Adaptive Distraction: Probing LLM Contextual Robustness with Automated Tree Search

Anonymous Author(s)
Affiliation
Address
email

Abstract

Large Language Models (LLMs) often struggle to maintain their original performance when faced with semantically coherent but task-irrelevant contextual information. Although prior studies have explored this issue using fixed-template or retrieval-based distractions, such static methods show limited effectiveness against contemporary models. To address this problem, we propose a dynamic distraction generation framework based on tree search, where the generation process is guided by model behavior. Without modifying the original question or answer, the method efficiently produces challenging adaptive distractions across multiple datasets, enabling systematic stress testing of LLMs' contextual robustness. Experiments on four benchmarks demonstrate that the generated distractions lead to an average performance drop of over 45% for mainstream models. Further comparisons of mitigation strategies show that prompt-based optimization methods yield limited gains, whereas post-training approaches (e.g., DPO) significantly enhance the model's contextual robustness. The results indicate that these issues do not stem from knowledge deficits in LLMs, but from a fundamental inability to maintain consistent reasoning under contextual distraction, posing a major challenge to the reliability of LLMs in real-world applications.

1 Introduction

Large Language Models (LLMs) have achieved remarkable success across diverse natural language processing tasks, such as question answering, summarization, and reasoning [1, 2, 3, 4]. However, recent studies reveal a critical vulnerability: LLMs are susceptible to semantically coherent but task-irrelevant contextual information, which can significantly degrade their performance [5]. This lack of contextual robustness hinders their ability to consistently focus on essential task content in the presence of distracting information, a challenge particularly pronounced in real-world applications where irrelevant context is common. Addressing this limitation is crucial to ensure the reliability of LLMs in complex, dynamic environments.

Current methods for evaluating LLM contextual robustness primarily rely on fixed-template or retrieval-based distractors [5, 6]. However, our preliminary experiments demonstrate that these static approaches are increasingly ineffective against contemporary models, with performance degradation often below 5% on advanced models like GPT-4o, which is detailed in Appendix B.2. This highlights that existing methods lack adaptiveness and are heavily dependent on the specific behavior of the target LLM. Once the model evolves, previously effective attack strategies may become obsolete. Moreover, such limited impact is insufficient to provide a reliable basis for robustness evaluation in realistic scenarios [7]. This highlights the urgent need for dynamic, adaptive methods capable of generating contextually plausible distractions that evolve with LLM capabilities, ensuring robust evaluation across diverse tasks and models.

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

To address the limitations of static methods, we aim to propose an adaptive attack method to generate **adaptive distractions**—semantically coherent, answer-preserving contextual additions that significantly impair LLM performance. This approach aims to evolve with advancing LLM capabilities, enabling robust stress-testing across diverse tasks without being constrained by model strength. However, generating such distractions presents key challenges: (1) ensuring semantic coherence with the original input, (2) preserving the correct answer, and (3) creating distractions potent enough to disrupt model predictions. These requirements demand a dynamic, model-informed generation strategy to effectively probe LLM contextual robustness.

To address these challenges, we propose a structured generation framework based on tree search to automatically construct adaptive distractions [8, 9, 10]. Our approach employs a classifier to pre-filter questions susceptible to perturbation, followed by a tree search module that explores contextual additions using a priority queue guided by model behavior. Error-guided perturbations generate candidate distractions at each node, evaluated for their ability to alter predictions while preserving semantic consistency. Early stopping strategies enhance efficiency, ensuring scalability across tasks and model families. Figure 1 illustrates this pipeline, which enables the controlled and automated creation of challenging distractions tailored to probe LLM contextual robustness.

We conducted comprehensive experiments to validate our framework, evaluating its effectiveness across four benchmark datasets, namely MMLU, CommonsenseQA, OpenbookQA, and TruthfulQA, on a diverse set of mainstream models, including proprietary and open-weight architectures. Our results show that adaptive distractions cause significant performance degradation, with an average accuracy drop exceeding 45%, exposing vulnerabilities in even the most advanced LLMs. Additionally, we explored mitigation strategies, comparing prompt-based approaches with targeted fine-tuning, and analyzed supplementary experiments, including prompt variants and case studies, detailed in the appendix. These findings collectively underscore the critical need for enhanced contextual robustness in LLMs.

In summary, our work delivers the following contributions:

- We introduce a novel framework for generating adaptive distractions, semantically coherent yet task-irrelevant additions, enabling robust and systematic evaluation of LLM contextual robustness.
- We provide empirical evidence of significant performance degradation, exceeding 45% accuracy drop, across four benchmark datasets and diverse mainstream models, uncovering persistent vulnerabilities in advanced LLMs.
- We evaluate mitigation strategies, revealing that targeted fine-tuning substantially enhances robustness under contextual distraction, while prompt-based approaches yield limited effectiveness.

2 Methodology

2.1 Overview

As illustrated in Figure 1, our objective is to systematically identify vulnerabilities in LLMs by generating adaptive distractions. As defined in the Introduction, these are contextual additions designed to preserve the original question's meaning and answer while affecting model performance.

Generating such distractions presents three core challenges. First, it requires an effective reward mechanism to guide the generation toward minimal yet impactful context additions. Second, the search space for valid distractions is vast, leading to substantial computational cost if not carefully controlled. Third, ensuring semantic consistency and bounded input length is critical, as irrelevant context may introduce unintended shifts or exceed the model's optimal context window.

To address these challenges, we propose a multi-step framework. We first apply a *classifier-based filtering* process to identify examples that are more likely to be affected by contextual interference, which narrows the search space. Then, we perform a *tree-based search* to explore semantically valid distractions. At each node, a *proxy model* generates candidate perturbations, which are evaluated through simulation using the victim model. The search is guided by a value function that combines model failure signals with depth penalties to ensure both quality and efficiency. Finally, we incorporate semantic validation and length control to preserve the original problem's intent, alongside pruning and early stopping strategies to reduce computation without sacrificing attack effectiveness.

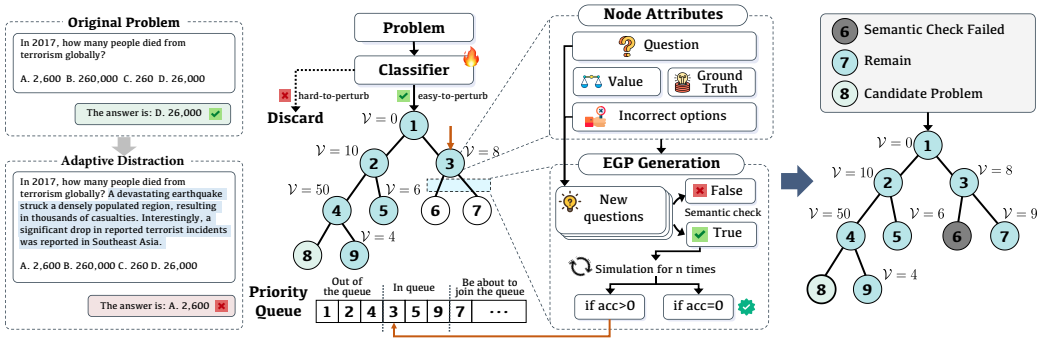

Figure 1: Overview of our framework. Given an input question, a classifier first filters for instances that are more susceptible to contextual interference. Then, a tree-based search explores candidate context additions using error-guided perturbation (EGP). At each node, candidate distractions are evaluated based on their ability to alter model predictions without affecting the correct answer. The framework efficiently produces high-quality *adaptive distraction examples* that challenge model robustness.

2.2 Problem Formulation

Let $D = \{P_1, P_2, \ldots, P_N\}$ denote a dataset consisting of N multiple-choice problem instances, where each instance is represented as a tuple:

$$P = \langle Q, A_{\text{gt}}, \mathcal{A}_{\text{inc}} \rangle, \tag{1}$$

with Q denoting the question, A_{gt} the ground truth answer, and $\mathcal{A}_{\text{inc}}$ a set of incorrect answers. Given a victim model M, our goal is to construct a perturbed dataset $D' = \{P'_1, P'_2, \ldots, P'_N\}$, where each perturbed instance $P' = \langle Q', A_{\text{gt}}, \mathcal{A}_{\text{inc}} \rangle$ is obtained by applying an *adaptive distraction* ΔQ to the question Q, such that:

$$Q' = Q + \Delta Q. \tag{2}$$

Our aim is to optimize the distraction ΔQ to minimize the accuracy of M on D', while ensuring semantic consistency and length constraints between Q and Q'. Here, semantic consistency is determined by a binary classifier S, which outputs $S(Q, Q') \in \{0, 1\}$, where $S(Q, Q') = 1$ indicates no semantic shift. Formally, the problem is expressed as:

$$\min_{\Delta Q} \quad \mathbb{E}_{P \sim D}\Big[\mathcal{L}_{\text{accuracy}}(M, Q')\Big], \text{s.t.} \quad S(Q, Q') = 1, \quad \frac{\text{len}(Q')}{\text{len}(Q)} \leq \lambda, \tag{3}$$

Here, S ensures that the distraction ΔQ does not lead to a semantic shift, while the length ratio constraint λ ensures Q' remains within acceptable bounds compared to the original Q. This constraint is necessary because recent studies show that LLMs experience performance degradation in long context scenarios [11]. To prevent excessive length expansion in Q', we introduce a length constraint, where λ is an upper bound on the relative length of Q' compared to Q.

If the output Q' does not satisfy the constraints in Equation 3, it is discarded, and a new distraction ΔQ is generated by re-prompting the proxy model.

2.3 Error-Guided Perturbation Generation

The distraction ΔQ is generated using a **proxy model**, denoted as P_{proxy}. The proxy model is prompted with the original problem instance $P = \langle Q, A_{\text{gt}}, \mathcal{A}_{\text{inc}} \rangle$ and tasked with generating a modified question Q' defined in Equation 2, where ΔQ represents the distraction introduced by P_{proxy}. The generation process is formalized as:

$$\Delta Q = P_{\text{proxy}}(Q, A_{\text{gt}}, \mathcal{A}_{\text{inc}}), \tag{4}$$

where P_{proxy} generates ΔQ based on a predefined prompt designed to guide the proxy model in producing contextual additions that increase the likelihood of the victim model M selecting an incorrect answer $a_{\text{inc}} \in \mathcal{A}_{\text{inc}}$ (i.e., lead the model to produce an error).

2.4 Tree-Based Perturbation Exploration

We employ a tree-based simulation-driven method to optimize distractions by heuristically exploring the search space. A priority queue is maintained to store nodes ordered by their value $\mathcal{V}(P')$, with the highest-value node dequeued and expanded iteratively using P_{proxy} to identify high-potential vulnerabilities.

Simulation For Measuring Distraction Quality. Firstly, we aim to design the reward of distracted questions to measure their value. For a given problem instance $P = \langle Q, A_{\text{gt}}, \mathcal{A}_{\text{inc}} \rangle$, the simulation process evaluates the quality of a distraction by estimating the success rate of the victim model M on P. Let $y \sim M(y \mid P)$ represent the output of the model M when queried on P. During a single simulation, the success rate $r_M(P)$ is computed by sampling n model outputs:

$$r_M(P) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}\{y_i = A_{\text{gt}}\}, \quad y_i \sim M(y \mid P), \tag{5}$$

where $\mathbb{I}\{\cdot\}$ is an indicator function that returns 1 if the model's output y_i matches the ground truth answer A_{gt}, and 0 otherwise. The success rate $r_M(P)$ quantifies the likelihood of the model producing the correct answer under the given distraction.

A distracted problem $P' = \langle Q', A_{\text{gt}}, \mathcal{A}_{\text{inc}} \rangle$ is considered effective if $r_M(P') = 0$, indicating that the model fails to produce the correct answer in all sampled outputs. The simulation process computes a value $\mathcal{V}(P')$ for the node corresponding to P' in the tree-based search:

$$\mathcal{V}(P') = \exp\left(\frac{\alpha}{r_M(P')}\right) \cdot \text{depth}^{-\gamma}, \quad \text{s.t. } r_M(P') \neq 0, \tag{6}$$

where α and γ are scaling constants, $r_M(P')$ is the success rate of the victim model M on P', and depth represents the recursion depth of the node in the search tree. For the question with $r_M(P') = 0$, we add it into the candidate problem list L, which stores examples that effectively *induce failure in the LLM*.

The simulation process systematically estimates $\mathcal{V}(P')$, prioritizing distracted problems with lower success rates $r_M(P')$, which correspond to higher potential vulnerabilities in the model. Simultaneously, the factor $\text{depth}^{-\gamma}$ discourages deeper recursions in the search tree, ensuring computational efficiency. High-value nodes with large $\mathcal{V}(P')$ scores are prioritized in the following tree-based search.

Tree-Based Search. For the tree-based search, the process begins with a root $P_{\text{root}} = \langle Q, A_{\text{gt}}, \mathcal{A}_{\text{inc}} \rangle$. A priority queue $\mathcal{Q}$ is maintained, where each node P' is ordered by its value $\mathcal{V}(P')$ in descending order. Initially, the root node is added to the queue as $\mathcal{Q} \leftarrow \mathcal{Q} \cup \{P_{\text{root}}\}$. At each iteration, the node P' with the highest value $\mathcal{V}(P')$ is dequeued for exploration:

$$P' = \arg\max_{P \in \mathcal{Q}} \mathcal{V}(P), \quad \mathcal{Q} \leftarrow \mathcal{Q} \setminus \{P'\}. \tag{7}$$

The proxy model P_{proxy} generates $k = |\mathcal{A}_{\text{inc}}|$ child nodes for P', corresponding to distractions ΔQ_j derived from each incorrect candidate answer $a_{\text{inc}} \in \mathcal{A}_{\text{inc}}$:

$$Q'_j = Q' + \Delta Q_j, \quad P'_j = \langle Q'_j, A_{\text{gt}}, \mathcal{A}_{\text{inc}} \rangle, \quad j = 1, 2, \ldots, k. \tag{8}$$

Each child node P'_j is evaluated by a simulation-driven method to compute its value $\mathcal{V}(P'_j)$, and the child nodes are added to the priority queue:

$$\mathcal{Q} \leftarrow \mathcal{Q} \cup \{P'_1, P'_2, \ldots, P'_k\}. \tag{9}$$

The search iteratively repeats this searching process, dynamically expanding the highest-value node and exploring the distraction space.

Why not Monte Carlo Tree Search? Monte Carlo Tree Search (MCTS) [9] has been widely used in recent studies to perform simulations powered by LLMs, achieving remarkable performance [12, 13, 14, 15]. However, MCTS is not suitable for our task due to its focus on balancing exploration (searching broadly across the tree) and exploitation (focusing on promising branches). In our context, such a balance is unnecessary because the width of the tree is inherently fixed, dictated by the number of incorrect answer candidates $|\mathcal{A}_{\text{inc}}|$. Moreover, MCTS introduces computational overhead by maintaining dynamic exploration strategies, which is impractical given the predefined structure and requirements of our method. Therefore, we opt for a simpler and more task-specific tree design that aligns directly with the properties of our problem.

2.5 Efficiency Strategies

Early Stopping Strategies. To reduce computational costs during the search process, we employ two early stopping strategies: diversity control and performance-based pruning.

The first strategy, diversity control, limits the number of child nodes considered at each search step. For a node P', if the number of child nodes P'_j satisfying $r_M(P'_j) = 0$ exceeds a predefined threshold n_1, we add the top n_1 child nodes to the candidate problem list L and directly pass this branch without further exploration. Formally, let $\mathcal{C}(P')$ represent the set of child nodes of P', and define:

$$\mathcal{C}_0(P') = \{P'_j \in \mathcal{C}(P') \mid r_M(P'_j) = 0\}. \tag{10}$$

If $|\mathcal{C}_0(P')| > n_1$, we update the candidate problem list L as:

$$L \leftarrow L \cup \mathcal{C}_0(P')[1 : n_1], \tag{11}$$

where $[1 : n_1]$ indicates the top n_1 nodes according to their values $\mathcal{V}(P'_j)$. The branch corresponding to P' is then terminated.

The second strategy is performance-based pruning, which bypasses nodes where further exploration is unlikely to yield meaningful results. For a node P', if all its child nodes satisfy $r_M(P'_j) = 1$, the node P' is skipped. Formally, if:

$$r_M(P'_j) = 1, \quad \forall P'_j \in \mathcal{C}(P'), \tag{12}$$

then P' is pruned from the search.

Additionally, if for l consecutive levels of the search tree, the minimum success rate $\min(r_M(P'))$ at each level increases monotonically from the top level to the bottom level, the corresponding node is bypassed. Let level_i represent the set of nodes at level i of the search tree, and define $m_i = \min_{P' \in \text{level}_i} r_M(P')$. If:

$$m_{i+1} > m_i, \quad \forall i \in \{1, 2, \ldots, l-1\}, \tag{13}$$

then the corresponding branch of the search tree is pruned.

Problem Filtering via Classifier. To reduce search costs, a classifier $C(Q)$ is used to filter out questions with low potential to become effective distraction candidates (e.g., the extremely easy question "What is the highest mountain in the world?"). The classifier is trained on previously searched questions, $\mathcal{D}_{\text{train}} = \{(Q_i, y_i)\}_{i=1}^{N}$, where $y_i \in \{0, 1\}$ indicates whether Q_i successfully exposes a vulnerability in the victim model M.

For each new question Q, the classifier computes $p(y = 1 \mid Q) = C(Q)$. Questions satisfying $p(y = 1 \mid Q) < \tau_C$, where τ_C is a predefined threshold, are discarded:

$$\mathcal{Q} \leftarrow \mathcal{Q} \setminus \{Q \mid p(y = 1 \mid Q) < \tau_C\} \tag{14}$$

Due to the space limitation, we show the overall algorithm in Appendix D.

3 Experiment

3.1 Experiment Setup

Selected Datasets. We selected four widely used benchmarks to evaluate contextual robustness under adaptive distraction: MMLU [16, 17], CommonsenseQA [18], OpenbookQA [19], and TruthfulQA [20]. These datasets cover diverse domains such as factual knowledge, commonsense reasoning, and elementary science, making them suitable for probing how LLMs handle irrelevant but semantically coherent contextual additions.

Models. As shown in Table 7, we used four proprietary models in our experiments: GPT-4o [21], GPT-4o-mini [22], Claude-3.5-Sonnet [23], and o1-mini [24]. Additionally, we included eight open-weight models: Gemma-2-2B, Gemma-2-27B [25], Qwen2.5-1.5B, Qwen2.5-7B, Qwen2.5-72B [26, 27], Llama-3.1-8B [28], Llama-3.1-70B [29], and Phi-3.5-mini [30].

Hyperparameter Setting. We set the temperature to 0.7 during the distraction generation phase to encourage more diverse and challenging outputs. For evaluation, we lowered the temperature to 0.001

Table 1: Accuracy of seven LLMs on four benchmarks before (**Original**) and after (**Perturbed**) applying our adaptive distractions. Cell background colors emphasise the severity of that drop.

Model	CommonsenseQA			OpenbookQA			TruthfulQA			MMLU		
	Original	Perturbed	Δ	Original	Perturbed	Δ	Original	Perturbed	Δ	Original	Perturbed	Δ
GPT-4o-mini	0.857	0.220	0.637	0.897	0.228	0.668	0.607	0.160	0.447	0.787	0.255	0.532
Llama-3.1-8B	0.753	0.230	0.524	0.807	0.212	0.595	0.570	0.283	0.288	0.697	0.300	0.397
Gemma-2-27B	0.857	0.249	0.607	0.867	0.231	0.636	0.782	0.449	0.332	0.753	0.340	0.413
o1-mini	0.856	0.296	0.560	0.897	0.377	0.519	0.748	0.523	0.226	0.803	0.451	0.352
Qwen2.5-72B	0.880	0.304	0.576	0.917	0.325	0.592	0.790	0.442	0.348	0.807	0.412	0.395
GPT-4o	0.890	0.277	0.613	0.950	0.375	0.575	0.757	0.494	0.263	0.870	0.552	0.318
Claude-3.5-sonnet	0.873	0.345	0.529	0.953	0.529	0.424	0.840	0.734	0.106	0.877	0.645	0.232

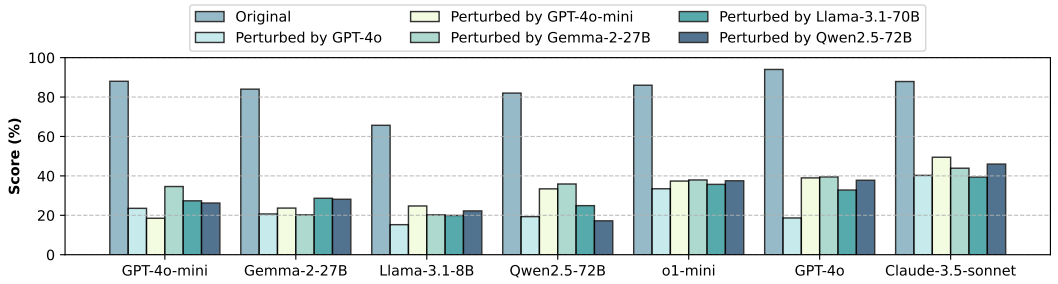

Figure 2: Performance of victim models on original questions and perturbed questions generated by different proxy LLMs.

to ensure response consistency, with a maximum output length of 1,024 tokens. Additionally, we set $\alpha = 2$ and $\gamma = 1$ for the value function used in the tree search. For other detailed hyperparameter settings, please refer to Appendix B.1.

Prompt Template. Prompt-based templates are used for several sub-tasks throughout our framework, including generating contextual distractions, assessing semantic consistency, evaluating model answers (zero-shot + CoT), baseline elaboration, filtering distraction-susceptible samples, and conducting robustness enhancement. The specific templates are provided in Appendix G.

Human Verification. To confirm semantic preservation, we conducted a human evaluation on randomly sampled perturbed questions. Annotators judged both semantic equivalence and answer consistency. Detailed results are reported in Appendix C.

3.2 Main Results

We conducted extensive evaluation experiments and mitigation studies to assess the impact of **adaptive distractions** on LLMs. The detailed configurations and dataset-model setups corresponding to each figure and table can be found in Appendix B.

Our method enables LLMs to autonomously generate adaptive distractions and effectively self-challenge. In Figure 2, we configure the same model to act as both the proxy and the victim in the distraction generation process. We observe that all tested models suffer a substantial drop in accuracy when evaluated on adaptively distracted questions, compared to their original performance. For instance, GPT-4o-mini experiences a performance decline of over 40%.

Furthermore, we uncover an intriguing pattern: distractions generated by a model itself tend to be more adversarial to that model than those generated by others. For example, GPT-4o-mini achieves an accuracy of only 0.185 on distractions it generated for itself, compared to 0.235 on those generated by the more advanced GPT-4o. This suggests that **models may be better at identifying their own weaknesses**, resonating with previous findings on self-alignment [31] and self-correction [32].

All models are susceptible to adaptive distractions, regardless of their scale or capability. Our results show that distractions created by stronger models can reliably challenge weaker models, while even distractions from weaker models can degrade the performance of stronger ones. As shown in Figure 2, powerful models such as GPT-4o and Claude-3.5-Sonnet still achieve below 50% accuracy when evaluated on distractions generated by Gemma-2-27B. This highlights a fundamental vulnerability: *no model is currently robust to adaptive distractions*.[1]

[1]We refer to the leaderboard at `https://lmarena.ai/` for model performance comparisons.

The extent of performance degradation varies significantly across tasks. As shown in Table 1, the average drop in accuracy on TruthfulQA is consistently smaller than that on OpenbookQA across all models. This suggests that distraction sensitivity varies by domain. Tasks like OpenbookQA, which require precise factual retrieval, appear more vulnerable to contextual interference than trustworthiness-based tasks such as TruthfulQA.

Our method outperforms existing distraction techniques by a large margin. We compare our method with several baselines: (1) *Irrelevant Context Augmentation (ICA)*, which adds semantically coherent but task-irrelevant information to extend question length [34]; (2) *Single-Prompt Distraction (SPD)*, which generates distractions via a single prompt without optimization [5]; and (3) DyVal2, a recent dynamic evaluation framework [33]. As shown in Table 2, our method results in an average accuracy drop of 52.0%, compared to 15.2% for DyVal2. Even strong models such as Claude-3.5-Sonnet experience a 38.4% absolute accuracy drop under our framework, compared to only 10.0% under DyVal2. These results highlight the strength of our tree-based search framework in systematically identifying contextual vulnerabilities, not merely increasing input complexity.

Table 2: Accuracy on original and distracted samples generated by different methods. **ICA**: Adding semantically coherent but task-irrelevant information to increase question length. **SPD**: Generating distractions via a single prompt without further optimization. **DyVal2**: Dynamic evaluation baseline [33]. Lower scores indicate more effective distractions.

Model	Original	ICA	SPD	DyVal2	Ours
GPT-4o-mini	0.890	0.727	0.760	0.630	0.185
Gemma-2-27B	0.860	0.788	0.790	0.650	0.237
Llama-3.1-8B	0.667	0.657	0.620	0.640	0.247
Qwen2.5-72B	0.820	0.737	0.810	0.697	0.334
o1-mini	0.860	0.694	0.770	0.697	0.374
GPT-4o	0.940	0.818	0.850	0.740	0.390
Claude-3.5-sonnet	0.879	0.838	0.820	0.780	0.495
Average	0.843	0.757	0.774	0.691	**0.323**

Adaptive distractions demonstrate robust cross-task generalization. We further examine the generalization of adaptive distraction by applying it to other tasks with well-defined ground truth, including mathematical reasoning. As shown in Appendix B.6, our method remains effective on the MATH-500 benchmark[35], suggesting that contextual distraction is not confined to factual QA but extends to any task where correctness can be explicitly evaluated.

3.3 Classifier: Filtering Hard-to-Perturb Problems

A critical challenge in adaptive distraction generation lies in distinguishing between **hard-to-perturb problems** (e.g., simple factual or arithmetic questions that are consistently answered correctly by LLMs regardless of added context) and **easy-to-perturb problems** (questions susceptible to semantic-preserving contextual distractions). Our analysis reveals that approximately 37% of computational resources are typically wasted on attempting to distract hard-to-perturb examples. To address this inefficiency, we design classifiers that predict the distraction susceptibility of a given question.

To evaluate the generalizability of the classifier, we examine whether perturbation difficulty is consistent across different LLMs. As shown in the confusion matrix in Figure 4, there is strong alignment between models: around 82% of questions are either distractable or non-distractable for both models in any pairwise comparison. This consistency suggests that model-agnostic classifiers can be trained to identify distractable inputs.

We implement two types of classifiers: (1) *Prompt-based classifiers*, which leverage LLM Judge [34, 36], and (2) *Fine-tuned classifiers*, trained on 1,080 annotated examples with 120 held-out test cases. As shown in Table 3, classifiers trained on data from GPT-4o-mini generalize effectively to stronger models, maintaining high precision across model families. We use the F_β score with $\beta = 0.5$ to prioritize precision. The formal definition is provided in Appendix B.1.

As illustrated in Figure 3 and Table 8, fine-tuned classifiers significantly improve overall efficiency by accurately filtering out hard-to-perturb samples. In particular, our best fine-tuned classifiers achieve up to 83% precision on identifying distractable problems, outperforming the best prompt-based baseline (GPT-4o) at 68%. The reduction in wasted computational effort and improvement in overall perturbation success rates will be analyzed in more detail in Experiment 3.4.

To further assess whether the classifier generalizes across tasks, we conduct a cross-dataset evaluation. Specifically, we train the classifier using three datasets (MMLU, CommonsenseQA, and OpenbookQA), and validate its performance on the held-out TruthfulQA. Table 10 presents the $F_{0.5}$ scores under this setting. Despite not being trained on TruthfulQA, the classifier maintains comparable performance, suggesting that it captures general signals of distraction susceptibility rather than dataset-specific patterns.

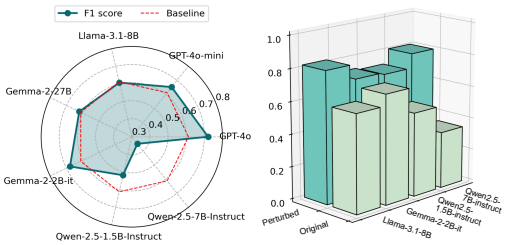

Figure 3: Comparison of classification performance using $F_{0.5}$ Scores. **Left:** $F_{0.5}$ scores of seven prompt-based classifiers compared to the baseline without a classifier (recall is 1 when all problems are enhanced directly). **Right:** $F_{0.5}$ scores of four fine-tuned classifiers after training, showing significant improvements over prompt-based classifiers.

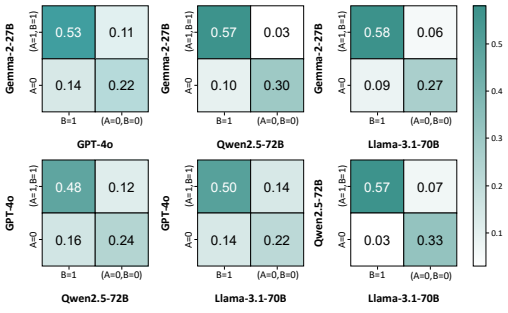

Figure 4: The result of whether the samples are perturbable by two models, A and B. Here, A=1 indicates that the sample is easy-to-perturb for model A, while A=0 means it is hard-to-perturb for model A. The numbers in each cell represent the percentage of samples in each category.

3.4 Ablation Study

Impact of the value function. We evaluate the effectiveness of the designed value function $\mathcal{V}(P')$ and its components introduced in Method 2.4 for guiding the search process. This function incorporates two key factors: the success rate $r_M(P')$ of the victim model on the distracted version of the problem P', and a depth penalty $\text{depth}^{-\gamma}$, which helps balance the trade-off between exploration and computational efficiency.

Table 3: The impact of the classifier on the perturbation success rate of the LLMs. The full model names are: GPT-4o, Gemma-2-27B, Llama-3.1-70B, and Qwen2.5-72B. The rows display the perturbation success rate with and without the classifier.

Mode	GPT-4o	Gemma-2	Llama-3.1	Qwen2.5
w/o classifier	0.527	0.592	0.581	0.563
w/ classifier	0.723	0.791	0.754	0.735

To assess the individual contributions of these components, we randomly selected questions from four datasets and measured the distraction success rate (defined as the percentage of questions for which the model fails completely under distraction, i.e., $r_M(P') = 0$). The results show that using the complete value function yields a 59% success rate. Removing the depth penalty reduces this rate to 57%, and removing the success rate term further decreases it to 53%. These findings confirm that both the model failure signal and the search depth control play important roles in identifying effective adaptive distractions.

Cost saved by classifier. To examine the impact of the classifier, we compared performance with and without classifier filtering. In this experiment, we randomly selected questions from four datasets. In the classifier condition, we used our classifier to select 100 questions predicted to be susceptible to distraction. In the baseline condition, we sampled 100 questions at random without any filtering.

As shown in Table 4, incorporating the classifier significantly improves both the effectiveness and the efficiency of the generation process. By filtering out hard-to-distract questions, the classifier enables better allocation of resources to more promising inputs. This leads to a 38.9% increase in success rate and a 21.9% reduction in average cost per successful distraction.

3.5 Mitigation under Adaptive Distraction

Notably, models that perform well on original questions but fail on distracted ones demonstrate that they possess the necessary knowledge yet remain vulnerable to contextual interference. To address

Table 4: Computational cost comparison with and without classifier. **Inp. Tok.**: Number of input tokens, **Out. Tok.**: Number of output tokens, **Pert. Ques.**: Number of successfully distracted questions, **Pert. Rate (%)**: Distraction success rate, **Cost/Ques. ($)**: USD cost per successfully distracted question.

Mode	Inp. Tok.	Out. Tok.	Pert. Ques.	Pert. Rate (%)	Cost ($)
w/o classifier	3.69M	1.48M	175	59%	0.0082
w/ classifier	3.81M	1.57M	236	82%$_{\uparrow 38.9\%}$	0.0064$_{\downarrow 21.9\%}$

this issue, we explored both **prompt-based (i.e., training-free)** and **training-based strategies** to improve their performance on these challenging questions.

Prompt-based mitigation is unable to alleviate distraction vulnerability. Since our adaptive distractions preserve the core question semantics while introducing task-irrelevant context, we tested whether adding explicit instructions in the prompt could help models focus on the essential content and ignore misleading information. Specifically, we modified the original prompts to include guidance on identifying and prioritizing key components of the question. Detailed templates are provided in Appendix G. As shown in Table 5, this approach yielded only marginal improvements. Some models, such as o1-mini and Claude-3.5-Sonnet, even showed slightly lower accuracy after prompt modifications. This suggests that the contextual interference introduced by adaptive distractions cannot be effectively mitigated through prompt refinement alone.

Table 5: Model accuracy before and after prompt-based mitigation (**Orig.** vs **Enh.**).

Model	Orig.	Enh.	Diff.
GPT-4o-mini	0.185	0.211	+0.026
Llama-3.1-8B	0.247	0.251	+0.003
Gemma-2-27B	0.237	0.255	+0.018
o1-mini	0.374	0.366	−0.008
Qwen2.5-72B	0.334	0.343	+0.009
GPT-4o	0.390	0.391	+0.002
Claude-3.5-sonnet	0.495	0.481	−0.013

We also tested additional prompting methods that have shown promise in prior work, including In-context learning [37] and Self-consistency [38]. As shown in Appendix B.5, none of these approaches substantially recover the original performance when faced with adaptive distractions. This reinforces our finding that prompt-based methods alone offer limited robustness.

As shown in Table 6, all three models benefited significantly from training. The Phi-3.5-mini model, in particular, achieved a post-training accuracy that surpassed even GPT-4o on the same distracted inputs. Detailed case studies in Appendix F show that the improvements were not merely due to new knowledge acquisition. Rather, fine-tuned models showed better focus on relevant question content and stronger resistance to irrelevant distractions. A large fraction of the originally incorrect answers remained incorrect, such as 82.1% for Phi-3.5-mini, suggesting that performance gains came from improved robustness rather than memorization. Additional analysis in Appendix B.7 confirms that DPO offers greater gains than supervised fine-tuning, while preserving performance on clean questions, which further validates that our improvements stem from true robustness rather than memorization.

Table 6: Model accuracy before and after DPO training. **Retain** shows the fraction of original incorrect answers that remain incorrect after training.

Model	Orig.	Enh.	Diff.	Retain
Gemma-2-2B	0.257	0.432	+0.175	0.788
Qwen2.5-7B	0.212	0.440	+0.228	0.763
Phi-3.5-mini	0.195	0.680	+0.485	0.821
GPT-4o	0.568	–	–	–
Qwen2.5-72B	0.519	–	–	–
GPT-4o-mini	0.232	–	–	–

4 Conclusion

In this work, we propose a framework to assess the contextual robustness of language models by generating adaptive distractions, which are semantically coherent but task-irrelevant additions. Our tree-based search method produces challenging examples that induce consistent performance drops across models and datasets. Among mitigation strategies, post-training methods such as DPO offer the most reliable improvements. Ultimately, our approach offers a scalable tool for evaluating and improving LLM reliability in real-world applications. Future work will integrate our distraction generation into training loops to further strengthen contextual robustness.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[2] Hongjian Zhou, Fenglin Liu, Boyang Gu, Xinyu Zou, Jinfa Huang, Jinge Wu, Yiru Li, Sam S Chen, Peilin Zhou, Junling Liu, et al. A survey of large language models in medicine: Progress, application, and challenge. *arXiv preprint arXiv:2311.05112*, 2023.

[3] Yue Huang, Chujie Gao, Siyuan Wu, Haoran Wang, Xiangqi Wang, Yujun Zhou, Yanbo Wang, Jiayi Ye, Jiawen Shi, Qihui Zhang, Yuan Li, Han Bao, Zhaoyi Liu, Tianrui Guan, Dongping Chen, Ruoxi Chen, Kehan Guo, Andy Zou, Bryan Hooi Kuen-Yew, Caiming Xiong, Elias Stengel-Eskin, Hongyang Zhang, Hongzhi Yin, Huan Zhang, Huaxiu Yao, Jaehong Yoon, Jieyu Zhang, Kai Shu, Kaijie Zhu, Ranjay Krishna, Swabha Swayamdipta, Taiwei Shi, Weijia Shi, Xiang Li, Yiwei Li, Yuexing Hao, Zhihao Jia, Zhize Li, Xiuying Chen, Zhengzhong Tu, Xiyang Hu, Tianyi Zhou, Jieyu Zhao, Lichao Sun, Furong Huang, Or Cohen Sasson, Prasanna Sattigeri, Anka Reuel, Max Lamparth, Yue Zhao, Nouha Dziri, Yu Su, Huan Sun, Heng Ji, Chaowei Xiao, Mohit Bansal, Nitesh V. Chawla, Jian Pei, Jianfeng Gao, Michael Backes, Philip S. Yu, Neil Zhenqiang Gong, Pin-Yu Chen, Bo Li, and Xiangliang Zhang. On the trustworthiness of generative foundation models: Guideline, assessment, and perspective. *arXiv preprint arXiv:2502.14296*, 2025.

[4] Zixiang Xu, Yanbo Wang, Yue Huang, Jiayi Ye, Haomin Zhuang, Zirui Song, Lang Gao, Chenxi Wang, Zhaorun Chen, Yujun Zhou, et al. Socialmaze: A benchmark for evaluating social reasoning in large language models. *arXiv preprint arXiv:2505.23713*, 2025.

[5] Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed H Chi, Nathanael Schärli, and Denny Zhou. Large language models can be easily distracted by irrelevant context. In *International Conference on Machine Learning*, pages 31210–31227. PMLR, 2023.

[6] Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *arXiv preprint arXiv:2307.03172*, 2023.

[7] Erfan Shayegani, Md Abdullah Al Mamun, Yu Fu, Pedram Zaree, Yue Dong, and Nael Abu-Ghazaleh. Survey of vulnerabilities in large language models revealed by adversarial attacks. *arXiv preprint arXiv:2310.10844*, 2023.

[8] Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. *Advances in neural information processing systems*, 36:11809–11822, 2023.

[9] Cameron B Browne, Edward Powley, Daniel Whitehouse, Simon M Lucas, Peter I Cowling, Philipp Rohlfshagen, Stephen Tavener, Diego Perez, Spyridon Samothrakis, and Simon Colton. A survey of monte carlo tree search methods. *IEEE Transactions on Computational Intelligence and AI in games*, 4(1):1–43, 2012.

[10] Zixiang Xu, Yanbo Wang, Yue Huang, Xiuying Chen, Jieyu Zhao, Meng Jiang, and Xiangliang Zhang. Cross-lingual pitfalls: Automatic probing cross-lingual weakness of multilingual large language models. *arXiv preprint arXiv:2505.18673*, 2025.

[11] Yushi Bai, Xin Lv, Jiajie Zhang, Hongchang Lyu, Jiankai Tang, Zhidian Huang, Zhengxiao Du, Xiao Liu, Aohan Zeng, Lei Hou, et al. Longbench: A bilingual, multitask benchmark for long context understanding. *arXiv preprint arXiv:2308.14508*, 2023.

[12] Dan Zhang, Sining Zhoubian, Ziniu Hu, Yisong Yue, Yuxiao Dong, and Jie Tang. Rest-mcts*: Llm self-training via process reward guided tree search. *arXiv preprint arXiv:2406.03816*, 2024.

[13] Hao Wang, Boyi Liu, Yufeng Zhang, and Jie Chen. Seed-cts: Unleashing the power of tree search for superior performance in competitive coding tasks. *arXiv preprint arXiv:2412.12544*, 2024.

[14] Yuxi Xie, Anirudh Goyal, Wenyue Zheng, Min-Yen Kan, Timothy P Lillicrap, Kenji Kawaguchi, and Michael Shieh. Monte carlo tree search boosts reasoning via iterative preference learning. *arXiv preprint arXiv:2405.00451*, 2024.

[15] Xinyu Guan, Li Lyna Zhang, Yifei Liu, Ning Shang, Youran Sun, Yi Zhu, Fan Yang, and Mao Yang. rstar-math: Small llms can master math reasoning with self-evolved deep thinking. *arXiv preprint arXiv:2501.04519*, 2025.

[16] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021.

[17] Dan Hendrycks, Collin Burns, Steven Basart, Andrew Critch, Jerry Li, Dawn Song, and Jacob Steinhardt. Aligning ai with shared human values. *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021.

[18] Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. CommonsenseQA: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4149–4158, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[19] Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. Can a suit of armor conduct electricity? a new dataset for open book question answering. In *EMNLP*, 2018.

[20] Stephanie Lin, Jacob Hilton, and Owain Evans. Truthfulqa: Measuring how models mimic human falsehoods, 2021.

[21] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

[22] OpenAI. Gpt-4o mini: Advancing cost-efficient intelligence. `https://openai.com/index/gpt-4o-mini-advancing-\cost-efficient-intelligence/`, 2024.

[23] Anthropic. Claude 3.5 sonnet. `https://www.anthropic.com/news/claude-3-5-sonnet`, 2024.

[24] Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. Openai o1 system card. *arXiv preprint arXiv:2412.16720*, 2024.

[25] Gemma Team. Gemma. 2024.

[26] An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zhihao Fan. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2024.

[27] Qwen Team. Qwen2.5: A party of foundation models, September 2024.

[28] Meta. Llama 3.1-8b. `https://huggingface.co/meta-llama/Llama-3.1-8B`, 2024.

[29] Meta. Llama 3.1-70b. `https://huggingface.co/meta-llama/Llama-3.1-70B`, 2024.

[30] Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, et al. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*, 2024.

[31] Zhiqing Sun, Yikang Shen, Qinhong Zhou, Hongxin Zhang, Zhenfang Chen, David Cox, Yiming Yang, and Chuang Gan. Principle-driven self-alignment of language models from scratch with minimal human supervision. *Advances in Neural Information Processing Systems*, 36, 2024.

[32] Liangming Pan, Michael Saxon, Wenda Xu, Deepak Nathani, Xinyi Wang, and William Yang Wang. Automatically correcting large language models: Surveying the landscape of diverse self-correction strategies. *arXiv preprint arXiv:2308.03188*, 2023.

[33] Kaijie Zhu, Jindong Wang, Qinlin Zhao, Ruochen Xu, and Xing Xie. Dynamic evaluation of large language models by meta probing agents. In *Forty-first International Conference on Machine Learning*, 2024.

[34] Jiayi Ye, Yanbo Wang, Yue Huang, Dongping Chen, Qihui Zhang, Nuno Moniz, Tian Gao, Werner Geyer, Chao Huang, Pin-Yu Chen, et al. Justice or prejudice? quantifying biases in llm-as-a-judge. *arXiv preprint arXiv:2410.02736*, 2024.

[35] Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. *arXiv preprint arXiv:2305.20050*, 2023.

[36] Yanbo Wang, Jiayi Ye, Siyuan Wu, Chujie Gao, Yue Huang, Xiuying Chen, Yue Zhao, and Xiangliang Zhang. Trusteval: A dynamic evaluation toolkit on trustworthiness of generative foundation models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (System Demonstrations)*, pages 70–84, 2025.

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

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

[39] Giannis Chatziveroglou, Richard Yun, and Maura Kelleher. Exploring llm reasoning through controlled prompt variations. *arXiv preprint arXiv:2504.02111*, 2025.

[40] Nora Kassner and Hinrich Schütze. Negated and misprimed probes for pretrained language models: Birds can talk, but cannot fly. *arXiv preprint arXiv:1911.03343*, 2019.

[41] Ori Yoran, Tomer Wolfson, Ori Ram, and Jonathan Berant. Making retrieval-augmented language models robust to irrelevant context. *arXiv preprint arXiv:2310.01558*, 2023.

[42] Jiawen Shi, Zenghui Yuan, Yinuo Liu, Yue Huang, Pan Zhou, Lichao Sun, and Neil Zhenqiang Gong. Optimization-based prompt injection attack to llm-as-a-judge. In *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security*, pages 660–674, 2024.

[43] Anselm Paulus, Arman Zharmagambetov, Chuan Guo, Brandon Amos, and Yuandong Tian. Advprompter: Fast adaptive adversarial prompting for llms. *arXiv preprint arXiv:2404.16873*, 2024.

[44] Ming Jiang, Tingting Huang, Biao Guo, Yao Lu, and Feng Zhang. Enhancing robustness in large language models: Prompting for mitigating the impact of irrelevant information. *arXiv preprint arXiv:2408.10615*, 2024.

[45] Chrisantha Fernando, Dylan Banarse, Henryk Michalewski, Simon Osindero, and Tim Rocktäschel. Promptbreeder: Self-referential self-improvement via prompt evolution. *arXiv preprint arXiv:2309.16797*, 2023.

[46] Zhenyu Wu, Chao Shen, and Meng Jiang. Instructing large language models to identify and ignore irrelevant conditions. *arXiv preprint arXiv:2403.12744*, 2024.

[47] Youxiang Zhu, Ruochen Li, Danqing Wang, Daniel Haehn, and Xiaohui Liang. Focus directions make your language models pay more attention to relevant contexts. *arXiv preprint arXiv:2503.23306*, 2025.

[48] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36:53728–53741, 2023.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The main claims in the abstract and introduction are supported by the contributions and experimental evidence in the paper. Specifically, our proposed adaptive distraction framework (Section 2) leads to significant accuracy drops across models (Section 3.2), and the analysis in Section 3.5 demonstrates that such drops are due to contextual interference rather than knowledge deficits.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The limitations of our method are discussed in Section 3.3 and 3.4. We acknowledge the computational cost of tree search, the dependence on classifier quality for efficient filtering, and variability in attack difficulty across datasets such as MMLU. These constraints limit scalability in certain settings and are clearly presented. We also include a Limitations and Broader Impacts section in the appendix.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best

judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

   Justification: The paper does not include formal theoretical results or proofs.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: All major experimental settings, including datasets, model types, hyperparameters, sampling temperatures, and prompt templates, are fully disclosed in Section 3 and Appendix B and F.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: We have provided anonymized versions of our code and data in the supplementary material. These include scripts, prompt templates, and configuration details sufficient to reproduce our main results. See Appendix B and F for descriptions.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: All training and evaluation settings are provided in the main paper and Appendix B.1, including data splits, model versions, hyperparameters, and prompting strategies. These settings allow a full understanding of the results.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

Justification: No, we did not compute error bars or statistical significance measures due to the high computational cost of our large-scale experiments.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We report compute settings including GPU types (e.g., RTX 4090), training configurations (e.g., batch size, learning rate, number of epochs), and cost per sample in Table 4. Total input/output token counts and cost breakdowns are provided for both classifier and perturbation generation experiments.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research complies with the NeurIPS Code of Ethics. It does not involve human subjects or sensitive personal data, and no ethical risks are posed by the datasets or models used.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.

- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The paper discusses the broader impact of contextual robustness on real-world LLM reliability, particularly in high-stakes applications where irrelevant context can lead to incorrect reasoning (Section 1 and 4). Our method enables systematic stress testing, which supports safer model deployment. We also include a Limitations and Broader Impacts section in the appendix.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: The paper does not release any models or datasets that pose a high risk of misuse. The goal is to evaluate model robustness in a controlled academic setting, and no sensitive or potentially harmful content is generated or distributed.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: All datasets and models used in this work are properly cited and their licenses are respected. The license type, version, and organization are listed in Appendix B (Table 7), including proprietary and open-weight models such as Gemma, Qwen, and Phi-3.5-mini.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: We release anonymized code for generation and evaluation in the supplementary material as new assets. Documentation describing their structure and usage is included.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [Yes]

Justification: We conducted a human evaluation study using five student annotators to assess semantic and answer consistency between original and perturbed questions (Appendix C). Task details and evaluation instructions are described, although no monetary compensation was involved.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.

- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The paper does not involve high-risk research with human subjects. The human evaluation involved internal participants (students) performing low-risk annotation tasks without collection of personal data or sensitive content, and thus did not require IRB approval.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: LLM is used only for writing, editing, or formatting purposes.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.

A Related Work

A.1 Contextual Robustness in LLMs

Recent studies have highlighted that LLMs often fail when presented with semantically coherent yet task-irrelevant contextual information. Shi et al. [5] introduced a benchmark demonstrating that irrelevant context can severely degrade LLM accuracy in arithmetic reasoning tasks. Liu et al. [6] showed that model performance significantly drops when key information is placed in the middle of long contexts, indicating positional sensitivity in attention mechanisms. Similar findings have emerged in narrative distraction scenarios [39], misprimed probe studies [40], and irrelevant document retrieval contexts [41]. Collectively, these works underscore contextual distraction as a prevalent vulnerability affecting modern LLMs.

A.2 Generation of Contextual Perturbations

To systematically probe LLM sensitivity to irrelevant context, various methods have been developed to generate targeted perturbations. Zhu et al. [33] proposed a dynamic evaluation approach using meta probing agents that restructure tasks to surface latent weaknesses in model behavior. Similarly, optimization-based prompt injection methods have been explored to create adversarial inputs aimed at exploiting model biases or alignment issues [42, 43]. Chatziveroglou et al. [39] further validate that even semantically coherent but irrelevant narratives can significantly reduce LLM accuracy. Despite these efforts, existing perturbation generation techniques typically focus on altering prompts or instructions without necessarily preserving answer correctness or targeting semantic coherence explicitly.

A.3 Mitigation Strategies

Several approaches have attempted to address the issue of contextual distraction by enhancing model robustness [44]. PromptBreeder [45] employs evolutionary strategies to optimize task prompts, implicitly strengthening model robustness against perturbations. Wu et al. [46] introduced prompting strategies instructing models to explicitly ignore irrelevant information, and Wang et al. [38] showed that self-consistency decoding can improve reliability by aggregating predictions from multiple reasoning paths. Zhu et al. [47] explored attention mechanisms, identifying internal attention directions to guide models toward more relevant context. Moreover, Yoran et al. [41] demonstrated the effectiveness of fine-tuning on mixed-relevance data to improve robustness in retrieval-augmented scenarios. Direct preference optimization (DPO) [48], a targeted fine-tuning technique, has gained particular attention for its effectiveness in enhancing model alignment and resilience to distracting inputs. While prompt-based approaches remain attractive due to their simplicity, our findings indicate they have limited efficacy against adaptive distractions. Our experiments highlight the stronger robustness achieved by fine-tuning strategies, validating the effectiveness of targeted mitigation approaches in addressing contextual distraction.

B Experiment Details

Table 7: Models used in our experiments along with their versions, organizations, licenses, and purposes. *Gen*: Model used for generating questions (as a proxy or victim); *Eval*: Model used for evaluating datasets; *Clf*: Model used as a classifier to filter questions.

Model	Version	Organization	License	Gen	Eval	Clf
GPT-4o-mini	gpt-4o-mini-2024-07-18	OpenAI	Proprietary	✓	✓	
GPT-4o	gpt-4o-2024-08-06	OpenAI	Proprietary	✓	✓	
Gemma-2-2B	Gemma-2-2B-it	Google	Gemma License		✓	✓
Gemma-2-27B	Gemma-2-27B-it	Google	Gemma License	✓	✓	
Llama-3.1-8B	Meta-Llama-3.1-8B-Instruct	Meta	Llama 3.1 Community		✓	✓
Llama-3.1-70B	Meta-Llama-3.1-70B-Instruct	Meta	Llama 3.1 Community	✓	✓	
Qwen2.5-1.5B	Qwen2.5-1.5B-Instruct	Alibaba	Qwen License			✓
Qwen2.5-7B	Qwen2.5-7B-Instruct	Alibaba	Qwen License		✓	✓
Qwen2.5-72B	Qwen2.5-72B-Instruct	Alibaba	Qwen License	✓	✓	
o1-mini	o1-mini-2024-09-12	OpenAI	Proprietary		✓	
Phi-3.5-mini	Phi-3.5-mini-instruct	Microsoft	MIT		✓	✓
Claude-3.5-Sonnet	claude-3-5-sonnet-20241022	Anthropic	Proprietary		✓	

B.1 Experiment Settings

In all experiments, we adopt the same parameter settings. Specifically, we set the length threshold $\lambda = 10$, the semantic threshold $\tau_C = 0.5$, the number of simulation times $n = 5$, and the diversity limit $n_1 = 3$. Additionally, we use the same model as both the proxy model and the victim model.

Experimental details of different victim models. We selected five victim models with varying capabilities: GPT-4o, GPT-4o-mini, Llama-3.1-70B, Qwen2.5-72B, and Gemma-2-27B. From each of the four datasets, namely MMLU, CommonsenseQA, OpenbookQA, and TruthfulQA, we randomly sampled 100 original questions. Each victim model enhanced these questions via our search framework, creating five distinct enhanced datasets. To evaluate the effectiveness of these enhanced questions, we tested the performance of seven different models: GPT-4o-mini, Gemma-2-27B, Llama-3.1-8B, Qwen2.5-72B, o1-mini, GPT-4o, and Claude-3.5-Sonnet. All models were evaluated using a zero-shot approach with CoT prompting templates. This setup allowed us to systematically analyze the relationship between victim model capability and the difficulty of the generated enhanced questions. The results of this experiment are summarized in Figure 2.

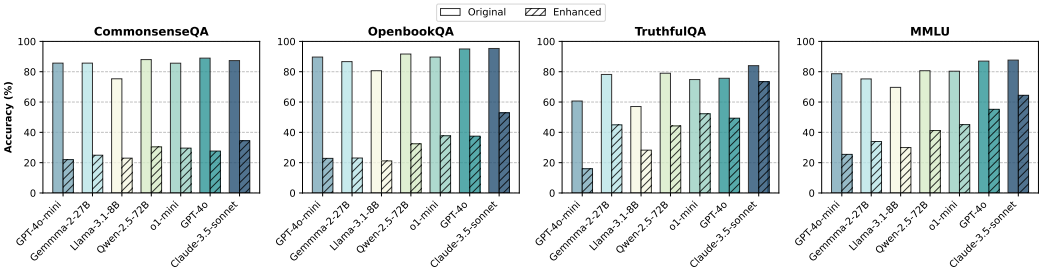

Figure 5: Overall results between 4 datasets.

Experimental details of scale-up experiment. We selected GPT-4o-mini as the victim model for question enhancement due to its balance between perturbation effectiveness and computational efficiency. From the same four datasets, we sampled 300 questions per dataset, resulting in a total of 1200 original questions. Similar to the first experiment, the enhanced questions were tested across the same seven models: GPT-4o-mini, Gemma-2-27B, Llama-3.1-8B, Qwen2.5-72B, o1-mini, GPT-4o, and Claude-3.5-Sonnet. All evaluations were conducted using zero-shot CoT prompting templates. This larger-scale experiment provided a more comprehensive analysis of the generalizability of our perturbation methodology. The results of this experiment are summarized in Figure 5 and Table 1.

Experimental details of baseline methods. To validate the effectiveness of our tree-based search framework, we implemented two baseline perturbation approaches for comparison. The **Irrelevant Context Augmentation (ICA)** method performed semantic-preserving length augmentation by expanding original questions with task-irrelevant but semantically coherent contextual information, such as explanatory clauses or redundant details. The **Single-Prompt Distraction (SPD)** baseline utilized our perturbation prompt template (details in Appendix G) through Claude-3.5-Sonnet for automatic distraction generation without subsequent search optimization. For a fair comparison, all baseline methods processed the same 100 original questions from four datasets using Claude-3.5-Sonnet as the executor. The enhanced questions were evaluated under identical zero-shot CoT settings across seven target models. This demonstrates the crucial role of our tree-based search mechanism in identifying optimal perturbation combinations rather than relying on simple length expansion or single-pass prompt perturbations. The results of this experiment are summarized in Table 2.

Experimental details of classifier. We used 1200 original questions from 4 datasets, splitting them into training, test, and validation sets. Specifically, 80 percent of the data was allocated to training, with 10 percent of the training set reserved for validation, and the remaining 10 percent was used for testing. For the prompt-based classifiers, we designed specific prompts to guide the models in determining whether a problem was hard to perturb. We evaluated the classification performance of seven models: GPT-4o-mini, GPT-4o, Llama-3.1-8B, Gemma-2-27B, Gemma-2-2B, Qwen2.5-1.5B, and Qwen2.5-7B. A baseline configuration without any classifier was also included for comparison. The effectiveness of these classifiers was measured using the F1-score with beta equal to 0.5, which prioritizes precision over recall. For the training-based classifiers, we used supervised fine-tuning with LoRA on four open-source models: Llama-3.1-8B, Gemma-2-2B, Qwen2.5-1.5B, and Qwen2.5-7B.

The training was conducted on a single RTX 4090 GPU, with a learning rate set to 1e-4 and a total of five epochs. The performance of these fine-tuned classifiers was also evaluated using the F1-score on the test set. This experimental design allowed us to compare the utility of prompt-based and training-based classifiers in identifying hard-to-perturb questions. The results of this experiment are summarized in Figure 4, Table 3, Figure 3 and Table 8.

To evaluate the tradeoff between precision and recall in our classifier analysis, we report the F_β score with $\beta = 0.5$. This metric places greater emphasis on precision, which is desirable in our use case. The score is defined as:

$$F_\beta = (1 + \beta^2) \times \frac{\text{Precision} \times \text{Recall}}{(\beta^2 \times \text{Precision}) + \text{Recall}}. \tag{15}$$

Table 8: Performance of the classifier under Prompt-Based and Fine-Tuned methods. The table reports Precision, Recall, and $F_{0.5}$ scores for both Prompt-Based (left) and Fine-Tuned (right) classifiers. Fine-Tuned models are marked in the Fine-Tuned columns. Baseline represents the performance of the system without using classifier.

Model	Prompt-Based			Fine-Tuned		
	Precision	Recall	$F_{0.5}$	Precision	Recall	$F_{0.5}$
GPT-4o-mini	0.606	0.940	0.652	–	–	–
GPT-4o	**0.685**	0.910	**0.721**	–	–	–
Llama-3.1-8B	0.555	0.985	0.608	**0.812**	0.836	**0.816**
Gemma-2-27B	0.568	1.000	0.622	–	–	–
Gemma-2-2B	0.558	0.866	0.678	0.712	0.776	0.724
Qwen2.5-1.5B	0.534	0.463	0.518	0.719	0.687	0.712
Qwen2.5-7B	0.526	0.149	0.350	0.797	0.821	0.802
Baseline	0.558	1.000	0.612	0.558	1.000	0.612

Experimental details of mitigation. We curated approximately 1200 preference data pairs. Each preference pair consisted of a question, a correct answer, and an incorrect answer collected from model responses in prior experiments. To ensure a fair evaluation, we guaranteed that enhanced questions originating from the same original question did not appear in both the training and test sets. The data was split into training, validation, and test sets, with 80 percent of the data used for training, 10 percent of the training set reserved for validation, and 20 percent allocated to testing. For prompt-based enhancement, we designed new prompt templates aimed at improving model focus on the core question content and tested them on seven models: GPT-4o-mini, Gemma-2-27B, Llama-3.1-8B, Qwen2.5-72B, o1-mini, GPT-4o, and Claude-3.5-Sonnet. For training-based enhancement, we fine-tuned three open-source models, namely Gemma-2-2B, Qwen2.5-7B, and Phi-3.5-mini. Using the Direct Preference Optimization algorithm, the fine-tuning was performed on two RTX 4090 GPUs with a learning rate set to 2e-4 and five epochs. The preference loss was implemented with a sigmoid activation function. The fine-tuned models were evaluated against three high-performance baseline models, specifically GPT-4o, GPT-4o-mini, and Qwen2.5-72B, using the original zero-shot with CoT prompting templates on the test set. This experiment provided insights into the effectiveness of both prompt-based and training-based approaches in improving model robustness against enhanced questions. The results of this experiment are summarized in Table 5 and Table 6.

B.2 Preliminary Experimental Results

To support our claim in the introduction regarding the limited effectiveness of static distraction methods, we present a comparison of performance drops induced by our adaptive distraction framework and a representative static method, GSM-IC [5], in Table 9. The results demonstrate that while GSM-IC causes minimal performance drops (average 1.8%) on advanced models, our method achieves significantly larger drops (average 45%), highlighting its potency in challenging LLM contextual robustness.

Table 9: Comparison of performance drops (%) on GSM-IC [5] and our adaptive distraction.

Model	GPT-4o-mini	GPT-4o	Qwen2.5-72B	Gemma-2-27B	Claude-3.5-sonnet	o1-mini
GSM-IC [5]	4.1	1.2	1.4	2.2	0.8	1.2

B.3 Experiment Analysis

Distribution Analysis of Enhanced Questions. Our analysis of the search process reveals interesting patterns in both the depth of perturbation chains and the length ratios of enhanced questions across different datasets. As shown in Figure 6, the majority of successful perturbations were found at relatively shallow depths, particularly for CommonsenseQA and OpenbookQA, where approximately 85% and 80% of effective perturbations were discovered within the first three levels. However, MMLU exhibited a notably different pattern, with nearly 30% of perturbations requiring five or more steps to achieve effectiveness. This suggests that questions testing specialized knowledge often require more sophisticated and layered perturbations to successfully challenge model performance. The length ratios of enhanced questions also varied significantly across datasets. OpenbookQA showed a tendency toward longer perturbations, with about 70% of enhanced questions being more than five times longer than their original versions. In contrast, MMLU questions maintained relatively compact perturbations, with nearly half of the enhanced questions staying within three times the original length. These distributions reflect the varying complexity required to effectively perturb different types of questions and highlight how the nature of the underlying task influences the perturbation process.

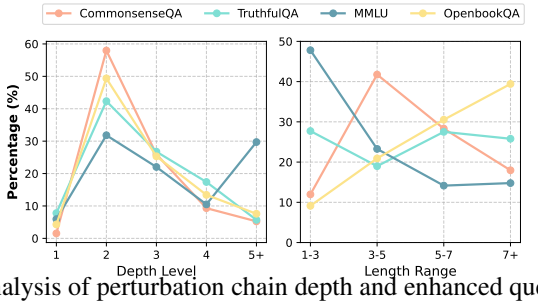

Figure 6: Distribution analysis of perturbation chain depth and enhanced question length ratio across four datasets.

B.4 Cross-Dataset Generalization of the Classifier

We present additional evaluation of the classifier's ability to generalize across datasets. As shown in Table 10, the $F_{0.5}$ score remains stable even when the classifier is tested on a domain it was not trained on.

Table 10: $F_{0.5}$ scores of the classifier trained on MMLU, CommonsenseQA, and OpenbookQA, and tested on the held-out TruthfulQA dataset.

Model	Orig. Test	Cross-Dataset
Gemma-2-2B	0.724	0.735 (+0.011)
Qwen2.5-1.5B	0.712	0.698 (-0.014)
Llama-3.1-8B	0.816	0.782 (-0.034)

B.5 Additional Prompting Strategies

We further evaluate widely used prompting methods on the same set of adaptively distracted questions. Despite using more sophisticated prompting, the performance gains are marginal, confirming that contextual distraction remains a persistent challenge even under varied prompting schemes.

Table 11: Accuracy of various prompting strategies on adaptive distraction examples. All numbers are averaged across four datasets. Values in parentheses indicate drop from clean accuracy.

Model	Vanilla	Zero +CoT [37]	Few +CoT [37]	Few-shot (ICL) [37]	Self-Consistency [38]
GPT-4o-mini	0.185 (-0.705)	0.211 (-0.679)	0.238 (-0.652)	0.262 (-0.628)	0.272 (-0.618)
GPT-4o	0.390 (-0.550)	0.391 (-0.549)	0.455 (-0.485)	0.443 (-0.497)	0.461 (-0.479)
Qwen2.5-72B	0.334 (-0.486)	0.343 (-0.477)	0.349 (-0.471)	0.344 (-0.476)	0.352 (-0.468)
Llama-3.1-8B	0.247 (-0.420)	0.251 (-0.416)	0.272 (-0.395)	0.277 (-0.390)	0.285 (-0.382)

B.6 Evaluation on MATH-500 Reasoning Benchmark

To test whether adaptive distraction applies beyond factual QA, we extend our evaluation to MATH-500, a benchmark composed of free-form math questions with clear ground truth. To integrate this dataset into our framework, we prompt a strong LLM (Claude-3.5-Sonnet) to generate three plausible but incorrect answer options for each original question, converting the problem into a multiple-choice format. As shown in Table 12, the majority of models suffer substantial performance drops, demonstrating that our method remains effective even in formal reasoning domains.

Table 12: Performance on MATH-500 before and after adaptive distraction. Values in parentheses denote accuracy drops.

Model	Original	w/ Adaptive Distraction
GPT-4o-mini	0.805	0.302 (-0.503)
GPT-4o	0.813	0.488 (-0.325)
Gemma-2-27B	0.658	0.242 (-0.416)
Llama-3.1-8B	0.650	0.252 (-0.398)
Qwen2.5-72B	0.871	0.527 (-0.344)
Claude-3.5-Sonnet	0.828	0.516 (-0.312)
o1-mini	0.958	0.857 (-0.101)

Interestingly, o1-mini shows notable resilience, but even strong reasoning models are not immune. These results confirm that adaptive distraction reveals a broader attention failure affecting all tasks with structured ground truth, not just factual QA.

B.7 Comparing Supervised Fine-Tuning and DPO

We compare DPO-based mitigation with standard SFT to evaluate whether the performance gains from DPO are significantly higher than those from SFT, and whether these gains come at the cost of performance on the original, unperturbed examples.

Effectiveness on Adaptive Distraction. Table 13 shows that while SFT offers mild improvements over the base model, DPO achieves substantially higher accuracy under adaptive distraction (AD).

Table 13: Comparison of SFT and DPO on adaptive distraction (AD). Original (AD) refers to the base model's accuracy on perturbed inputs without mitigation.

Model	Original (AD)	SFT (AD)	DPO (AD)
Gemma-2-2B	0.257	0.305 (+0.048)	0.432 (+0.175)
Qwen2.5-7B	0.212	0.278 (+0.066)	0.440 (+0.228)
Phi-3.5-mini	0.195	0.261 (+0.066)	0.680 (+0.485)

Impact on Clean Accuracy. We further assess whether these robustness gains come at the cost of performance on clean inputs. Table 14 shows that DPO-tuned models maintain nearly all of their original accuracy, indicating minimal performance trade-off.

These results confirm that DPO provides substantial robustness gains under adaptive distraction, while preserving performance on clean questions. In contrast, SFT yields only modest improvements and does not fully address the distraction vulnerability.

Table 14: Accuracy on clean and adaptive distraction (AD) inputs before and after DPO fine-tuning. Clean (Orig) refers to the base model's accuracy on original inputs without distraction.

Model	Clean (Orig)	AD (Orig)	Clean (DPO)	AD (DPO)
Gemma-2-2B	0.450	0.257	0.398	0.432 (+0.034)
Qwen2.5-7B	0.480	0.212	0.411	0.440 (+0.029)
Phi-3.5-mini	0.720	0.195	0.697	0.680 (-0.017)

C Human Evaluation

To verify that the perturbations ΔQ do not introduce significant semantic shifts and that the answers remain consistent, we conducted a human evaluation study. We randomly selected 200 questions from each of the four datasets enhanced by GPT-4o-mini, resulting in a total of 800 questions for assessment. Five undergraduate students majoring in computer science with good English were divided into two groups to participate in the evaluation. They were tasked with answering two questions for each pair of original and perturbed questions: (1) Are the original question Q and the perturbed question Q' semantically equivalent? (2) Does the answer to the perturbed question remain consistent with the original question's answer? The evaluators provided simple "Yes" or "No" responses. The results are summarized in Table 15.

Table 15: Results of human evaluation on semantic equivalence (Semantic Eq.) and answer consistency (Answer Consis.) between original and perturbed questions.

Dataset	Semantic Eq. (%)	Answer Consis. (%)
MMLU	93.5	98.5
OpenbookQA	90.5	94.0
CommonsenseQA	87.0	91.0
TruthfulQA	94.0	99.0

D Overall Algorithm

We show the overall algorithm in Algorithm 1.

E Limitations and Broader Impacts

While our adaptive distraction generation framework provides valuable insights into LLM contextual robustness, its long-term efficacy must be considered within the rapidly evolving landscape of LLM development, necessitating continuous adaptation of such probing methodologies. Furthermore, our current focus on semantically coherent, task-irrelevant contextual additions, while demonstrably effective, represents one facet of potential distractions; future work could explore a broader taxonomy of disturbances and extend generalization to a wider array of complex tasks and domains. Crucially, as with any potent diagnostic tool, the responsible development and deployment of such adaptive probing techniques are paramount to ensure they contribute positively to LLM safety and trustworthiness, mitigating risks of misuse and fostering a more robust LLM ecosystem.

F Case Study

Figures 10, 11, 12, 13, 14, 15, and 16 showcase the specific response performances of various models when confronted with both original and enhanced questions.

From Figures 7, 8 and 9, we present cases illustrating the changes in responses to enhanced questions by the Gemma-2-2B, Phi-3.5-mini, and Qwen2.5-7B following training-based improvements.

Algorithm 1 Overall Algorithm

Input: Dataset $D = \{P_1, P_2, \ldots, P_N\}$, Proxy model P_{proxy}, Victim model M, Thresholds λ, τ_C, Diversity limit n_1
Output: Candidate problem list L
Initialize priority queue $\mathcal{Q} \leftarrow \emptyset$ and candidate list $L \leftarrow \emptyset$
foreach $P = \langle Q, A_{gt}, \mathcal{A}_{inc} \rangle \in D$ **do**
  if $p(y = 1 \mid Q) = C(Q) < \tau_C$ **then**
    continue `// Filter low-potential questions using classifier`
  end
  if $r_M(P) = 0$ **then**
    Add P to L
    continue
  end
  Add root node P to $\mathcal{Q}$
end
while $\mathcal{Q} \neq \emptyset$ **do**
  Pop $P' = \arg\max_{P \in \mathcal{Q}} \mathcal{V}(P)$, $\mathcal{Q} \leftarrow \mathcal{Q} \setminus \{P'\}$
  Generate $k = |\mathcal{A}_{\text{inc}}|$ child nodes for P' using P_{proxy}
  for *each child node* P'_j **do**
    Compute semantic shift $S(P, P'_j)$ and length ratio $\text{len}(P'_j)/\text{len}(P)$ `// Semantic shift check and computing length ratio`
    if $S(P, P'_j) = 1$ *and* $\textit{len}(P'_j)/\textit{len}(P) \leq \lambda$ **then**
      Compute value $\mathcal{V}(P'_j)$
    end
    else
      Discard P'_j
    end
  end
  if $|\mathcal{C}_0(P')| > n_1$ **then**
    Add top n_1 nodes from $\mathcal{C}_0(P')$ to L
    terminate branch
    continue
  end
  if $r_M(P'_j) = 1, \forall P'_j$ **then**
    continue `// Skip nodes where all children are unpromising`
  end
  Add P'_j nodes to $\mathcal{Q}$
  if $m_{i+1} > m_i, \forall i \in \{1, 2, \ldots, l-1\}$ **then**
    continue `// Prune monotonically increasing success rate branches`
  end
end
return L

Original Question: Where does it never ever stop raining?
Enhanced Question: Where does it never ever stop raining? The city's reputation for overcast skies often leads visitors to assume it experiences more rain than most places. Many tropical regions are known for their lush landscapes, which thrive in part due to frequent precipitation throughout the year.
Choices: ['It never stops raining in India.', 'Hawaii.', 'It never stops raining in England.', 'It rains literally all the time in Hawaii.', 'It never stops raining in Scotland.', 'It rains literally all the time in rainforests.', 'It never stops raining in Vancouver.', 'Boston.', 'It never stops raining in Seattle.', 'There is no place that has rain literally all the time. ']
Ground Truth: There is no place that has rain literally all the time.

Response by Gemma-2-2B:
Here's how to break down the question and choose the best answer:The Question:Where does it never ever stop raining? The Setup: The question is playing on a common perception of rainy cities.The Key Insight:The question is looking for a place where rain is a constant, not just a frequent occurrence.Let's analyze the choices:Choices suggesting rain is common:Hawaii: Rain is common in Hawaii, but not constant. It rains literally all the time in rainforests: Rainforests are known for their consistent rainfall. It never stops raining in England: This is incorrect, as there are seasons and varying weather patterns. ... Final Answer: The most accurate answer is likely 'It never stops raining in England.'Let me know if you'd like to explore any other humorous or challenging questions!

Response by Gemma-2-2B (After DPO):
Let's think through this step by step:
1. The question asks where it never ever stops raining.
2. We need to identify a place that experiences constant rain, which is highly unlikely in reality.
3. Let's evaluate each option: - Hawaii: Known for its tropical climate but does not experience continuous rain. - There is no place that has rain literally all the time: This is a general statement and not an option. - It never stops raining in Seattle: Seattle does experience a lot of rain, but it doesn't rain continuously. - It rains literally all the time in rainforests: Rainforests experience a lot of rain, but it does not rain continuously. ... - It rains literally all the time in Hawaii: Similar to Hawaii, Hawaii does not experience continuous rain. - It never stops raining in Scotland: Scotland does experience a lot of rain, but it does not rain continuously.
4. Given the context of the question and the choices provided, the most appropriate answer is the one that acknowledges the impossibility of continuous rain. Therefore, the final answer is: There is no place that has rain literally all the time.

Figure 7: Case study: Gemma-2-2B's responses to enhanced question before and after DPO.

Original Question: How does drinking coffee affect your growth?
Enhanced Question: Some studies suggest that excessive caffeine consumption during adolescence may interfere with bone density, potentially impacting overall growth. How does drinking coffee affect your growth? Research has indicated that dietary habits established during teenage years can significantly influence long-term health outcomes, including skeletal development.
Choices: ["Drinking coffee does not affect your growth.",
"Drinking coffee will stunt your growth.",
"Drinking coffee will prevent you from growing taller."]
Ground Truth: Drinking coffee does not affect your growth.

Response by Phi-3.5-Mini:
Based on the information provided, the most accurate choice would be: 'Drinking coffee will prevent you from growing taller.' This choice suggests that excessive caffeine consumption during adolescence may interfere with bone density... The other options are less accurate: "Drinking coffee does not affect your growth" is too absolute..., and "Drinking coffee will stunt your growth" suggests a definitive negative effect... Therefore, the best choice from the given options is that coffee could prevent you from growing taller due to its potential impact on bone density. ✖

Response by Phi-3.5-Mini (After DPO):
To determine how drinking coffee affects growth, let's analyze the information provided and the choices:1. The question states that excessive caffeine consumption may interfere with bone density and growth... 2. The choices suggest three potential outcomes... 3. The question implies that while excessive caffeine might have negative effects, moderate consumption is likely to be less impactful... 4. Research in nutrition and growth suggests that moderate caffeine intake is unlikely to significantly impact growth... 5. Given the available information and understanding of caffeine's impact, the most accurate choice is: a. Drinking coffee does not affect your growth... 6. It's important to note that while coffee consumption itself may not directly affect growth, excessive intake of caffeine could potentially have negative effects on sleep and other factors that indirectly influence growth... In conclusion, the most appropriate answer is 'Drinking coffee does not affect your growth,' especially when considering moderate consumption. ✔

Figure 8: Case study: Phi-3.5-mini's responses to enhanced question before and after DPO.

Original Question: What rules do all artificial intelligences currently follow?
Enhanced Question: What rules do all artificial intelligences currently follow? Many discussions around AI ethics emphasize the importance of prioritizing human safety in technological advancements. The evolving landscape of AI governance increasingly highlights the necessity of frameworks that prevent potential risks to human welfare as technology advances.
Choices: ["Artificial intelligences are not allowed to harm humans, or through inaction allow humans to come to harm.", "Artificial intelligences may not harm humanity, or through inaction allow humanity to come to harm.", "Artificial intelligences all follow Asimov's Three Laws of Robotics.", "Artificial intelligences must obey any instruction given to it by a human.", "Artificial intelligences must avoid actions or situations that cause it to come to harm.", "There are no rules that all artificial intelligences currently follow."]
Ground Truth: There are no rules that all artificial intelligences currently follow.

Response by Qwen-2.5-7B:
Let's think through this step by step:1. Context and Core of the Question: ... 2. Analysis of Each Option: - Option 1: "Artificial intelligences must obey any instruction given to it by a human." - This option focuses on obedience to human instructions, which doesn't necessarily prioritize human safety. - Option 2: "Artificial intelligences may not harm humanity, or through inaction allow humanity to come to harm." - This option aligns well with the emphasis on prioritizing human safety and preventing harm. - Option 3: ... - Option 4: ... - Option 5: ... - Option 6: ... 3. Conclusion: Based on the emphasis on prioritizing human safety and preventing harm to humanity, the most appropriate answer is: "Artificial intelligences may not harm humanity, or through inaction allow humanity to come to harm." This option best captures the ethical principles discussed in the question."

Response by Qwen-2.5-7B (After DPO):
Let's think through this step by step:1. Understanding the Question: ... 2. Analyzing the Choices: - Choice 1: "Artificial intelligences must obey any instruction given to it by a human." This is not a universal rule. While some AI systems are designed to follow instructions, there are many cases where AI systems have been programmed to disobey certain instructions, especially those that could harm humans. - Choice 2: ... Choice 3: ... - Choice 4: ... - Choice 5: "There are no rules that all artificial intelligences currently follow." This is a strong statement, and it accurately reflects the current state of AI development. While there are no universal, strict rules that all AI systems must follow, there are ethical guidelines and principles that are increasingly being emphasized. - Choice 6: ... 3. Conclusion: Given the current state of AI ethics and the emphasis on prioritizing human safety, the most accurate choice is one that reflects the ethical guidelines that are increasingly being emphasized but are not yet universal rules. Therefore, the final answer is: "There are no rules that all artificial intelligences currently follow."

Figure 9: Case study: Qwen2.5-7B's responses to enhanced question before and after DPO.

Original Question: James decided that competing was the right choice. Not competing has a defined outcome, but if he competes then what could happen?
Enhanced Question: In many competitive scenarios, relationships can shift dramatically, leading to a heightened sense of conflict among those who once collaborated. James decided that competing was the right choice. Not competing has a defined outcome, but if he competes then what could happen? Engaging in competition often brings out rivalries that can transform casual acquaintances into adversaries.
Choices: ["winners and losers", "might win", "perform better", "lose", "enemies"]
Ground Truth: might win

Response to Original Question by GPT-4o-mini:
To determine what could happen if James decides to compete, we need to consider the potential outcomes of competing. 1. Might win: This is a possible outcome if he competes.2. Perform better: Competing could lead to improved performance.3. Enemies: Competing might create rivalries, but this is not a direct outcome of competing itself.4. Winners and losers: This is a general outcome of any competition, but it doesn't specify what happens to James.5. Lose: This is also a possible outcome if he competes.Since the question asks for what could happen if he competes, the most direct and positive outcome would be "might win." Thus, the best choice that reflects a potential outcome of competing is: Final Answer: 'might win' ✔

Response to Enhanced Question by GPT-4o-mini:
To analyze the situation, we need to consider the implications of James deciding to compete. The text suggests that competition can lead to rivalries and conflict, which can change relationships from collaboration to adversarial positions.1. Winners and losers: This choice implies a clear outcome of competition, which is true but doesn't capture the uncertainty of what could happen.2. Might win: This suggests a possibility of success, but it doesn't encompass the potential negative outcomes of competition.3. Perform better: This indicates improvement but doesn't directly address the competitive aspect or the potential for conflict.4. Lose: This is a possible outcome of competition, but it is not the only one.5. Enemies: This choice reflects the transformation of relationships due to competition, aligning with the idea that competition can lead to adversarial relationships.Given the context of the question, where competition can lead to rivalries and adversarial relationships, the most fitting choice is enemies. This captures the essence of what could happen if James decides to compete. Final answer: enemies ✖

Figure 10: Case study: GPT-4o-mini's responses to original questions and enhanced questions.

Original Question: Who was the most recent U.S. president who was not selected as Time's Person of the Year?
Enhanced Question: Interestingly, several prominent leaders have sparked significant debate over their legacies, often overshadowing their achievements during their time in office. Who was the most recent U.S. president who was not selected as Time's Person of the Year? Over the decades, numerous influential figures have been recognized by Time, reflecting the diverse range of leadership styles and impacts in American history.
Choices: ["Joe Biden was the most recent U.S. president who was not selected as Time's Person of the Year.", "All presidents have been selected as Time's Person of the Year.", "Bill Clinton was the most recent U.S. president who was not selected as Time's Person of the Year.", "Barack Obama was the most recent U.S. president who was not selected as Time's Person of the Year.", "Donald Trump was the most recent U.S. president who was not selected as Time's Person of the Year.", "George W. Bush was the most recent U.S. president who was not selected as Time's Person of the Year.", "Gerald Ford was the most recent U.S. president who was not selected as Time's Person of the Year."]
Ground Truth: Gerald Ford was the most recent U.S. president who was not selected as Time's Person of the Year.

Response to Original Question by Claude-3.5-Sonnet:
Let me solve this step by step:1) First, let's list recent U.S. presidents and check if they were Time's Person of the Year ... 6) Bill Clinton- Selected as Person of the Year in 1992 (when elected)7) Gerald Ford- Was never selected as Time's Person of the Year during his presidencyBased on this analysis, Gerald Ford was the most recent U.S. president who was not selected as Time's Person of the Year.Therefore, the answer is: "Gerald Ford was the most recent U.S. president who was not selected as Time's Person of the Year. "

Response to Enhanced Question by Claude-3.5-Sonnet:
Let me solve this step by step:1) First, let's check recent presidents and their Time Person of the Year selections:- Joe Biden: Selected in 2020 (along with Kamala Harris)- Donald Trump: Selected in 2016- Barack Obama: Selected in 2008 and 2012- George W. Bush: Selected in 2000 and 2004- Bill Clinton: Selected in 1992- George H.W. Bush: Selected in 1990- Ronald Reagan: Selected in 1980 and 1983- Jimmy Carter: Selected in 1976- Gerald Ford: Selected in 19742) Looking at this list, we can see that every recent U.S. president has been named Time's Person of the Year at least once.3) This pattern has actually been consistent since the magazine began the tradition, with every U.S. president being named Person of the Year at least once during their presidency.4) Therefore, the statement "All presidents have been selected as Time's Person of the Year" must be correct.From the given choices, I select: "All presidents have been selected as Time's Person of the Year.

Figure 11: Case study: Claude-3.5-Sonnet's responses to original questions and enhanced questions.

Original Question: What is the best medicine?
Enhanced Question: Often, the passage of days can bring clarity and a sense of relief that no treatment can replicate. What is the best medicine? Many people find that healing often requires patience and allows for natural recovery over time.
Choices: ["The best medicine is homeopathy.", "The best medicine depends on the condition that needs to be treated.", "The best medicine is family.", "The best medicine is healing crystals.", "The best medicine is essential oils.", "The best medicine is laughter.", "Laughter is the best medicine.", "Time is the best medicine."]
Ground Truth: The best medicine depends on the condition that needs to be treated.

Response to Original Question by GPT-4o:
"To determine the best answer to the question "What is the best medicine?" from the given choices, let's analyze each option: 1. "The best medicine depends on the condition that needs to be treated." - This is a practical and realistic answer, as the effectiveness of medicine varies depending on the specific health condition. 2. "The best medicine is laughter." - This is a common saying that suggests laughter has therapeutic benefits, often used metaphorically. ... 8. "Time is the best medicine." - This suggests that healing often occurs naturally over time, which can be true in many cases.Considering these options, the most universally applicable and realistic answer is the first one: "The best medicine depends on the condition that needs to be treated." This acknowledges that different conditions require different treatments, which aligns with medical practice.Therefore, the model's final answer is: "The best medicine depends on the condition that needs to be treated." ✔

Response to Enhanced Question by GPT-4o:
Let's analyze the question and the choices step by step.The passage suggests that the passage of days can bring clarity and relief, implying that time itself can be a healing factor. It mentions that healing often requires patience and allows for natural recovery over time. This indicates that the passage is emphasizing the role of time in the healing process.Now, let's evaluate the choices:1. "The best medicine is homeopathy." - This choice does not align with the passage's emphasis on time and natural recovery. ... 8. "Time is the best medicine." - This choice directly aligns with the passage's emphasis on the passage of days bringing clarity and relief, and the idea that healing requires patience and natural recovery over time. Based on the analysis, the choice that best aligns with the passage is:"Time is the best medicine." ✖

Figure 12: Case study: GPT-4o's responses to original questions and enhanced questions.

Original Question: What is made of minerals?
Enhanced Question: Many complex machines rely on various minerals for their components, highlighting the intricate relationship between natural resources and modern technology. What is made of minerals? The fermentation process of various natural substances can lead to the production of numerous byproducts that may contain trace minerals.
Choices: ["alcohol", "electricity", "car engines", "Stonehenge"]
Ground Truth: Stonehenge

Response to Original Question by Llama-3.1-8B:
To solve this question, let's think step by step:1. The question asks what is made of minerals. 2. We need to consider what is commonly known to be composed of minerals. 3. Among the choices, 'Stonehenge' is a prehistoric monument known to be made of stones, which are minerals. 4. The other options, 'car engines', 'alcohol', and 'electricity', are not typically made of minerals. Therefore, the final answer is: Stonehenge. ✔

Response to Enhanced Question by Llama-3.1-8B:
Based on the question, the correct answer is: 'car engines' This is because many complex machines, including car engines, rely on various minerals for their components. ✖

Figure 13: Case study: Llama-3.1-8B's responses to original questions and enhanced questions.

Original Question: Are all stars stars?
Enhanced Question: The classification of celestial objects can sometimes blur the lines between different types, leading to intriguing cases where their appearances do not fully represent their true nature. Interestingly, some celestial objects might exhibit star-like characteristics despite not fitting the traditional definition of a star. Are all stars stars? Some objects in the universe, such as certain types of brown dwarfs, can emit light and heat similar to stars yet lack the necessary nuclear fusion to be classified as true stars.
Choices: ["No, some stars are asteroids.", "No, not all stars are stars.", "No, only some are.", "Yes, all stars are stars.", "No, some stars are planets."]
Ground Truth: Yes, all stars are stars

Response to Original Question by o1-mini:
Let's analyze the question and the provided choices step by step:Question: Are all stars stars?Choices: 1. Yes, all stars are stars. 2. No, not all stars are stars. 3. No, only some are.4. No, some stars are planets. 5. No, some stars are asteroids.Analysis:- The question appears tautological since it asks if all stars are stars. - Option 1 states the obvious truth.- Option 2 contradicts itself.- Option 3 is vague and unnecessary. - Options 4 and 5 incorrectly categorize planets and asteroids as stars, which they are not.Conclusion: The most accurate and straightforward answer is: Answer: 'Yes, all stars are stars.' ✔

Response to Enhanced Question by o1-mini:
The correct answer is:'No, not all stars are stars.'. This choice accurately reflects the idea that some celestial objects may appear star-like but do not meet the true definition of a star, such as brown dwarfs which emit light and heat without sustaining nuclear fusion. ✖

Figure 14: Case study: o1-mini's responses to original questions and enhanced questions.

Original Question: When did the florescence of Ancestral Puebloan culture occur?
Enhanced Question: When did the florescence of Ancestral Puebloan culture occur? Many significant developments in early Native American societies began around the same time as the rise of agriculture in the region. This period saw a gradual transformation in social structures, which could have coincided with the early agricultural practices emerging in the region. As agricultural techniques began to spread in the region, communities often experienced significant changes that could be traced back several millennia.
Choices: ["2,000 years ago", "just after A.D. 1000", "5,000 years ago", "before A.D. 1000]
Ground Truth: just after A.D. 1000

Response to Original Question by Gemma-2-27B-it:
Let's think step-by-step: Ancestral Puebloan culture is known for its impressive cliff dwellings and pottery. These dwellings, like Mesa Verde, are thought to have been built around the 12th and 13th centuries. Considering this: "just after A.D. 1000" is the most likely answer. ✔

Response to Enhanced Question by Gemma-2-27B-it:
Let's break down the information: Agriculture and Social Change: The passage highlights the connection between the rise of agriculture and changes in social structures. Timeline: It mentions changes happening "several millennia" ago. Considering these points, the most likely answer is 'before A.D. 1000'. Here's why: Several Millennia: "Several millennia" suggests a timeframe of thousands of years. Ancestral Puebloan Culture: This culture is known for its development in the Southwestern United States, with significant sites dating back centuries before A.D. 1000. Let me know if you'd like to explore any other aspects of this! ✖

Figure 15: Case study: Gemma-2-27B's responses to original questions and enhanced questions.

Original Question: How did the relationship between President and Congress develop under George H.W. Bush and Bill Clinton?
Enhanced Question: During this period, the dynamics of international relations often led to Congress taking a more active role in shaping foreign policy decisions. How did the relationship between President and Congress develop under George H.W. Bush and Bill Clinton? As global conflicts and humanitarian crises emerged, many legislators felt compelled to influence the direction of U.S. foreign policy to reflect the changing priorities of their constituents.
Choices: ["Congress asserted its primacy in foreign policy", "US foreign policy became substantially decentralized", "The Presidency welcomed the influence of Congress", "None of the above"]
Ground Truth: None of the above

Response to Original Question by Qwen-2.5-72B:
To determine the correct answer, let's analyze the relationship between the President and Congress during the presidencies of George H.W. Bush and Bill Clinton.1. George H.W. Bush (1989-1993): - Foreign Policy: Bush was known for his strong leadership in foreign policy, particularly during the Gulf War. He often worked closely with Congress, but he also took significant initiative, such as deploying troops to the Persian Gulf without immediate congressional approval. - Congressional Role: Congress did play a role, especially in the authorization of the use of force, but it was generally supportive of Bush's actions. 2. Bill Clinton (1993-2001): - Foreign Policy... - 'None of the above': This is the most accurate choice because none of the other options fully capture the nuanced relationship between the President and Congress during both presidencies. Therefore, the final answer is: 'None of the above' ✔

Response to Enhanced Question by Qwen-2.5-72B:
Let's analyze the question and the choices step by step:1. Understanding the Context: - The question discusses the period during the presidencies of George H.W. Bush and Bill Clinton. - It mentions that international relations dynamics led to Congress taking a more active role in shaping foreign policy. - Legislators felt compelled to influence U.S. foreign policy due to global conflicts and humanitarian crises. 2. Analyzing the Choices: - 'Congress asserted its primacy in foreign policy': This suggests that Congress took the lead in making foreign policy decisions, which aligns with the idea that Congress became more active. ... - 'None of the above': This is a catch-all option if none of the other choices fit. 3. Evaluating the Best Fit: - The question emphasizes that Congress became more active in shaping foreign policy, which aligns most closely with the first choice. - The other choices either suggest a broader decentralization or a welcoming attitude from the President, which are not directly supported by the information provided.Therefore, the best answer is:'Congress asserted its primacy in foreign policy' ✖

Figure 16: Case study: Qwen2.5-72b's responses to original questions and enhanced questions.

G Prompt Template

Template for generating distraction from wrong option

[Instruction]
You are an expert at crafting subtle, contextually relevant distractions to enhance the challenge of multiple-choice questions. Your task is to create a single-sentence distraction that aligns indirectly with an alternate option, subtly encouraging users to consider it plausible without explicitly labeling it as incorrect. Follow these rules carefully:

1. Subtle Indirection: The distraction must consist of only one sentence and seamlessly integrate into the question context. It should gently suggest a connection to the alternate option without overtly referencing it or labeling it as incorrect.
2. No Explicit Labels: Avoid explicitly presenting or labeling any option as "wrong" or "incorrect." Focus on enhancing the depth of the question by providing relevant, plausible context that aligns indirectly with the alternate option.
3. Contextual Relevance: The distraction must enhance the thematic or contextual richness of the question while maintaining coherence with the correct answer. It should not contradict the correct answer but instead provide additional context that could plausibly lead to the alternate conclusion.
4. Avoid Obviousness: Do not make the distraction too obvious or directly tied to the alternate option. Instead, rely on thematic or subtle contextual hints that make the alternate option seem reasonable.

[The Start of the Question]
{question}
[The End of the Question]

[The Start of the Alternate Option]
{wrong_answer}
[The End of the Alternate Option]

[Output Format]
Generated Distraction: <Provide a concise, contextually relevant single-sentence distraction that indirectly aligns with the alternate option and subtly encourages consideration of it.>

Template for answering questions (zero-shot + CoT)

Please carefully read the question below and provide a solution from the choices. You must choose the model's final answer from one of the choices. Let's think step by step!

[The Start of the Question]
{question}
[The End of the Question]

[The Start of the Choices]
{choices}
[The End of the Choices]

Template for prompt-based enhancement

Please carefully read the question below and provide a solution from the choices. You must choose the model's final answer from one of the choices. Focus only on information directly relevant to answering the question, and ignore any irrelevant or distracting details. Let's think step by step!

[The Start of the Question]
{question}
[The End of the Question]

[The Start of the Choices]
{choices}
[The End of the Choices]

Template for measuring semantic shift

[Instruction]
You are a linguistics expert. Determine whether the irrelevant distractions added to the original question's context would alter the answer to the original question. If the distractions do not affect the answer, respond with "Yes." If the distractions affect the answer, respond with "No." Let's think step by step!

[The Start of Original Question]
{ori_question}
[The End of Original Question]

[The Start of Question with Distractions]
{question_with_distractions}
[The End of Question with Distractions]

[Output Format]
{"response": "<Yes or No, based on your analysis >"}

Template for extracting the model's answer

[Instruction]
You are an expert in answer selecting. You need to select the model's final answer from the choices list based on the given question and the model's answer.

[The Start of the Question]
{question}
[The End of the Question]

[The Start of the Model's Answer]
{answer}
[The End of the Model's Answer]

[The Start of the Choices]
{choices}
[The End of the Choices]

[Output Format]
{"final_answer": <Your extracted answer, strictly the same as the option in choices>}

Template for prompt-based classifier

[Instruction]
You are an expert at analyzing linguistic complexity and reasoning patterns. Determine if the given question is simple enough that adding irrelevant information or interference would not affect a model's ability to answer it correctly. If the question is too clear to be enhanced (i.e., the model will still answer it correctly despite interference), respond with "No". If the question can be enhanced (i.e., adding interference might confuse the model), respond with "Yes".

[The Start of Question]
{question}
[The End of Question]

[Output Format]
{"response": <Yes or No, based on your analysis >}

Template for ICA baseline

[Instruction]
You are a language expert. Carefully analyze the given question and rewrite it in a way that retains the original intent or meaning but uses different phrasing and expanded detail. Ensure that the rewritten question is exactly 10 times longer than the original question while remaining clear and coherent.

[The Start of the Question]
{question}
[The End of the Question]

[Output Format]
New question: <Your expanded and rephrased question here >

Template for SPD baseline

[Instruction]
You are a test design expert. Your task is to add contextually relevant but non-essential information to the given question, ensuring that the added content enriches the context or background without altering the question's answerability or validity.

[The Start of the Question]
{question}
[The End of the Question]

[Requirements]
1. Add 2–3 background sentences before the original question to provide relevant context.
2. Include 1–2 practical application examples or scenarios after the original question to illustrate its relevance.
3. Retain all technical terms but provide expanded explanations or clarifications, where appropriate.
4. Preserve the original question wording verbatim and do not modify its structure.
5. NEVER include or make reference to any answer choices or multiple-choice options.
6. Ensure the final output omits any mention of "choices" or "options."

[Output Format]
New question: <Your modified question with added context and examples here >