions?
Options:
A. 10
B. 6
C. 4
D. 8

Eliminate the candidate option set

First, observe options A (10) and D (8). Based on the given conditions and preliminary analysis, we can confidently rule these out.

Rearrange the problem input

Question:
Let $f(x) = ax^2 + bx + c$ be a quadratic function, known to attain a minimum value at some point within the interval x∈[1,5]. …
Options:
A. 6
B. 4

Final output

We examine these by substituting them back into the equations to see if they satisfy all given conditions.After careful calculation and validation, option B (4) proves to meet all conditions.

Assume $f(x)$ is a quadratic function. Recheck the conditions$f(1)$+$f(5)$=8 and the property of the minimum value. Attempt to set up a coherent equation involving a,b,c.

Using conditions and the minimum pfunction, attempt to …

So the correct answer is B(6).

Figure 2: Illustration of the EoT framework's reasoning process. Initially, EoT utilizes various reasoning chains to generate the raw outputs, which are processed through the Local-Scoring module to obtain the initial probability distribution. The Corrector then adjusts this distribution to produce an exclusion-based distribution. The Evaluator assesses whether the LLM has sufficient confidence to select the correct answer. If not, the Recombinator excludes the lowest-scoring option and reconstructs the input, ensuring that the LLMs do not refocus on the eliminated distractor options.

defined as follows:

$$P(r \mid \mathcal{T}, q, \mathcal{X}) = \prod_{i=1}^{|\mathcal{R}|} P_{\mathcal{M}}(r_i \mid \mathcal{T}, q, \mathcal{X}, r_{<i}) \tag{3}$$

$$P(o \mid \mathcal{T}, q, \mathcal{X}, r) \;=\; P_{\mathcal{M}}(o \mid \mathcal{T}, q, \mathcal{X}, r) \tag{4}$$

where $\mathcal{R} = \{r_1, r_2, r_3 \dots\}$ denotes the set of reasoning steps, and $r_{<i}$ indicates that the generation of r_i depends on the preceding $i - 1$ chains.

4 Methodology

4.1 Overview

In light of the existing challenges associated with MCQs and the limitations of current methodologies, we propose the EoT framework. This approach is specifically designed to enhance the performance of LLMs in MCQ tasks by systematically eliminating incorrect options. The core idea of EoT is to eliminate distractors from the set of possible answers, thereby increasing the likelihood of identifying the correct answer. Let C denote the set of excluded incorrect options, and let the refined options set be $\mathcal{X}' = \mathcal{X} \setminus C$. Given the question q and options $\mathcal{X}'$, the adjusted probability distribution P_{EoT} is defined as:

$$\begin{aligned} P_{EoT}(o \mid \mathcal{T}, q, \mathcal{X}', C) = P_{\text{exclude}}(C \mid \mathcal{T}, q, \mathcal{X}) \cdot \\ \frac{P_{\mathcal{M}}(o \mid \mathcal{T}, q, \mathcal{X}')}{\sum_{d_k \in \mathcal{X}'} P_{\mathcal{M}}(d_k \mid \mathcal{T}, q, \mathcal{X}')} \end{aligned} \tag{5}$$

where $C = \{c_1, c_2, \dots, c_j\}, j < |\mathcal{X}|$. $P_{\text{exclude}}(c_j \mid \mathcal{T}, q, X)$ represents the confidence of LLMs in excluding the option c_j.

The exclusion process adjusts the probability distribution by removing the influence of the incorrect options C, thereby increasing the relative probability of selecting the correct answer o to some extent.

4.2 System Architecture

The EoT framework is designed to enhance the decision-making accuracy of LLMs by systematically eliminating distractor options, thereby guiding the models to focus their attention on the most plausible choices. This process unfolds through two complementary phases: **Exclusionary Confidence Calibration** (§ 4.2.1) and **Dynamic Confidence Gap Decision** (§ 4.2.2). Together, these phases emulate human-like exclusion strategies, alleviating the cognitive burden on LLMs and improving their performance on MCQ tasks.

As illustrated in Figure 2, EoT functions as a plug-and-play module that can be easily integrated into existing prompting strategies. It identifies and removes the most likely incorrect options to refine the answer set, directing the model's attention toward the correct choices. Initially, EoT generates raw sampling results using different reasoning strategies, which are processed by the Local-Scoring module. This module computes the initial distribution of the LLM's confidence across the options. The Corrector module then adjusts the probability distribution by excluding certain options. EoT evaluates whether the LLM has sufficient confidence in its answers, based on a precomputed threshold. If confidence is insufficient, the Recombinator module refines the input, preventing the LLM from refocusing on the eliminated distractors.

4.2.1 Exclusionary Confidence Calibration

This phase systematically identifies and removes the answer options that are most likely to be incorrect, subsequently adjusting the probability distribution after exclusion.

Initial Probability Computation. Given a prompt $\mathcal{T}$, a question q, and a set of options $\mathcal{X} = \{d_1, d_2, \ldots, d_n\}$, EoT computes the initial probability distribution as follows:

$$P_{observed}(d_i \mid \mathcal{T}, q, \mathcal{X}) = \frac{exp(f(\mathcal{T}, q, d_i))}{\sum\limits_{d_k \in \mathcal{X}} exp(f(\mathcal{T}, q, d_k))} \quad (6)$$

where $f(\mathcal{X}, q, d_i)$ represents the score of token d_i in the raw output of the LLMs.

Confidence for Excluding Options. The confidence $P_{\text{exclude}}(c_j \mid \mathcal{T}, q, \mathcal{X})$ for excluding an option c_j is defined as the LLMs' certainty in rejecting c_j as a distractor, expressed as:

$$P_{\text{exclude}}(c_j \mid \mathcal{T}, q, \mathcal{X}) = 1 - P_{\text{observed}}(d_j \mid \mathcal{T}, q, \mathcal{X}) \quad (7)$$

Refined Probability Distribution. After excluding a set of options $C = \{c_1, c_2, \ldots, c_j\}$, the refined set of candidate options is $\mathcal{X}' = \mathcal{X} \setminus C$. Let $\widetilde{P}_{\text{EoT}}$ denote the EoT-adjusted relative scoring after exclusion. We define the adjusted probability for each remaining choice d_j as:

$$\widetilde{P}_{EoT}(d_i \mid \mathcal{T}, q, \mathcal{X}') = P_{\text{observed}}(d_i \mid \mathcal{T}, q, \mathcal{X}) \cdot P_{\text{exclude}}(C \mid \mathcal{T}, q, \mathcal{X}) \quad (8)$$

where $P_{\text{exclude}}(C \mid \mathcal{T}, q, \mathcal{X})$ is the product of the confidence scores for all excluded options:

$$P_{\text{exclude}}(C \mid \mathcal{T}, q, \mathcal{X}) = \prod_{c_j \in C} P_{\text{exclude}}(c_j \mid \mathcal{T}, q, \mathcal{X}) \quad (9)$$

4.2.2 Iterative Confidence Refinement

This phase evaluates whether the LLMs have sufficient confidence to provide a final answer based on the refined probability distribution $P^{\sim}_{\mathcal{M}}$ obtained in the previous phase. It also determines whether re-evaluation of $\mathcal{X}'$ is necessary to ensure the robustness of EoT.

Confidence Gap Calculation. In each iteration, EoT assesses the confidence difference Δ between the highest-probability option o_α and the second-highest-probability option o_β:

$$\Delta = \widetilde{P}_{EoT}(o_\alpha \mid \mathcal{T}, q, \mathcal{X}') - \widetilde{P}_{EoT}(o_\beta \mid \mathcal{T}, q, \mathcal{X}') \quad (10)$$

If Δ is below a predefined threshold θ, the model samples additional reasoning chains to gather more information about the remaining options. These reasoning chains are integrated into the decision-making process, updating $P_{observed}(d_i \mid \mathcal{T}, q, \mathcal{X}')$. The process terminates when $\Delta > \theta$ or when the maximum number of iterations N_{max} is reached, with the final answer being $o = o_\alpha$.

To adapt as the number of options decreases, we update the threshold using the following formula:

$$\theta_{new} = \theta + (\theta_{max} - \theta) \cdot (1 - \frac{N_{current}}{N_{initial}})^p \quad (11)$$

where $N_{current}$ denotes the number of remaining options in the current state, $N_{initial}$ represents the initial number of options, θ_{max} is the maximum threshold value , and p is a tuning factor.

Algorithm 1 Reasoning with EoT

```
Require: Question $q$, option set $\mathcal{X}$, model $P_{\mathcal{M}}$,
    threshold $\theta$, maximum iterations $N_{\text{max}}$
Ensure: $o$
$N \leftarrow 0$
Compute $P_{\text{obs}}(d_i \mid q, \mathcal{X})$, for all $d_i \in \mathcal{X}$
while $N < N_{\text{max}}$ do
    Compute $P_{\text{exc}}(c_i \mid \mathcal{T}, q, \mathcal{X})$, for all $c_i \in \mathcal{X}$
    $\mathcal{X}' \leftarrow \mathcal{X} \setminus \{c_j\}$ {Remove $c_j$ based on a
        predefined criterion}
    Compute $\widetilde{P}(d_i \mid \mathcal{T}, q, \mathcal{X}')$, for all $d_i \in \mathcal{X}'$
    $\Delta \leftarrow \widetilde{P}(o_\alpha) - \widetilde{P}(o_\beta)$
    if $\Delta \geq \theta$ then
        return $o \leftarrow o_\alpha$
    else
        $\mathcal{X} \leftarrow \mathcal{X}'$, Update $\theta$
    end if
    $N \leftarrow N + 1$
end while
return $o$
```

5 Experimental Setup

5.1 Datasets

To comprehensively evaluate the effectiveness of our proposed EoT framework, we conducted experiments on MCQ benchmark datasets. These datasets span diverse domains and difficulty levels, enabling us to validate the model's performance across a variety of tasks. **MMLU-Pro** (Massive Multi-Task Language Understanding Professional). An enhanced version of MMLU, MMLU-Pro focuses on professional domains such as law, medicine, engineering, and finance. The questions are more difficult and are presented in a 10-option MCQs format. **CommonsenseQA** (CSQA). A dataset emphasizing commonsense reasoning, with questions in a 5-option MCQs format. **ARC** (AI2 Reasoning Challenge). Derived from U.S. elementary and middle school science exams, ARC includes 7,787 science reasoning questions and is primarily used to assess the model's performance in scientific knowledge and complex reasoning tasks. **AQuA** (Algebra Question Answering with Rationales). This dataset contains nearly 100,000 algebra questions, each accompanied by detailed solution steps. It aims to evaluate the model's problem-solving and logical reasoning abilities in mathematical tasks. **GSM8K-MC**. This dataset is a MCQs adaptation of the original GSM8K dataset, containing 8,000 elementary-level mathematics problems designed to assess the model's mathematical reasoning capabilities.

5.2 Backbone LLMs

In the main experiments, we selected a series of representative LLMs. The models used in our experiments include LLaMA-2-7B-Chat-HF, LLaMA-2-13B-Chat-HF, LLaMA-3-8B-Instruct, and Phi-3.5-mini, Yi-1.5-9B-Chat, as well as the recently popular o1 model, specifically the Marco-o1 model. These models are of medium scale, covering a range of sizes and training paradigms, which aligns with the requirements of our study.

5.3 Baseline

The performance gains of the EoT framework, as a plug-and-play approach, we compared it against the following baseline methods. We adhere to the original experimental settings and utilize the official implementations of these methods to ensure fairness and comparability:

- Chain-of-Thought (CoT): Generates intermediate reasoning steps to help the model derive answers step by step.
- Zero-Shot CoT: Triggers the model's reasoning process through simple instructions without using examples.
- ComplexCoT: Selects the most effective reasoning examples based on complexity to improve multi-step reasoning tasks.

5.4 Hyperparameter Settings

The threshold θ plays a critical role in the EoT framework. To determine the optimal threshold, we used 5% of the total samples in each dataset as a validation subset. For each option in this 5% subset, we recorded the model's confidence scores. Using these confidence scores as priors, we performed Maximum Likelihood Estimation (MLE) to identify the threshold θ that maximizes the performance of the exclusion method on the validation subset. During the exclusion process, θ is dynamically adjusted to ensure that the model maintains the same level of confidence after excluding options as it did in the initial state. In our experiments, we treat θ_{max} and p as hyperparameters, with θ_{max} set to 0.85 and p set to 2.

5.5 Evaluation Metrics

We evaluate model performance using the following metrics:

Accuracy: This is the percentage of correctly answered questions out of the total number of questions. It serves as the primary performance indicator and is computed as:

$$\mathcal{A}cc = \frac{\sum_{i=1}^{N} \mathbb{I}(o_i = o_i^*)}{N} \tag{12}$$

where $\mathbb{I}(\cdot)$ is the indicator function that returns 1 if the ground truth answer o_i equals the EoT's answer o_i^* and 0 otherwise, and N is the total number of questions.

Exclusion Accuracy: Denoted as $\mathcal{E}_i$, this metric represents the accuracy of the i-th exclusion operation. It is defined as:

$$\mathcal{E}_i = \frac{\sum_{j=1}^{M} \mathbb{I}(o_j \notin C)}{M} \tag{13}$$

where M is the number of MCQs requiring the i-th elimination operation, and C is the set of excluded options.

Model	MMLU-PRO	ARC	CSQA	GSM8K-MC	AQuA
Llama-2-7b-chat-hf	18.04	55.02	54.77	24.66	18.90
Llama-2-7b-chat-hf-EoT	**19.72**	**55.45**	**56.00**	**25.14**	**21.25**
Llama-2-13b-chat-hf	19.76	60.77	63.07	27.39	19.29
Llama-2-13b-chat-hf-EoT	**22.14**	**62.23**	**67.27**	**27.55**	**22.44**
Llama-3-8b-instruct	37.43	82.40	76.97	35.61	29.92
Llama-3-8b-EoT	**39.02**	**83.09**	**78.29**	**38.20**	**33.07**
Phi-3.5-mini	40.91	86.26	72.94	37.98	25.20
Phi-3.5-mini-EoT	**41.60**	**86.43**	**73.52**	**41.17**	**26.77**
Yi-1.5-9B-Chat	40.84	88.44	81.74	46.73	29.92
Yi-1.5-9B-Chat-EoT	**42.36**	**89.53**	**82.82**	**47.96**	**32.71**
Marco-o1	43.25	89.61	81.65	46.27	27.95
Marco-o1-EoT	**43.96**	**90.21**	**82.51**	**47.19**	**31.10**

Table 2: Performance of various models with and without EoT under the 5-shot setting (Acc).

Model	Prompt	EoT	Dataset				
			ARC	MMLU-PRO	CSQA	GSM8K-MC	AQuA
Llama-3-8B-Instruct	Zero-shot CoT	✗	80.55	36.10	75.12	34.50	27.50
		✓	82.00	37.80	76.40	36.75	30.45
			(+1.45)	(+1.70)	(+1.28)	(+2.25)	(+2.95)
	Complex CoT	✗	81.20	36.85	76.30	35.00	28.60
		✓	82.80	38.70	77.85	37.85	31.75
			(+1.60)	(+1.85)	(+1.55)	(+2.85)	(+3.15)
	CoT	✗	83.00	38.00	78.10	37.10	30.10
		✓	84.20	40.30	79.50	39.90	34.20
			(+1.20)	(+2.30)	(+1.40)	(+2.80)	(+4.10)
Phi-3.5-mini-instruct	Zero-shot CoT	✗	85.00	39.50	71.50	36.50	23.80
		✓	85.70	40.70	72.40	39.50	25.50
			(+0.70)	(+1.20)	(+0.90)	(+3.00)	(+1.70)
	Complex CoT	✗	85.50	40.00	72.20	37.80	24.80
		✓	86.00	41.50	73.10	40.30	26.10
			(+0.50)	(+1.50)	(+0.90)	(+2.50)	(+1.30)
	CoT	✗	86.10	40.30	73.00	39.00	25.50
		✓	87.00	42.20	74.50	42.00	28.00
			(+0.90)	(+1.90)	(+1.50)	(+3.00)	(+2.50)
Marco-o1	Zero-shot CoT	✗	88.00	42.00	80.00	45.00	26.50
		✓	89.20	43.10	82.20	46.30	28.90
			(+1.20)	(+1.10)	(+2.20)	(+1.30)	(+2.40)
	Complex CoT	✗	89.00	42.80	81.80	46.00	27.60
		✓	89.80	43.60	82.94	47.50	30.80
			(+0.80)	(+0.80)	(+1.14)	(+1.50)	(+3.20)
	CoT	✗	89.40	43.00	82.13	46.80	28.50
		✓	90.40	44.50	83.01	48.50	32.20
			(+1.00)	(+1.50)	(+0.88)	(+1.70)	(+3.70)

Table 3: Performance improvements obtained by applying EoT to different LLMs and prompting methods. The results indicate that EoT yields additional gains, particularly on more challenging tasks (Acc).

6 Experimental Result

We comprehensively evaluated the proposed EoT method across various LLMs and compared its performance with mainstream baseline approaches. Since GPT-3.5-turbo does not provide direct access to output probabilities, we employed a multi-turn dialogue approach (details in Appendix D.3).

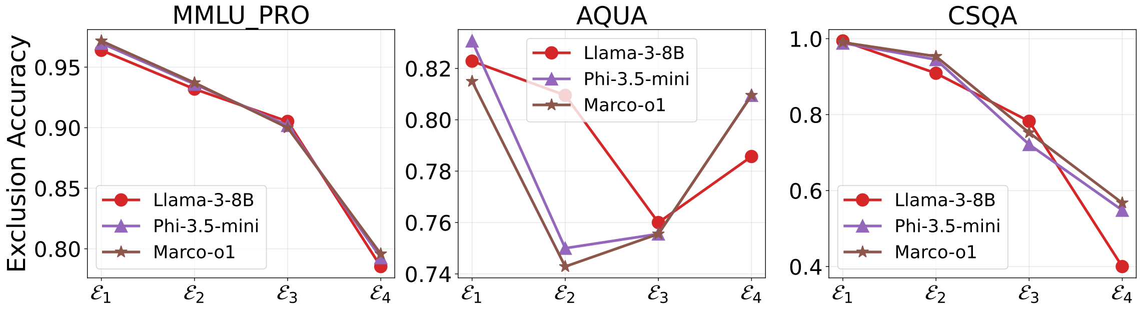


Figure 3: Under the optimal threshold θ with standard prompt settings, the Exclusion Accuracy of EoT across different numbers of exclusions.

Table 2 presents the performance comparison of models under standard prompts. Overall, incorporating EoT consistently improves accuracy, particularly on challenging MCQ tasks. For example, on the AQuA dataset, EoT achieved a 6.84 percentage point improvement. Similarly, on the GSM8K-MC and MMLU-Pro datasets, the gains were 2.85 and 2.3 percentage points, respectively. These results suggest that EoT effectively mitigates the “cognitive load” imposed by distractor options, thereby supporting our hypothesis that LLMs tend to overemphasize distractors, leading to suboptimal performance on difficult samples.

Table 3 highlights the performance of EoT when combined with mainstream prompting strategies as a plug-and-play module. The results indicate that EoT still yields significant performance improvements when integrated with these strategies. Specifically, LLMs exhibit better performance with the addition of prompting strategies, and this improvement is further amplified when EoT is applied. Although the magnitude of improvement is smaller compared to scenarios without any prompting, EoT still performs better on more difficult tasks.

7 Analysis

In this section, we analyze the behavior of the EoT framework under different experimental settings, focusing on two key aspects: the exclusion accuracy over successive rounds of elimination and the sensitivity of EoT's performance to the threshold θ.

7.1 Exclusion Accuracy Analysis

Figure 3 illustrates how the exclusion accuracy—computed only on those samples where the model's confidence in the answer falls below a predefined threshold—changes with the number of exclusion rounds across various datasets. The results reveal that, during MCQ reasoning tasks, the probability of correctly excluding a distractor significantly exceeds the probability of directly selecting the correct answer. This observation validates the core premise of EoT: systematically eliminating distractors can enhance overall performance.

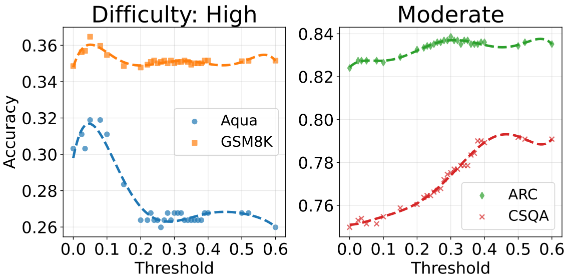


Figure 4: The impact of the threshold θ on the performance of EoT across tasks of varying difficulty levels (5-shot Llama-3-8B-Instruct).

However, for simpler tasks such as those in CSQA, exclusion accuracy declines sharply after more than two exclusion rounds—sometimes falling below the baseline performance. This suggests that for tasks where LLMs already exhibit strong commonsense reasoning, excessive exclusion may introduce hallucination issues. In contrast, for more challenging tasks like AQuA, while exclusion accuracy initially decreases, it later improves as more distractors are removed, indicating that further elimination helps the model to better differentiate between correct and incorrect options.

7.2 Threshold Sensitivity Analysis

Figure 4 examines the impact of the threshold θ on the performance of EoT, using the 5-shot Llama-3-8B-Instruct setting as an example. For high-difficulty tasks, optimal performance is achieved with relatively low thresholds. This is likely because challenging MCQs feature complex distractors that narrow the confidence gap between distractors and the correct option, necessitating a smaller threshold to iteratively eliminate distractors. Conversely, for standard tasks where LLMs can more readily distinguish between correct and distractor options, a larger threshold is more effective.

These findings support our strategy of using 5% of each dataset to determine the optimal threshold θ, ensuring that the exclusion process consistently maintains a balanced confidence level.

8 Conclusion

In this work, we introduced the EoT framework, a novel prompting strategy aimed at enhancing the performance of LLMs on MCQ tasks by systematically eliminating incorrect options. Inspired by human reasoning strategies, EoT reduces cognitive load by redirecting the model's attention away from distractors, allowing it to focus more effectively on analyzing the relevant remaining choices. Our extensive experiments across a variety of MCQ datasets demonstrate that EoT significantly improves reasoning accuracy, particularly for challenging tasks. We hope our framework will inspire future research into incorporating human-inspired reasoning strategies in LLMs.

Limitations

Our research on the EoT framework has demonstrated significant efficacy in reducing the cognitive load of large language models (LLMs) and enhancing performance on MCQ tasks. However, there are numerous promising avenues for further exploration to expand upon these findings. For instance, a particularly promising direction involves extending the application of EoT from multiple-choice questions to other reasoning tasks and domains. The core principles of systematic exclusion hold potential for adaptation in areas such as knowledge graphs and retrieval-augmented generation, thereby better simulating human-like reasoning across diverse scenarios.

Acknowledgments

This research was supported by the National Key Research and Development Program of China (No.2023YFC3304500), the Guizhou Provincial Major Scientific and Technological Program (No.[2024]003), and the Guizhou Provincial Key Technology R&D Program (No.[2023]448).

Ethics Statement

This research presents no ethical concerns. All experiments were conducted using publicly available datasets exclusively for research purposes. We carefully reviewed the selected datasets to ensure they do not contain unethical content, private information, or sensitive topics. The foundation models used in this study are openly accessible and have been employed in accordance with their research-oriented licenses. Moreover, AI assistants were utilized in the writing and refinement of this paper.

References

Nishant Balepur, Shramay Palta, and Rachel Rudinger. 2024. It‘s not easy being wrong: Large language models struggle with process of elimination reasoning. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 10143–10166, Bangkok, Thailand. Association for Computational Linguistics.

Zheng Chu, Jingchang Chen, Qianglong Chen, Weijiang Yu, Tao He, Haotian Wang, Weihua Peng, Ming Liu, Bing Qin, and Ting Liu. 2024. Navigate through enigmatic labyrinth a survey of chain of thought reasoning: Advances, frontiers and future. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1173–1203.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*.

Yao Fu, Hao Peng, Ashish Sabharwal, Peter Clark, and Tushar Khot. 2022. Complexity-based prompting for multi-step reasoning. In *The Eleventh International Conference on Learning Representations*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

Jie Huang and Kevin Chen-Chuan Chang. 2022. Towards reasoning in large language models: A survey. *arXiv preprint arXiv:2212.10403*.

Jaehun Jung, Lianhui Qin, Sean Welleck, Faeze Brahman, Chandra Bhagavatula, Ronan Le Bras, and Yejin Choi. 2022. Maieutic prompting: Logically consistent reasoning with recursive explanations. *arXiv preprint arXiv:2205.11822*.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.

Yifei Li, Zeqi Lin, Shizhuo Zhang, Qiang Fu, Bei Chen, Jian-Guang Lou, and Weizhu Chen. 2023. Making language models better reasoners with step-aware verifier. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5315–5333.

Zhiyuan Li, Hong Liu, Denny Zhou, and Tengyu Ma. 2024. Chain of thought empowers transformers to solve inherently serial problems. *arXiv preprint arXiv:2402.12875*.

Wang Ling, Dani Yogatama, Chris Dyer, and Phil Blunsom. 2017. Program induction by rationale generation: Learning to solve and explain algebraic word problems. *arXiv preprint arXiv:1705.04146*.

Chenkai Ma and Xinya Du. 2023. POE: Process of elimination for multiple choice reasoning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4487–4496, Singapore. Association for Computational Linguistics.

Julian Michael, Salsabila Mahdi, David Rein, Jackson Petty, Julien Dirani, Vishakh Padmakumar, and Samuel R Bowman. 2023. Debate helps supervise unreliable experts. *arXiv preprint arXiv:2311.08702*.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. Rethinking the role of demonstrations: What makes in-context learning work?

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Shuofei Qiao, Yixin Ou, Ningyu Zhang, Xiang Chen, Yunzhi Yao, Shumin Deng, Chuanqi Tan, Fei Huang, and Huajun Chen. 2022. Reasoning with language model prompting: A survey. *arXiv preprint arXiv:2212.09597*.

Daking Rai and Ziyu Yao. 2024. An investigation of neuron activation as a unified lens to explain chain-of-thought eliciting arithmetic reasoning of llms. *arXiv preprint arXiv:2406.12288*.

Mirac Suzgun and Adam Tauman Kalai. 2024. Meta-prompting: Enhancing language models with task-agnostic scaffolding. *arXiv preprint arXiv:2401.12954*.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2018. Commonsenseqa: A question answering challenge targeting commonsense knowledge. *arXiv preprint arXiv:1811.00937*.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B Hashimoto. 2023. Stanford alpaca: An instruction-following llama model.

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher Manning. 2023. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5433–5442, Singapore. Association for Computational Linguistics.

Yongqi Tong, Yifan Wang, Dawei Li, Sizhe Wang, Zi Lin, Simeng Han, and Jingbo Shang. 2023. Eliminating reasoning via inferring with planning: A new framework to guide llms' non-linear thinking. *arXiv preprint arXiv:2310.12342*.

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

Peifeng Wang, Zhengyang Wang, Zheng Li, Yifan Gao, Bing Yin, and Xiang Ren. 2023. Scott: Self-consistent chain-of-thought distillation. *arXiv preprint arXiv:2305.01879*.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, et al. 2024. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. *arXiv preprint arXiv:2406.01574*.

Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. 2021. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. 2022a. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022b. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Chenrui Zhang, Lin Liu, Chuyuan Wang, Xiao Sun, Hongyu Wang, Jinpeng Wang, and Mingchen Cai.

2024a. Prefer: Prompt ensemble learning via feedback-reflect-refine. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 19525–19532.

Ziyin Zhang, Lizhen Xu, Zhaokun Jiang, Hongkun Hao, and Rui Wang. 2024b. Multiple-choice questions are efficient and robust llm evaluators. *arXiv preprint arXiv:2405.11966*.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*.

Yu Zhao, Huifeng Yin, Bo Zeng, Hao Wang, Tianqi Shi, Chenyang Lyu, Longyue Wang, Weihua Luo, and Kaifu Zhang. 2024. Marco-o1: Towards open reasoning models for open-ended solutions. *arXiv preprint arXiv:2411.14405*.

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

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023b. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623.

Wanjun Zhong, Ruixiang Cui, Yiduo Guo, Yaobo Liang, Shuai Lu, Yanlin Wang, Amin Saied, Weizhu Chen, and Nan Duan. 2023. Agieval: A human-centric benchmark for evaluating foundation models. *arXiv preprint arXiv:2304.06364*.

Ce Zhou, Qian Li, Chen Li, Jun Yu, Yixin Liu, Guangjing Wang, Kai Zhang, Cheng Ji, Qiben Yan, Lifang He, et al. 2023. A comprehensive survey on pretrained foundation models: A history from bert to chatgpt. *arXiv preprint arXiv:2302.09419*.

A Evaluation Data and Statistics

The datasets used in this study can be accessed via the links below:

- **MMLU_PRO**: https://github.com/TIGER-AI-Lab/MMLU-Pro
- **ARC**: https://allenai.org/data/arc
- **CSQA**: https://allenai.org/data/commonsenseqa
- **GSM8K-MC**: https://huggingface.co/datasets/guipenedo/gsm8k-mc
- **AQuA**: https://github.com/google-deepmind/AQuA

B Evaluated Open-Source Models

We obtained the open-source LLMs used in our experiments through the following means:

- **Llama-2-7b-chat-hf** https://huggingface.co/meta-llama/Llama-2-7b-chat-hf
- **Llama-2-13b-chat-hf** https://huggingface.co/meta-llama/Llama-2-13b-chat-hf
- **Llama-3-8b-instruct** https://huggingface.co/meta-llama/Meta-Llama-3-8B-Instruct
- **Phi-3.5-mini-instruct** https://huggingface.co/microsoft/Phi-3.5-mini-instruct
- **Yi-1.5-9B-Chat** https://huggingface.co/01-ai/Yi-1.5-9B-Chat
- **Marco-o1** https://huggingface.co/AIDC-AI/Marco-o1

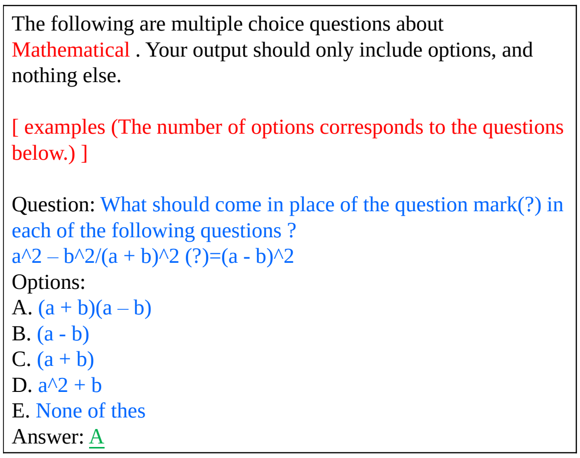
C Experimental Setup Details

We conducted our experiments using two NVIDIA A6000 GPUs (48GB each) to facilitate the inference of LLMs in the 7B-14B parameter range.

Name	Domain	Format	Quantity	Year	Difficulty
MMLU-Pro	Law, Medicine, Engineering, Finance	10-choice	12,000	2024	High
CommonsenseQA (CSQA)	Commonsense Reasoning	5-choice	12,102	2019	Moderate
ARC	Science Reasoning	4-choice	7,787	2018	Moderate
AQuA	Mathematical Problem-Solving	5-choice	100,000	2017	High
GSM8K-MC	Mathematical Reasoning	4-choice	8,500	2021	High

Table 4: Benchmark Datasets Overview.

C.1 Prompts Used in Experiments



The following are multiple choice questions about Mathematical . Your output should only include options, and nothing else.

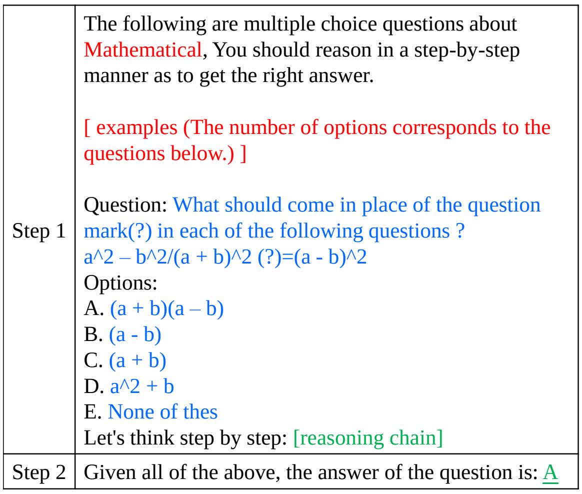
[examples (The number of options corresponds to the questions below.)]

Question: What should come in place of the question mark(?) in each of the following questions ?
a^2 – b^2/(a + b)^2 (?)=(a - b)^2
Options:
A. (a + b)(a – b)
B. (a - b)
C. (a + b)
D. a^2 + b
E. None of thes
Answer: A

Figure 5: The input format for open-source models (e.g., LLaMA and Phi) is structured as follows: Black text represents the template input. Red text denotes the task-specific template (if no task-specific information is available, the dataset's default template is used). Green text indicates the position where the next-token prediction probabilities for option IDs are utilized as the observed prediction distribution. It is important to note that the input is preprocessed using the dialogue template recommended by the respective LLMs before being fed into the model.

C.2 Selecting Examples for 5-Shot Prompting

The examples used in the prompt were randomly drawn from the dataset's non-numerical reasoning questions. We chose to exclude numerical reasoning items because, in those cases, the EoT exclusion strategy fails to yield any additional inferential benefit: the model typically retains similar levels of uncertainty across all answer choices. In contrast, non-numerical reasoning questions align more naturally with the elimination logic of EoT, allowing the model to leverage this strategy more effectively.



Step 1	The following are multiple choice questions about Mathematical, You should reason in a step-by-step manner as to get the right answer. [examples (The number of options corresponds to the questions below.)] Question: What should come in place of the question mark(?) in each of the following questions ? a^2 – b^2/(a + b)^2 (?)=(a - b)^2 Options: A. (a + b)(a – b) B. (a - b) C. (a + b) D. a^2 + b E. None of thes Let's think step by step: [reasoning chain]
Step 2	Given all of the above, the answer of the question is: A

Figure 6: Input format for Chain-of-Thought (CoT) prompting, divided into two stages. In the first stage, the LLM generates a reasoning chain (with the temperature set to 0, or 0.8 for CoT-SC). In the second stage, the generated reasoning chains are aggregated to produce the final prediction distribution.

D More Experimental Results

D.1 Examples of EoT in Action

Input	**Question:** Two balls A and B rotate along a circular track. Ball A makes 2 full rotations in 26 minutes. Ball B makes 5 full rotation in 35 minutes. If they start rotating now from the same point, when will they be at the same starting point again? **Options:** A: 1 hour and 31 minute B: 2 hour and 31 minute C: 3 hour and 31 minute D: 4 hour and 31 minute E: 5 hour and 31 minute
Original	B: 0.278, A: 0.245, D: 0.191, C: 0.168, E: 0.116
EoT-1	A: 0.378, B: 0.334, C: 0.178, D: 0.108
EoT-2	A: 0.486, B: 0.429, C: 0.084
EoT-3	A: 0.798, B: 0. 201

Figure 7: An example from the AQuA dataset. The most probable option from the original prediction distribution is incorrect. Here, $EoT-i$ denotes the application of the exclusion method i times. After eliminating distractor options, the LLM is better able to distinguish the correct answer from the remaining distractors.

D.2 Discussion of the Additional Overhead Introduced by EoT

EoT operates iteratively but incurs only modest overhead, since it computes choice probabilities solely from the logits of the last token in the LLM's original output, without requiring additional text generation. We evaluated this on LLaMA3-8B and measured inference time before and after integrating EoT. The results as 5 are as follows. These findings indicate that the extra computational cost introduced by EoT across various tasks is limited and acceptable given the performance gains it yields.

Task	Standard	EoT
ARC	1.41	1.51
CSQA	1.02	1.19
MMLU-PRO	1.96	2.64
AQuA	0.95	1.54
GSM8K-MC	1.51	2.77

Table 5: Average inference time per MCQ (in seconds) with and without EoT (5-shot, LLaMA3-8B).

D.3 Implementing the EoT process with Non-Open-Source Models

For non-open-source LLMs (e.g., GPT), we implement the concept of EoT using Chain-of-Thought prompting. In this approach, the LLM is guided to generate an exclusion process that mimics exclusion-based reasoning. However, this is not an optimal solution. Our experiments reveal that the generated exclusion process may still exhibit bias, as the model continuously perceives the presence of distractors. This can influence the reasoning trajectory and introduce unintended biases into the exclusion process.

E The Potential of EoT in Other Tasks

Although EoT was originally developed for multiple-choice questions, its core principle, progressivelyy eliminating distractors and focusing on the relatively correctcandidates, canan be fruitfully applied to a much wider range of tasks. Inspired by the way the GSM8K-MC dataset was constructed, we can adapt non-MCQ tasks to an EoT-style pipeline in three stages:

Candidate Sampling We first prompt the LLM to sample repeatedly on the same input, generating a diverse set of answer candidates along with their chain-of-thoughts. These outputs then serve as the input pool for EoT.

Stage-wise Exclusion We apply EoT to this pool in successive rounds: in each round, we discard answers that are contradictory, logically flawed, or obviously incorrect, retaining only those that remain comparatively plausible at that stage.

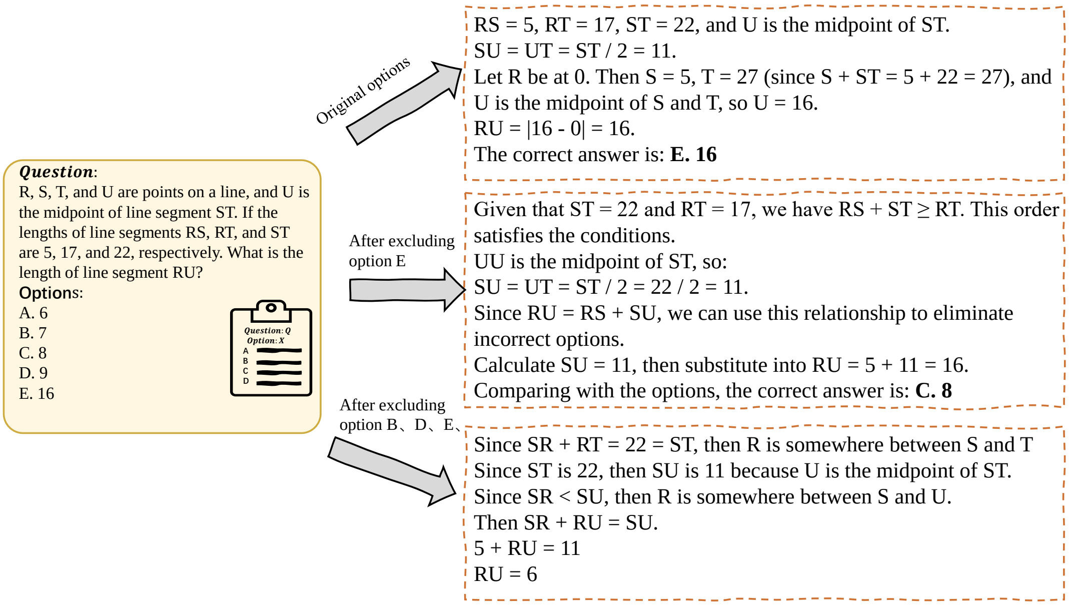
Answer Refinement Finally, we use the surviving candidates as a foundation to guide the model toward producing a single, polished final answer.

By reframing open-ended QA, mathematical reasoning, code generation, and other tasks as a "sample–exclude–iterate" process, we leverage EoT's strength in step-by-step pruning to substantially reduce the model's cognitive load and boost both accuracy and robustness in complex reasoning scenarios.

Step 1	{ "role": "system", "content": "The following are multiple choice questions about subject, You need to start by eliminating the one option out of the four that least fits the question as Output 1. Then, from the remaining three options, eliminate the one that least fits as Output 2. Finally, from the last two options, eliminate the one that least fits and provide the correct answer as Output 3. " }
	
	[examples (The number of options corresponds to the questions below.)]
	
	{ "role": "user", "content": """Question: {Question}
	Options:
	{Question} """ }
	
	{ "role": "assistant", "content": "Output 1, Output 2, Output 3" }

Step 2	Given all of the above, the answer of the question is: A

Figure 8: When utilizing non-open-source LLMs, we are unable to directly access the model's raw output probabilities. Therefore, we employ GPT-3.5-turbo to generate three outputs based on the exclusion-based reasoning approach to simulate the exclusion process.



Question:
R, S, T, and U are points on a line, and U is the midpoint of line segment ST. If the lengths of line segments RS, RT, and ST are 5, 17, and 22, respectively. What is the length of line segment RU?
Options:
A. 6
B. 7
C. 8
D. 9
E. 16

RS = 5, RT = 17, ST = 22, and U is the midpoint of ST.
SU = UT = ST / 2 = 11.
Let R be at 0. Then S = 5, T = 27 (since S + ST = 5 + 22 = 27), and U is the midpoint of S and T, so U = 16.
RU = |16 - 0| = 16.
The correct answer is: **E. 16**

Given that ST = 22 and RT = 17, we have RS + ST ≥ RT. This order satisfies the conditions.
UU is the midpoint of ST, so:
SU = UT = ST / 2 = 22 / 2 = 11.
Since RU = RS + SU, we can use this relationship to eliminate incorrect options.
Calculate SU = 11, then substitute into RU = 5 + 11 = 16.
Comparing with the options, the correct answer is: **C. 8**

Since SR + RT = 22 = ST, then R is somewhere between S and T
Since ST is 22, then SU is 11 because U is the midpoint of ST.
Since SR < SU, then R is somewhere between S and U.
Then SR + RU = SU.
5 + RU = 11
RU = 6

Figure 9: An example illustrating the exclusion process. By systematically ruling out distractor options, the LLM correctly infers the answer.