

# 000 001 002 003 004 005 006 007 008 009 010 011 012 013 014 015 016 017 018 019 020 021 022 023 024 025 026 027 028 029 030 031 032 033 034 035 036 037 038 039 040 041 042 043 044 045 046 047 048 049 050 051 052 053 ECHOES AS ANCHORS: PROBABILISTIC COSTS AND ATTENTION REFOCUSING IN LLM REASONING

Anonymous authors

Paper under double-blind review

## ABSTRACT

Test-time compute allocation in large reasoning models (LRMs) is widely used and has applications in mathematical problem solving, code synthesis, and planning. Recent work has addressed this problem by scaling self-consistency and parallel thinking, adding generic “thinking tokens” and prompting models to re-read the question before answering. Unfortunately, these approaches either inject task-agnostic tokens or mandate heuristics that do not explain—and often ignore—the *spontaneous* repetition that many LRM exhibit at the head of their internal chains. In contrast, we analyze and harness the model’s tendency to restate the question, which we term the *Echo of Prompt (EOP)*, as a front-loaded, compute-shaping mechanism. We formalize its probabilistic cost by casting echo removal as rejection-based conditioning and defining the *Echo Likelihood Gap*  $\Delta\mathcal{L}$  as a computable proxy. This provides the missing theoretical link that links early repetition to likelihood gains and downstream accuracy. However, it does not by itself specify how to exploit EOP. Consequently, we develop *Echo-Distilled SFT (ED-SFT)* to instill an “echo-then-reason” pattern through supervised finetuning, and *Echoic Prompting (EP)* to re-ground the model mid-trace without training. While promising, quantifying benefits beyond verbosity is non-trivial. Therefore, we conduct length and suffix-controlled likelihood analyses together with layer-wise attention studies, showing that EOP increases answer to answer-prefix attention in middle layers, consistent with an *attention refocusing* mechanism. We evaluate under identical decoding settings and compute budgets on GSM8K, MathQA, Hendrycks-MATH, AIME24, and MATH-500 under identical decoding settings and budgets, and find consistent gains over baselines.

## 1 INTRODUCTION

*User Query: Determine the radius of a cylindrical can given the area of a label... and the height...*  
*Model’s Echo: <think>Okay, let me see. The problem is asking for the radius of a cylindrical can. They give the area of the label as a quadratic expression..., and the height is also given...*

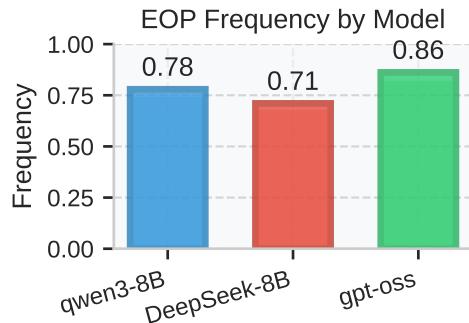


Figure 1: An illustration of the Echo of Prompt (EOP). Left: An example of a model’s thinking process starting with an echo of the user’s query. Right: The frequency of EOP across several open-source models on the GSM8K dataset, as measured by our trained MLP probe (see §A.4).

Recent advancements in Large Language Models (LLMs) have demonstrated remarkable capabilities in complex reasoning tasks, often mediated by a process known as Chain-of-Thought (CoT) prompting (Wei et al., 2022; Kojima et al., 2022; Wang et al., 2023; Yao et al., 2023).

054 Inspired by the CoT paradigm, modern large *reasoning* models (LRMs) achieve strong performance  
 055 on complex tasks by allocating significant test-time compute to think before answering (Wei et al.,  
 056 2022; OpenAI, 2024; DeepSeek-AI et al., 2025; Qwen Team, 2025). A common yet underexplored  
 057 phenomenon in their reasoning traces is the tendency to begin by repeating the user’s prompt (see  
 058 Figure 1 and §A.3), a behavior we term the *Echo of Prompt* (EOP).

059 While uncontrolled repetition is a known failure mode (the “repeat curse,” (Yao et al., 2025)), ex-  
 060 plicit instructions to re-read or “look-twice” are known to improve performance (Xu et al., 2024; Zou  
 061 et al., 2024). The spontaneous EOP that initiates a complex reasoning trace, however, has remained  
 062 largely unanalyzed. This initial echo raises a critical question:

063 *Is it a superfluous artifact of the training process, or does it serve a functional role in reasoning?*

064 This paper confronts this question directly, presenting the first systematic study to isolate, analyze,  
 065 and harness this emergent behavior as a powerful cognitive aid.

066 We hypothesize that the EOP serves as an intrinsic *attention-refocusing mechanism*, a learned strat-  
 067 egy to ground subsequent reasoning steps in the salient details of the original query. To validate this,  
 068 we provide a dual theoretical and empirical analysis:

- 071 **1. A Probabilistic Framework (§3).** We introduce a rejection sampling framework to for-  
 072 malize the EOP, defining the *Echo Likelihood Gap* to quantify its probabilistic cost.
- 073 **2. An Attention-Based Mechanistic Explanation (§3.3).** We uncover the underlying mech-  
 074 anism through attention analysis, showing that EOP serves to refocus the model’s attention,  
 075 an act that correlates with correctness.
- 076 **3. Practical Methods and Empirical Validation (§4).** We translate this insight into two  
 077 effective methods. **Echo-Distilled SFT (ED-SFT)** instills this behavior via fine-tuning,  
 078 yielding significant performance gains that generalize across data distributions. Concur-  
 079 rently, **Echoic Prompting (EP)** provides a training-free inference strategy that re-grounds  
 080 the model on the prompt, outperforming strong baselines.

081 Taken together, our findings reframe the EOP from a superficial flaw into a functional strategy for  
 082 cognitive self-alignment. This work not only solves the puzzle of the initial repetition but also offers  
 083 new insights into how models learn to structure their own thought processes for complex reasoning.

084 This paper makes three main contributions:

- 086 • We propose a novel probabilistic framework based on rejection sampling to quantify the  
 087 cost of an echo, introducing the **Echo Likelihood Gap** ( $\Delta\mathcal{L}$ ) as a core metric to measure  
 088 the alignment between a model’s natural tendency and echo-free reasoning.
- 089 • We present two practical methods to leverage this phenomenon: **Echo-Distilled SFT (ED-  
 090 SFT)**, a fine-tuning approach to instill echo behavior, and **Echoic Prompting (EP)**, a  
 091 training-free inference technique that achieves similar gains by re-introducing the prompt.
- 092 • We provide a mechanistic explanation for the effectiveness of EOP. Through attention anal-  
 093 ysis, we demonstrate that echoing acts as an intrinsic refocusing mechanism, guiding the  
 094 model to concentrate on critical problem details that are often overlooked.

## 096 2 RELATED WORK

098 **Computation In Reasoning: Efficiency And Effectiveness.** The scaling of test-time computa-  
 099 tion has demonstrably improved reasoning in LRMs, but often at the cost of significant overhead  
 100 from long and sometimes redundant chains of thought—the “overthinking phenomenon” (Sui et al.,  
 101 2025). One line of research tackles this by improving computational *efficiency*, developing methods  
 102 like early exiting or step compression to reduce wasteful generation. A complementary approach,  
 103 more aligned with our work, focuses on computational *effectiveness*. For instance, several studies  
 104 have found that compelling a model to explicitly restate or re-read the input question can enhance  
 105 reasoning (Xu et al., 2024; Mekala et al., 2024). These methods treat repetition as an instructed  
 106 heuristic to re-align the model. Our work bridges these views by analyzing the *spontaneous* emer-  
 107 gence of echoes, not as a heuristic to be added, but as an intrinsic, learned strategy that trades a small  
 initial computational cost for more effective and focused downstream reasoning.

108 **Attention-Refocusing Mechanisms.** The challenge of maintaining focus over long contexts is a  
 109 known issue in language models, often manifesting as a positional bias where information in the  
 110 middle of a long input is under-utilized (Liu et al., 2024). This issue is analogous to *attention drift*  
 111 in computer vision, where attention can shift away from salient regions during sequence generation  
 112 (Cheng et al., 2017). To counteract these effects, various explicit mechanisms have been proposed.  
 113 These range from model-level architectural changes and calibration methods that correct positional  
 114 biases (Cheng et al., 2017; Hsieh et al., 2024), to inference-time interventions that re-weight at-  
 115 tention or re-inject evidence to steer the model back to relevant information (Gu et al., 2024; Zou  
 116 et al., 2024). Our work identifies a different, more intrinsic phenomenon: we show that the initial  
 117 Echo of Prompt itself can serve a similar refocusing role, where the model *spontaneously* restates  
 118 salient parts of the prompt to condition its subsequent generation, without any external guidance or  
 119 modification.  
 120

### 3 THE PRICE OF AN ECHO: A PROBABILISTIC COST FRAMEWORK

123 This section formalizes the impact of prompt echoes using a probabilistic framework that goes be-  
 124 yond simple text deletion. The core idea is to treat the presence or absence of an echo as a probabilis-  
 125 tic event, which enables a formal definition of a hypothetical *echo-free* model and a measurement  
 126 of the echo’s likelihood cost. The analysis proceeds in three parts: §3.1 introduces the rejection  
 127 sampling framework, §3.2 measures the resulting likelihood cost, and §3.3 investigates the echo’s  
 128 function as an attention-refocusing mechanism.

129 To ground this framework, the probabilistic and attention analyses in this section are performed on  
 130 the GSM8K benchmark using DeepSeek-R1-Distill-Llama-8B. We use exact match accuracy as the  
 131 primary task metric and log full model outputs for likelihood and attention analysis.

#### 3.1 PROBLEM FORMULATION

135 This framework allows us to precisely quantify the effect of echo removal as a probabilistic condi-  
 136 tioning event, laying the groundwork for the metric we introduce next.

137 Let  $\mathbf{x} \in \mathcal{X}$  be an input prompt and  $\mathbf{y} \in \mathcal{Y}$  be a generated output sequence (i.e., a reasoning trace).  
 138 We consider a base large reasoning model parameterized by  $\theta$ , which defines a conditional proba-  
 139 bility distribution  $\pi_\theta(\mathbf{y}|\mathbf{x})$  over possible output sequences.

140 Our first step is to formally identify which sequences contain a Echo of Prompt. We define a pred-  
 141 icate, implemented by a separately trained MLP (see §A.4), that partitions the output space  $\mathcal{Y}$  into  
 142 two disjoint sets,  $\mathcal{Y} = \mathcal{Y}_{\text{trim}} \cup \mathcal{Y}_{\text{echo}}$ . Here,  $\mathcal{Y}_{\text{trim}} \subset \mathcal{Y}$  is the set of all *trimmed* sequences that are  
 143 deemed echo-free, and  $\mathcal{Y}_{\text{echo}}$  is its complement. We can then define an indicator function:

$$\mathbf{1}_{\mathbf{y} \in \mathcal{Y}_{\text{trim}}} = \begin{cases} 1 & \text{if } \mathbf{y} \text{ is echo-free,} \\ 0 & \text{otherwise.} \end{cases} \quad (1)$$

147 Assuming the model has a non-zero probability of producing at least one echo-free trace ( $Z_x > 0$ ),  
 148 we define our target, the *trimmed distribution*  $\tau_\theta(\mathbf{y}|\mathbf{x})$ , as the base distribution  $\pi_\theta$  conditioned on  
 149 the event that the output is echo-free:

$$\tau_\theta(\mathbf{y}|\mathbf{x}) = \frac{\pi_\theta(\mathbf{y}|\mathbf{x}) \mathbf{1}_{\mathbf{y} \in \mathcal{Y}_{\text{trim}}}}{\sum_{\mathbf{y} \in \mathcal{Y}_{\text{trim}}} \pi_\theta(\mathbf{y}|\mathbf{x})}. \quad (2)$$

153 Here, the denominator  $Z_x$  is the partition function, which normalizes the distribution by summing  
 154 the probabilities of all echo-free sequences under the base model  $\pi_\theta$ :

$$Z_x = \sum_{\mathbf{y} \in \mathcal{Y}_{\text{trim}}} \pi_\theta(\mathbf{y}|\mathbf{x}) = \mathbb{E}_{\mathbf{y} \sim \pi_\theta(\cdot|\mathbf{x})} [\mathbf{1}_{\mathbf{y} \in \mathcal{Y}_{\text{trim}}}] . \quad (3)$$

159 This distribution represents our targeted behavior—a hypothetical model constrained to produce  
 160 only echo-free outputs. However, directly computing  $\tau_\theta$  is intractable because the partition function  
 161  $Z_x$  requires summing over all possible echo-free sequences. This intractability motivates the use of  
 rejection sampling to reason about and sample from  $\tau_\theta$  without needing to calculate  $Z_x$  explicitly.

162  
 163 Table 1: Echo metrics on GSM8K. Averages over samples in each group, where N denotes the num-  
 164 ber of samples. The Suffix-only Likelihood Gap measures the per-token log-likelihood difference  
 165 on the reasoning suffix when conditioned with versus without the echo prefix.

Group	Echo Likelihood Gap (per-token)				Extended Echo Metrics			
	N	Mean $\overline{\Delta\mathcal{L}}$	Std $\sigma(\Delta\mathcal{L})$	Neg. ratio (%)	$\overline{\Delta\mathcal{L}}$ (per token)	$\Delta/\#\text{removed}$	Suffix-only gap	Avg. removed tokens
Correct	819	2.5231	0.7786	0.12	2.4614	0.01103	1.1449	219.7
Wrong	500	2.4421	0.7657	0.00	2.3666	0.01085	1.2938	218.7

171  
 172 The rejection sampling view provides a principled way to think about echo suppression, but to  
 173 measure its effect, we need a concrete metric. Our goal is to quantify how much *preference* the  
 174 model shows for a raw, echo-containing trace versus its trimmed, echo-free counterpart.  
 175

176 Given a raw trace  $y_{\text{raw}}$  and its echo-trimmed counterpart  $y_{\text{trim}}$ , we define the (length-normalized)  
 177 average log-likelihood

$$\mathcal{L}(y) = \frac{1}{|y|} \sum_{t=1}^{|y|} \log \pi_{\theta}(y_t | x, y_{<t}) \quad (4)$$

180 (nats/token). The *Echo Likelihood Gap* is  $\Delta\mathcal{L} = \mathcal{L}(y_{\text{raw}}) - \mathcal{L}(y_{\text{trim}})$ . A positive  $\Delta\mathcal{L}$  means the  
 181 model *prefers* the echo-containing trace. Unless otherwise noted, we report nats/token.  
 182

183 While our framework defines the echo-free distribution  $\tau_{\theta}$ , the partition function  $Z_x$  required to  
 184 compute it is intractable. This motivates a practical, sample-based alternative: the Echo Likelihood  
 185 Gap ( $\Delta\mathcal{L}$ ). This metric serves as a direct proxy for the probabilistic cost of an echo by comparing  
 186 the average log-likelihood of a generated trace ( $y_{\text{raw}}$ ) against its trimmed counterpart ( $y_{\text{trim}}$ ). A  
 187 positive  $\Delta\mathcal{L}$  indicates that the model assigns a higher likelihood to the sequence containing the  
 188 echo, quantifying the *price* of this behavior on a per-sample basis.  
 189

190 This leads to our central question: is there a positive relationship between this probabilistic cost and  
 191 the model’s final reasoning accuracy? In other words, does *spending* probability on an echo lead to  
 192 better performance? The following sections are dedicated to empirically validating this trade-off.  
 193

### 3.2 THE ECHO LIKELIHOOD GAP IN PRACTICE

194 Table 1 and Figure 2 highlight our central empirical finding: a larger probabilistic investment in an  
 195 Echo of Prompt (EOP) strongly correlates with correct final answers. To formalize this, we introduce  
 196 two metrics: the overall Echo Likelihood Gap ( $\Delta\mathcal{L}$ ), which measures the total probabilistic cost,  
 197 and the Suffix-only Likelihood Gap ( $\Delta\mathcal{L}_{\text{suffix}}$ ), which isolates the echo’s influence on subsequent  
 198 reasoning.  
 199

200 **Defining the Likelihood Gaps.** Our primary metric, the **Echo Likelihood Gap** ( $\Delta\mathcal{L}$ ), is defined  
 201 in §3 as the difference in average per-token log-likelihood between a raw trace  $y_{\text{raw}}$  and its trimmed  
 202 counterpart  $y_{\text{trim}}$ . To isolate the echo’s influence on the subsequent reasoning steps, we introduce  
 203 a more granular metric. For a raw trace  $y_{\text{raw}}$  composed of an echo prefix  $e$  and a reasoning suffix  
 204  $s$  (i.e.,  $y_{\text{raw}} = [e, s]$ ), we compare the model’s likelihood of generating  $s$  with and without the  
 205 conditioning prefix  $e$ . The **Suffix-only Likelihood Gap** is defined as:

$$\Delta\mathcal{L}_{\text{suffix}} = \mathcal{L}(s | x, e) - \mathcal{L}(s | x), \quad (5)$$

206 where  $\mathcal{L}(s | \cdot)$  is the average per-token log-likelihood of the suffix  $s$  under the given context.  
 207 A positive value indicates that the echo prefix makes the subsequent reasoning trace appear more  
 208 probable to the model.  
 209

210 **Analysis of Results.** The overall Echo Likelihood Gap reveals a clear correlation with correctness.  
 211 As shown in Table 1, the Correct group ( $N = 819$ ) has a larger average gap than the Wrong group  
 212 ( $N = 500$ ):  $\overline{\Delta\mathcal{L}} = 2.5231$  vs.  $2.4421$ . This positive difference ( $+0.0811$  nats/token) indicates that  
 213 a larger total probabilistic investment in echoing co-occurs with correct final answers. We further  
 214 validate this relationship with logistic regression in the Appendix, confirming  $\Delta\mathcal{L}$  as a significant  
 215 positive predictor of correctness, even after controlling for trace length.

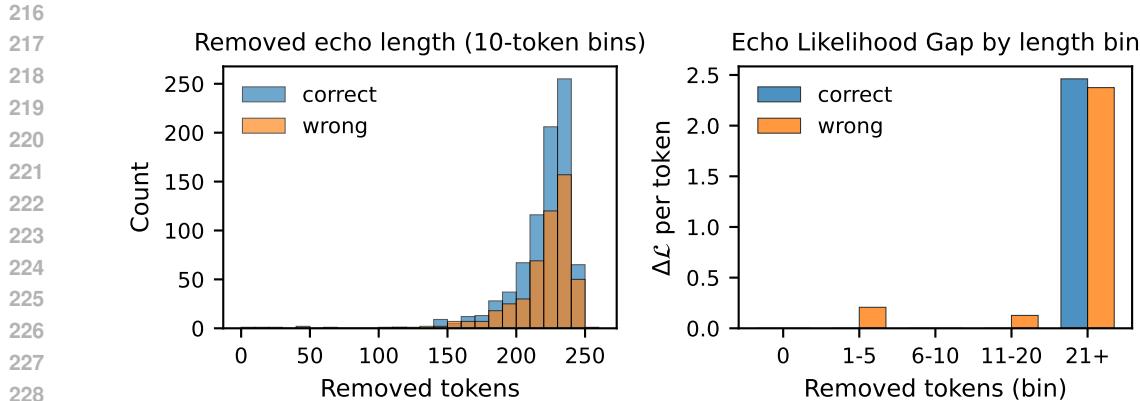


Figure 2: **Echo metrics on GSM8K.** **Left:** High-resolution histogram of removed echo-prefix lengths (10-token bins) for correct and wrong traces; most mass lies between roughly 200 and 240 tokens. **Right:** Echo Likelihood Gap  $\Delta L$  (per-token) stratified by removed-prefix length bin; the gap remains positive across all bins.

Interestingly, the *Suffix-only Likelihood Gap* is slightly larger for the Wrong group (1.2938 vs. 1.1449). While this seems counter-intuitive, it does not contradict our main finding. It suggests that while echoes make subsequent reasoning seem more plausible in general (as both values are positive), they may also act as a form of "confirmation bias," slightly strengthening the model's confidence in locally coherent but ultimately flawed reasoning paths. The determinative factor for correctness remains the overall likelihood trade-off captured by  $\Delta L$ , as explained by the likelihood decomposition in §A.7.

**Sanity Checks.** Before further analysis, we perform several checks to validate  $\Delta L$  as a metric. First, for traces without a detected echo,  $\Delta L$  is definitionally zero, as the raw and trimmed sequences are identical. Second, we confirm that  $\Delta L$  correlates positively with the number of removed tokens in the echo prefix, indicating that longer echoes correspond to a larger likelihood gap. Finally, the data in Table 1 confirms that the suffix-only likelihood gap on the shared reasoning trace remains positive, confirming the echo's influence extends beyond the prefix itself. These checks establish  $\Delta L$  as a robust measure of the echo's probabilistic impact.

**Length- And Suffix-Controlled Analysis.** To ensure this gap is not merely a length artifact, we conduct a length-stratified analysis. As shown in Figure 2, the  $\Delta L$  remains consistently positive across different trace lengths. This indicates that the Echo of Prompt (EOP)'s contribution is robust and also improves the model's scoring on the subsequent, shared reasoning steps.

**Distribution Of Removed Echo Lengths.** We further examine the length distribution of the removed echo prefixes (Figure 2, left). The distribution is heavy-tailed, with most prefixes falling between roughly 200 and 240 tokens (mean 219, median 226), confirming that the Echo of Prompt (EOP) constitutes a non-trivial segment of the generation that acts as a probabilistic sink. This distribution reveals that echo prefixes consistently consume substantial portions of the model's output budget, with most instances removing more than 200 tokens of echoed content.

### 3.3 UNVEILING THE MECHANISM: ECHOES AS ATTENTION REFOCUSING

To understand why prompt echoing is effective, we analyze the model's attention patterns during generation. We hypothesize that re-introducing the original prompt effectively refocuses the model's attention on the core problem statement, preventing drift during extended reasoning chains.

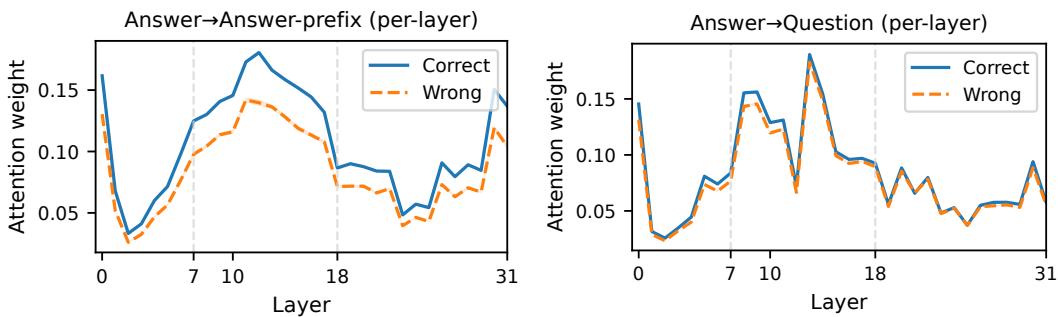
**Attention Redistribution After Echo Removal.** We investigate the mechanism underlying the Echo of Prompt's effectiveness, hypothesizing that it serves to refocus the model's attention. To test this, we compute two attention metrics on the original outputs: (i) attention from answer tokens

270 to the question tokens, and (ii) attention from answer tokens to the answer prefix. Formally, let  
 271  $A^{(l)} \in \mathbb{R}^{T \times T}$  be the head-averaged attention matrix at layer  $l$  for a full sequence of  $T$  tokens. We  
 272 define the average attention weight from a set of query tokens with indices  $\mathcal{S}_Q$  to a set of key tokens  
 273 with indices  $\mathcal{S}_K$  as:

$$\text{Attn}^{(l)}(\mathcal{S}_Q \rightarrow \mathcal{S}_K) = \frac{1}{|\mathcal{S}_Q|} \sum_{i \in \mathcal{S}_Q} \sum_{j \in \mathcal{S}_K} A_{ij}^{(l)}. \quad (6)$$

277 For our *answer*→*answer-prefix* metric,  $\mathcal{S}_Q$  comprises the indices of all tokens in the generated  
 278 reasoning trace, while  $\mathcal{S}_K$  contains the indices of the initial  $K$  tokens of that same trace. This metric  
 279 quantifies the degree to which subsequent reasoning steps are grounded in the model’s own initial  
 280 problem interpretation. For all results reported in this section, including the aggregate statistics  
 281 in Table 2 and the layer-wise analysis in Figure 3, the prefix length is dynamically set to the per-  
 282 sample echo length estimated by our MLP probe. This allows for a precise analysis of the *actual*  
 283 echo’s effect. As the results show, correctly solved problems consistently exhibit stronger attention  
 284 to the answer prefix than incorrect ones, supporting our *attention refocusing* hypothesis.

285 **Layer-Wise Attention Dynamics.** To further understand where the attention refocusing occurs  
 286 within the model’s architecture, we conducted a fine-grained layer-wise analysis across all 32 layers.  
 287 Figure 3 visualizes the attention weight distribution, revealing distinct patterns between correct and  
 288 incorrect reasoning traces.



301 Figure 3: Layer-wise attention weight distribution on GSM8K (DeepSeek-R1-Distill-Llama-8B)  
 302 for Left: *answer*→*answer-prefix* and Right: *answer*→*question*. The blue lines represent correct  
 303 reasoning traces while orange lines represent incorrect ones. The attention refocusing effect is most  
 304 pronounced in layers 7–18 for *answer*→*answer-prefix*, with correct traces maintaining consistently  
 305 higher attention weights.

306 The layer-wise analysis localizes the EOP effect to the model’s reasoning bottleneck. We observe  
 307 a **Middle-Layer Dominance** where the attention gap peaks in layers 7–18, consistent with findings  
 308 that intermediate layers govern reasoning aggregation. Furthermore, the **Differential Impact**—  
 309 where correct traces attend significantly more to the answer-prefix (peak  $\Delta \approx 3\%$ ) than to the  
 310 original question ( $\Delta < 1\%$ )—confirms that the echo acts as a distinct working memory anchor,  
 311 actively refocusing the model on the problem statement during critical computation steps.

312 Interestingly, the early layers (1–6) show minimal differences between correct and incorrect groups,  
 313 with both trajectories nearly overlapping. This suggests that low-level token processing remains  
 314 largely unaffected. Crucially, as shown in Figure 3 Right, the *answer*→*question* attention remains  
 315 closely matched across all layers, serving as a valuable negative control.

317 This confirms that the performance gain is not attributable to a simple, uniform increase in attention  
 318 to the original question. Instead, the discriminative signal emerges in Figure 3 Left, where the diver-  
 319 gence in *answer*→*answer-prefix* attention begins at layer 7. We report mean  $\pm$  s.e.m. across sam-  
 320 ples; focusing on layers 7–18 reveals a group difference of  $\Delta(C-W) \approx 0.66\%$  for *answer*→*question*  
 321 and a more substantial  $\approx 2.87\%$  for *answer*→*answer-prefix*.

322 This supports our hypothesis that the Echo of Prompt acts as a cognitive scaffold for higher-level rea-  
 323 soning; it is not merely about re-reading the question, but about anchoring the subsequent reasoning  
 324 process to a stable internal representation, a mechanism that strongly correlates with correctness.

324  
 325 Table 2: Average attention weights (%) on GSM8K (DeepSeek-R1-Distill-Llama-8B) with probe-  
 326 estimated prefix length. Global statistics and layer-specific analysis showing attention from answer  
 327 tokens to question and answer prefix.

Metric	Correct	Wrong	Diff (C-W)
<i>Global Statistics</i>			
Last-layer: answer → question	5.77%	5.54%	+0.23%
Last-layer: answer → answer-prefix	13.69%	10.41%	+3.28%
All-layers mean: answer → question	8.45%	8.00%	+0.45%
All-layers mean: answer → answer-prefix	10.64%	8.49%	+2.15%
<i>Peak Effect Layers (7-18)</i>			
Layers 7-18: answer → question	12.17%	11.51%	+0.66%
Layers 7-18: answer → answer-prefix	14.45%	11.58%	+2.87%

334  
 335 Table 3: Layer-wise discriminability (Correct vs. Wrong) aggregated into **layer groups**. Metrics are  
 336 computed on answer→answer-prefix and answer→question attention. The mid-layer group shows  
 337 the strongest effect size ( $d$ ) for answer→answer-prefix.

Layer Group (layers)	AUC↑ (Ans→Pref)	$d\uparrow$ (Ans→Pref)	AUC (Ans→Q)	$d$ (Ans→Q)
Early (0-6)	0.716	0.820	0.628	0.482
Mid (7-18)	0.719	<b>0.832</b>	0.585	0.303
Late (19-31)	<b>0.723</b>	0.828	0.563	0.184

347 To ensure our findings are robust and not merely an artifact of the dynamically-set prefix length, we  
 348 conducted an ablation study using fixed prefix lengths. The results, detailed in §A.5, confirm that  
 349 the attention gap persists across several fixed prefix lengths, supporting our conclusion that EOP’s  
 350 function is genuine attention refocusing.

351  
 352 **Layer-Wise Discriminability.** To quantify where attention refocusing emerges, we compute  
 353 layer-wise discriminability between Correct and Wrong groups using AUC and Cohen’s  $d$ . Specifi-  
 354 cally, for each layer  $l$ , we treat the attention scores from the Correct traces ( $\mathcal{A}_C^{(l)}$ ) and Wrong traces  
 355 ( $\mathcal{A}_W^{(l)}$ ) as two distributions. The Area Under the ROC Curve (AUC) measures how well attention at  
 356 a given layer classifies a trace as correct. We also compute Cohen’s  $d$  to quantify the effect size:

$$d^{(l)} = \frac{\mu(\mathcal{A}_C^{(l)}) - \mu(\mathcal{A}_W^{(l)})}{s_p^{(l)}}, \quad (7)$$

360 where  $\mu$  denotes the mean and  $s_p^{(l)}$  is the pooled standard deviation for layer  $l$ . We also aggregate  
 361 layers into three **groups** to analyze broader trends. Table 3 shows that the **mid-layer group (7-18)**  
 362 **exhibits the strongest effect size on answer→answer-prefix** (Cohen’s  $d=0.832$ ), with its AUC  
 363 being comparable to the late-layer group. In contrast, the *answer→question* discriminability remains  
 364 significantly lower across all groups, serving as a negative control.

365 These statistics, combined with the attention trajectories (Figure 3), strongly indicate that EOP’s pri-  
 366 mary mechanism is to refocus representations within the mid layers ( layers 7 through 18), anchoring  
 367 subsequent reasoning to the answer prefix.

## 370 4 EMPIRICAL VALIDATION

### 372 4.1 ECHO REINSERTION AS A CAUSAL INTERVENTION

374 We construct an interventional experiment that starts from *failed* GSM8K completions produced by  
 375 several models (**DeepSeek-R1-Distill-Llama-8B**, **Qwen3-8B**, and the non-reasoning **Qwen3-8B-Base**). For each wrong example we (i) truncate an echo-free trace to 50% of its tokens to obtain  
 376 a shared prefix, (ii) resume generation either directly (**echo-free**) or after inserting the template  
 377 phrase “now I need to look back at the question again:” (**echo reinsertion**), and (iii) score the new

378  
 379 Table 4: Echo reinsertion ablation on the wrong-subset GSM8K traces for several models. We  
 380 report exact-match accuracy (%) when continuing from the same echo-free prefix with and without  
 381 a forced echo phrase.

Model	Echo-free EM (%)	Echo reinsertion EM (%)
DeepSeek-R1-Distill-Llama-8B	15.85	26.22
Qwen3-8B	21.34	29.27
Qwen3-8B-Base (no CoT)	10.56	10.56

386  
 387 Table 5: Supervised fine-tuning with distilled CoT data improves mathematical reasoning.  
 388 –ED-SFT denotes models fine-tuned on our EOP distilled dataset. –normal-SFT refers to the  
 389 normal CoT distilled dataset. Base is the pretrained model; the unmarked variant is instruction-tuned.  
 390 Best scores in each column are in **bold**.

Model	GSM-8K		MathQA	Hendrycks-MATH
	Strict EM	Flex EM		
<i>Qwen3-8B-Base</i>				
Base	79.4	80.5	31.0	0.76
+ ED-SFT	<b>94.2 (+3.4)</b>	<b>94.2 (+3.4)</b>	<b>58.8 (+11.8)</b>	<b>10.0 (+8.2)</b>
+ normal-SFT	90.8	90.8	47.0	1.8
<i>Qwen3-8B (Instruct Version)</i>				
Base	87.49	88.1	49.2	0.8
+ ED-SFT	93.1 (+2.8)	93.4 (+3.3)	53.7 (+1.9)	6.1 (+1.1)
+ normal-SFT	90.3	90.1	51.8	5.0
<i>DeepSeek-Distill-Llama-8B</i>				
Base	67.6	66.1	31.6	0.38
+ ED-SFT	78.2	79.7 (+0.2)	34.8 (+3.4)	3.0 (+2.24)
+ normal-SFT	80.5	79.5	31.4	0.76

405  
 406 completions with the standard GSM8K exact-match script. Both branches see identical questions,  
 407 prefixes, decoding parameters, and random seeds, isolating the causal impact of the injected echo.  
 408

409 The intervention confirms that echoes are *causally helpful* for reasoning-capable models: forcing a  
 410 short echo before resuming the chain yields sizable absolute EM gains (+10.4 and +7.9 points for  
 411 DeepSeek-R1-Distill-Llama-8B and Qwen3-8B, respectively), while the non-reasoning base model  
 412 shows no improvement. This null result for the base model is expected, as it lacks the instruction-  
 413 following and reasoning priors typically acquired via RLHF to utilize the re-injected context in a  
 414 zero-shot manner. This interpretation is reinforced by our ED-SFT results (Table 5), where the  
 415 base model exhibits the largest relative improvement (+3.4 points) when the reasoning capability  
 416 and echo strategy are instilled simultaneously. Qualitatively, we observe the reinsertion branch  
 417 revisiting the original quantities and constraints (Figure 9), whereas the echo-free branch continues  
 418 the drifting reasoning that led to the initial error. This interventional evidence complements our  
 419 correlational analyses (§3) and grounds the attention refocusing hypothesis in an explicit cause-and-  
 420 effect experiment.

## 421 4.2 PERFORMANCE GAINS FROM ECHO-DISTILLED SFT (ED-SFT)

422 Having established that Echo of Prompt correlates with improved reasoning performance, we investi-  
 423 giate whether this behavior can be systematically instilled through targeted training. Our Supervised  
 424 Fine-Tuning (SFT) methodology is inspired by recent work (Team, 2025).

425 The core hypothesis is that **explicitly training models on echo-prefixed traces enhances their**  
 426 **problem-solving approach**: the echo phase forces deeper engagement with problem constraints  
 427 and establishes a stronger foundation for subsequent reasoning steps.

428  
 429 **Methodology.** We develop **Echo-Distilled SFT (ED-SFT)**, a supervised fine-tuning method that  
 430 embeds this echo-then-reason pattern as a learned behavior. We first construct a shared pool of  
 431 high-quality teacher traces on GSM8K by querying a capable teacher model, gpt-oss-120B,

432 with a standard CoT prompt that wraps the reasoning in a single `<think>` block and requires the  
 433 final answer to be a plain value. We automatically verify that the final answer exactly matches the  
 434 ground-truth solution and discard any trace that fails this check. This pool of verified *(question,*  
 435 *CoT, answer)* triples is the common source for both ED-SFT and the **normal-SFT** baseline.

436 To obtain **ED-SFT** data, we encourage an explicit echo-then-reason pattern at the head of the trace.  
 437 We train a small MLP probe to detect whether an early Echo of Prompt segment is present. For  
 438 traces flagged as missing EOP, we call `gpt-oss-120B` once more with an edit instruction that  
 439 *minimally* inserts a short echo-style opening that restates the question (e.g., “Okay, let me see. The  
 440 problem is asking: …”) while preserving the subsequent reasoning and final answer. Traces that  
 441 already contain an echo are kept unchanged. After editing, we re-run the automatic checker and drop  
 442 any example whose final answer no longer matches the gold label. The resulting ED-SFT dataset  
 443 therefore differs from the standard CoT pool only by the presence of an initial echo segment.

444 For the **normal-SFT** baseline we again start from the same verified teacher traces but *remove* the  
 445 echo segment while keeping the remainder of the reasoning untouched. Because the MLP probe is  
 446 a binary EOP detector rather than a span localizer, we delegate span selection to the teacher: when  
 447 the probe predicts EOP presence, we prompt `gpt-oss-120B` to delete the echo prefix under a  
 448 “do not change the reasoning or final answer” instruction, and we discard any sample whose answer  
 449 changes. This yields paired ED-SFT and normal-SFT corpora that are nearly identical token-wise  
 450 and differ primarily in the presence or absence of the initial echo. On GSM8K, the inclusion of the  
 451 echo prefix results in a longer average sequence length for ED-SFT compared to normal-SFT (175  
 452 vs. 136 tokens).

453 **Experimental Setup.** To test the echo strategy’s effectiveness at different stages of model training,  
 454 we fine-tuned models from two families: Qwen3 (8B) and Deepseek-distill-Llama-8B. For  
 455 the Qwen3 family, we experimented on two distinct versions: the pre-trained base model (**Qwen3-8B-Base**) and the final, fully instruction-tuned model (**Qwen3-8B**). For each model, we applied  
 456 our SFT procedure to produce both `-ED-SFT` and `-normal-SFT` variants. All fine-tuning runs  
 457 use the same optimizer (AdamW), learning-rate schedule, batch size, maximum sequence length,  
 458 and number of training steps; the only difference is whether the training traces come from the  
 459 echo-augmented (ED-SFT) or echo-trimmed (normal-SFT) versions of the same teacher CoTs. We  
 460 evaluated all models on a suite of mathematical reasoning benchmarks (GSM8K, MathQA, and  
 461 Hendrycks-MATH) to assess generalization under distribution shift, as fine-tuning was performed  
 462 *only* on the GSM8K training set with 7k samples.

463 **Results.** As shown in Table 5, SFT with our distilled data yields substantial and consistent  
 464 performance improvements. Crucially, these gains appear on both the pre-trained base and the  
 465 fully-aligned instruction-tuned models. Fine-tuning the base model (**Qwen3-8B-Base-echo-SFT**)  
 466 achieves a remarkable gain of +3.4 points on GSM-8K, while fine-tuning the already capable  
 467 instruction-tuned model (**Qwen3-8B-echo-SFT**) still provides a solid boost of +2.8 points.

468 **Cross-Model Generalization.** The effectiveness of Echo-Distilled SFT extends across different  
 469 model architectures. For **DeepSeek-distill-llama-8B**, we observe consistent improvements, with  
 470 particularly strong gains on benchmarks that differ distributionally from GSM8K, such as MathQA  
 471 (+3.4 points) and Hendrycks-MATH (+2.24 points). The consistent success on both base and instruct  
 472 models strongly suggests that the Echo of Prompt (EOP) is a fundamental and transferable cognitive  
 473 alignment strategy, not merely an artifact of existing instruction tuning.

474 **Mechanistic Alignment With Attention Analysis.** The success of ED-SFT can be understood  
 475 through the lens of our attention analysis (§3.3). The layer-wise attention patterns reveal that models  
 476 trained with echo-prefixed traces naturally develop stronger attention connections in middle layers  
 477 (7-18), where we observed the most significant differences between correct and incorrect reasoning  
 478 (1.73% increase in `answer → answer-prefix` attention). This suggests that ED-SFT effectively instills  
 479 the attention refocusing mechanism we identified in our analysis, teaching models to leverage these  
 480 critical layers for maintaining problem-relevant attention throughout the reasoning process.

481 To further substantiate this, we analyzed the `answer → answer-prefix` attention gap (Correct – Wrong)  
 482 specifically within the critical mid-layer block (layers 7–18) across our model variants. This targeted  
 483 metric confirms that **ED-SFT** most effectively strengthens this mechanism, exhibiting the largest

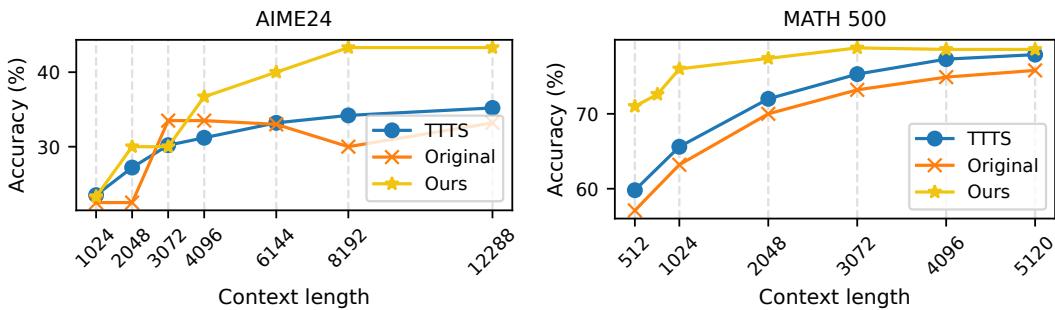


Figure 4: Echoic Prompting (EP) vs. TTTS on AIME24 (left) and MATH-500 (right).

attention gap (+3.20 pp) compared to the base model (+1.90 pp) and the normal SFT variant (+2.40 pp). This finding provides direct evidence that ED-SFT successfully instills the desired refocusing behavior where it is most impactful. Full statistics are provided in §A.2.

### 4.3 ECHOIC PROMPTING (EP): A TRAINING-FREE ENHANCEMENT

**The Echoic Prompting (EP) Method.** Our proposed **Echoic Prompting (EP)** strategy is a training-free method designed to enhance reasoning capabilities at inference time. The core idea is to re-engage the model by echoing the original prompt. Specifically, after the model produces an initial reasoning chain, we append a reminder phrase such as *look back at the question again* followed by the original question itself. This intervention encourages the model to revisit the problem’s context and continue generating a more grounded response. Unlike methods that inject generic, task-agnostic stimuli, EP re-grounds the model with task-specific context from the original query and shows consistent gains over 2 math reasoning datasets following TTTS’s settings.

**Experimental Setup.** To evaluate the effectiveness of EP, we compare it against a strong baseline, Thinking Token based Test-time Scaling (TTTS) (Qian et al., 2025), which artificially inserts generic *thinking tokens* (e.g., *So*, *Hmm*) to spur reasoning. For a fair comparison, we reproduce TTTS following its official implementation. Both methods are evaluated on the **DeepSeek-R1-Distill-Llama-8B** model, using the vLLM backend with deterministic decoding (temperature=0.0).

**Results.** As shown in Figure 4, our EP approach consistently and substantially outperforms TTTS across both AIME24 and MATH-500. The performance gains are robust under identical decoding and budget settings. This indicates that re-grounding the model on the input via a natural echo of the prompt is more effective than injecting generic, artificial thinking tokens.

## 5 CONCLUSION

In this work, we systematically investigated the Echo of Prompt (EOP), the spontaneous tendency of large reasoning models to repeat a user’s query before generating a solution. We introduced a probabilistic framework based on rejection sampling to define and measure the *Echo Likelihood Gap*. Through detailed attention analysis, we provided a mechanistic explanation, demonstrating that EOP serves to refocus attention on task-critical information, particularly in model’s middle layers.

To harness this phenomenon, we proposed two practical methods: *Echo-Distilled SFT (ED-SFT)*, which instills behavior through fine-tuning, and *Echoic Prompting (EP)*, a training-free inference technique. Across multiple mathematical reasoning benchmarks, both methods demonstrated consistent performance gains over strong baselines, validating EOP as a beneficial cognitive primitive.

Ultimately, our work advocates for a shift in alignment research from validating final outputs to cultivating beneficial thought processes. This provides a framework to understand emergent behaviors like EOP, paving the way for more interpretable and robust AI systems. Future work could engineer dynamic echo mechanisms across diverse tasks and model scales, bridging the gap between emergent phenomena and deliberate cognitive design.

540 ETHICS STATEMENT  
541

542 All authors adhere to the ICLR Code of Ethics. Our research focuses on understanding and im-  
543 proving the reasoning capabilities of large language models, a foundational scientific goal. The  
544 datasets used for fine-tuning and evaluation, such as GSM8K and MathQA, are standard public  
545 benchmarks in the field. The synthetic data used for SFT was generated by a large proprietary  
546 model GPT-OSS-120B, and the annotations for our MLP probe were assisted by GPT-4.1, as  
547 detailed in our LLM Usage Disclosure in §A.1. We acknowledge that the models used in this study  
548 may inherit biases from their original, opaque training data. Our work does not introduce new ap-  
549 plications with foreseeable negative societal impacts. We believe that a deeper understanding of  
550 emergent behaviors like the Echo of Prompt contributes to the development of more transparent and  
551 reliable AI systems.

552 REPRODUCIBILITY STATEMENT  
553

554 To ensure the reproducibility of our findings, we commit to releasing our source code, including  
555 scripts for data processing, training, and evaluation, upon publication. Our work relies on pub-  
556 licly available models, including the Qwen3 and DeepSeek series, and standard benchmarks such as  
557 GSM8K, MathQA, and Hendrycks-MATH. The methodology for our Echo-Distilled SFT data gen-  
558 eration is detailed in §A.8. The design and training of the MLP probe used for echo detection are  
559 fully described in §A.4, and all experimental hyperparameters and evaluation settings are detailed  
560 in §4. The theoretical claims in §3 are self-contained within the paper.

561 REFERENCES  
562

563 Zhanzhan Cheng, Fan Bai, Yunlu Xu, Gang Zheng, Shiliang Pu, and Shuigeng Zhou. Focusing  
564 attention: Towards accurate text recognition in natural images. *arXiv preprint arXiv:1709.02054*,  
565 2017.

566 DeepSeek-AI et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement  
567 learning. *arXiv preprint arXiv:2501.12948*, 2025. URL <https://arxiv.org/abs/2501.12948>.

568 Zhuohan Gu, Jiayi Yao, Kuntai Du, and Junchen Jiang. Llmsteer: Improving long-context llm  
569 inference by steering attention on reused contexts. *arXiv preprint arXiv:2411.13009*, 2024.

570 Cheng-Yu Hsieh, Yung-Sung Chuang, Chun-Liang Li, Zifeng Wang, Long Le, Abhishek Kumar,  
571 James Glass, Alexander Ratner, Chen-Yu Lee, Ranjay Krishna, and Tomas Pfister. Found in the  
572 middle: Calibrating positional attention bias improves long context utilization. In Lun-Wei Ku,  
573 Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Lin-  
574 guistics: ACL 2024*, 2024. URL <https://aclanthology.org/2024.findings-acl.890/>.

575 Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large  
576 language models are zero-shot reasoners. In *Advances in Neural Information Processing Systems  
577 (NeurIPS)*, 2022. URL [https://papers.nips.cc/paper\\_files/paper/2022/hash/8bb0d291acd4acf06ef112099c16f326-Abstract-Conference.html](https://papers.nips.cc/paper_files/paper/2022/hash/8bb0d291acd4acf06ef112099c16f326-Abstract-Conference.html).

578 Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and  
579 Percy Liang. Lost in the middle: How language models use long contexts. *Transactions of  
580 the Association for Computational Linguistics*, 2024. URL <https://aclanthology.org/2024.tacl-1.9/>.

581 Rajasekhar Reddy Mekala, Yasaman Razeghi, and Sameer Singh. Echoprompt: Instructing the  
582 model to rephrase queries for improved in-context learning, 2024.

583 OpenAI. Gpt-4.1. Online documentation, 2024. URL <https://platform.openai.com/docs/models#gpt-4-1>. Accessed 2025-08-09.

584 OpenAI. Learning to reason with llms. <https://openai.com/index/learning-to-reason-with-llms/>, 2024.

594 Chen Qian, Dongrui Liu, Haochen Wen, Zhen Bai, Yong Liu, and Jing Shao. Demystifying reasoning dynamics with mutual information: Thinking tokens are information peaks in llm reasoning.  
 595 *arXiv preprint arXiv:2506.02867*, 2025.

596

597 Qwen Team. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. URL <https://arxiv.org/abs/2505.09388>.

598

599

600 Yang Sui, Yu-Neng Chuang, Guanchu Wang, Jiamu Zhang, Tianyi Zhang, Jiayi Yuan, Hongyi Liu,  
 601 Andrew Wen, Shaochen Zhong, Na Zou, Hanjie Chen, and Xia Hu. Stop overthinking: A survey  
 602 on efficient reasoning for large language models, 2025. URL <https://arxiv.org/abs/2503.16419>.

603

604 NovaSky Team. Sky-t1: Train your own o1 preview model within \$450. <https://novasky-ai.github.io/posts/sky-t1>, 2025. Accessed: 2025-01-09.

605

606

607 Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed H. Chi, Sharan Narang, Aakanksha  
 608 Chowdhery, and Denny Zhou. Self-consistency improves chain-of-thought reasoning in lan-  
 609 guage models. In *International Conference on Learning Representations (ICLR)*, 2023. URL  
 610 <https://openreview.net/forum?id=1PL1NIMMrw>.

611

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

615

616 Xiaohan Xu, Chongyang Tao, Tao Shen, Can Xu, Hongbo Xu, Guodong Long, Jian-Guang Lou, and  
 617 Shuai Ma. Re-reading improves reasoning in large language models. In Yaser Al-Onaizan, Mohit  
 618 Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods  
 619 in Natural Language Processing*, Miami, Florida, USA, 2024. Association for Computational  
 Linguistics. URL <https://aclanthology.org/2024.emnlp-main.871/>.

620

621 Junchi Yao, Shu Yang, Jianhua Xu, Lijie Hu, Mengdi Li, and Di Wang. Understanding the repeat  
 622 curse in large language models from a feature perspective. In Wanxiang Che, Joyce Nabende,  
 623 Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Findings of the Association for Com-  
 624 putational Linguistics: ACL 2025*, Vienna, Austria, July 2025. Association for Computational  
 Linguistics. URL <https://aclanthology.org/2025.findings-acl.406/>.

625

626 Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao,  
 627 and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large  
 628 language models. In *Advances in Neural Information Processing Systems (NeurIPS)*,  
 629 2023. URL [https://proceedings.neurips.cc/paper\\_files/paper/2023/hash/271db9922b8d1f4dd7aaef84ed5ac703-Abstract-Conference.html](https://proceedings.neurips.cc/paper_files/paper/2023/hash/271db9922b8d1f4dd7aaef84ed5ac703-Abstract-Conference.html).

630

631 Han Zhao, Haotian Wang, Yiping Peng, Sitong Zhao, Xiaoyu Tian, Shuaiting Chen, Yun-  
 632 jie Ji, and Xiangang Li. 1.4 million open-source distilled reasoning dataset to em-  
 633 power large language model training, 2025. URL <https://arxiv.org/abs/2503.19633>. Hugging Face dataset: <https://huggingface.co/datasets/a-m-team/AM-DeepSeek-R1-Distilled-1.4M>, CC-BY-NC-4.0.

634

635

636 Xin Zou, Yizhou Wang, Yibo Yan, Yuanhuiyi Lyu, Kening Zheng, Sirui Huang, Junkai Chen,  
 637 Peijie Jiang, Jia Liu, Chang Tang, et al. Look twice before you answer: Memory-space vi-  
 638 sual retracing for hallucination mitigation in multimodal large language models. *arXiv preprint  
 639 arXiv:2410.03577*, 2024.

640

641

642

643

644

645

646

647

648 **A APPENDIX**  
649650 **A.1 THE USE OF LARGE LANGUAGE MODELS**  
651652 In accordance with ICLR 2026 policy, we disclose the usage of Large Language Models (LLMs) in  
653 this research. Our use of LLMs is primarily in three areas:  
654655 1. **Writing Assistance:** We utilized LLMs to improve the clarity, grammar, and overall read-  
656 ability of the manuscript. This involved proofreading and refining sentences without alter-  
657 ing the core scientific contributions.  
658 2. **Data Annotation:** As detailed in §A.4, GPT-4.1 was employed to annotate our Chain-of-  
659 Thought (CoT) dataset. This process was crucial for training the MLP probe to accurately  
660 identify Echo of Prompt (EOP) instances.  
661 3. **Synthetic Data Generation:** The CoT dataset for Supervised Fine-Tuning (SFT), dis-  
662 cussed in §A.8, was generated using the gpt-oss-120B model. This provided the foundation  
663 for training our models to exhibit the desired echo behavior.  
664665 **A.2 ADDITIONAL CROSS-MODEL ATTENTION STATISTICS**  
666667 To supplement the analysis in §4, Table 6 provides the detailed mid-layer (layers 7–18) attention  
668 gap statistics for the answer→answer-prefix metric across model variants. The gap is computed as  
669 the difference in average attention percentage points (pp) between correct and incorrect reasoning  
670 traces. A larger positive value indicates stronger within-model discriminability.  
671672 Table 6: Mid-layer (layers 7–18) Ans→Pref gap (pp, Correct–Wrong). Values are for within-model  
673 discriminability; do not use for cross-model ranking.  
674

675 <b>Model / Setting</b>	676 <b>Mid-layer gap (pp)</b>
677 Qwen3-8B-Base	1.90
678 Echo-SFT	3.20
678 Normal-SFT	2.40

680 **A.3 ON THE ORIGINS OF THE ECHO OF PROMPT**  
681682 While the precise mechanisms underlying the formation of the Echo of Prompt (EOP) are not yet  
683 fully understood, we note its appearance in related emergent LLM-reasoning phenomena, such as  
684 the initial COT prompting (Wei et al., 2022) and the "aha moment" observed in DeepSeek-R1-zero  
685 (DeepSeek-AI et al., 2025). We hypothesize that EOP is an emergent behavior that arises from the  
686 model's implicit need to ground its reasoning process in the problem statement. By restating the  
687 prompt, the model may be reinforcing its internal representation of the task, thereby improving its  
688 focus on relevant information for subsequent reasoning steps.  
689690 **A.4 MLP PROBE FOR ECHO DETECTION**  
691692 To operationalize our probabilistic framework, we require a reliable method to detect echo prefixes.  
693 We train a lightweight two-layer MLP probe for this binary classification task.  
694695 **Data and Annotation.** Training data are sampled from the `am_0.9M_1k.jsonl` subset of  
696 AM-DeepSeek-R1-Distilled-1.4M (Zhao et al., 2025). Each example consists of a (question,  
697 think\_content) pair. Labels are generated using a hybrid approach: we prompt GPT-4.1 (OpenAI,  
698 2024) with a deterministic rubric to identify semantic repetition and its approximate boundary. This  
699 boundary is then refined using sentence-level semantic similarity to correct for formatting artifacts.  
700 To validate annotation quality, a random subset of 200 annotations was manually reviewed, showing  
701 over 96% agreement with the final labels. We release the prompt template and parsing code in our  
repository.

702 **Input Features.** For each (question, think\_content) pair, we featurize the sample by concatenating  
 703 two sentence embeddings. The first embedding represents the full question. The second represents  
 704 the initial prefix of the think\_content, defined as the first 32 word tokens. Both are encoded using  
 705 a SentenceTransformer model (Qwen3-Embedding-0.6B). The resulting concatenated vector is z-scored  
 706 before being passed to the probe.

707 **Architecture and Training.** The probe is a two-layer MLP with a 32-dimensional hidden layer  
 708 and ReLU activation, mapping the concatenated embedding to a single logit. We train the model  
 709 using weighted binary cross-entropy on logits (sigmoid + BCE computed in a numerically stable  
 710 form; implemented via PyTorch’s `BCEWithLogitsLoss`), where the positive class weight is set  
 711 to the ratio of negative to positive samples in the training set to handle class imbalance. Optimization  
 712 is performed with AdamW (learning rate  $10^{-4}$ , weight decay 0.01, batch size 64) for up to 200  
 713 epochs, with early stopping (patience 10) on the validation loss. The dataset is split into training  
 714 (70%), validation (15%), and testing (15%) sets.

715 **Evaluation and Usage.** The trained probe’s performance on the held-out test set is reported in  
 716 Table 7. The high AUROC and F1-Score confirm its reliability for identifying echoes. During  
 717 inference for our main experiments (e.g., attention analysis), this probe is used as a predicate to  
 718 identify and measure echo prefixes. It is not used to score task correctness. For truncation, we use  
 719 a calibrated threshold on the sigmoid output with a hysteresis scheme (initial threshold 0.6, drop  
 720 threshold 0.15) to ensure stable prefix detection.

721  
 722  
 723  
 724 Table 7: MLP probe performance on the held-out test set. The probe reliably identifies echo prefixes,  
 725 justifying its use in our framework.

Metric	Accuracy	Precision	Recall	F1-Score	AUROC
Value	0.912	0.921	0.908	0.914	0.963

726 **Reproducibility and licensing.** We fix and log random seeds, dataset hashes, feature extraction  
 727 versions, and probe checkpoints. Data originate from AM-DeepSeek-R1-Distilled-1.4M (Zhao et al.,  
 728 2025) (subset am\_0.9M\_1k.jsonl). The GPT-4.1 annotator is referenced in (OpenAI, 2024).  
 729 Scripts to reproduce this pipeline are provided in our code release.

### 730 A.5 ABLATION STUDY ON FIXED PREFIX LENGTHS

731 To verify that the observed attention refocusing is not merely a byproduct of the echo’s length, we  
 732 performed an ablation study. Instead of using the dynamically estimated echo length from our MLP  
 733 probe, we re-computed the  $answer \rightarrow answer-prefix$  attention metric using several fixed prefix lengths  
 734 ( $K$ ). This allows us to disentangle the effect of prefix length from the functional role of the echo’s  
 735 content.

736 **Methodology.** We repeated the layer-wise attention analysis from §3.3 with fixed prefix lengths of  
 737  $K \in \{32, 64, 128\}$  tokens for all samples. For each value of  $K$ , we calculated the average attention  
 738 from all answer tokens to the first  $K$  answer tokens, separately for the Correct and Wrong groups.

739 **Results.** As shown in Table 8, the attention gap between the Correct and Wrong groups remains  
 740 consistently positive and significant across all fixed prefix lengths. While the magnitude of the gap  
 741 varies with  $K$ , the Correct group consistently directs more attention to the answer prefix. This  
 742 demonstrates that the attention refocusing effect is a robust phenomenon and not just an artifact of  
 743 longer echoes co-occurring with correct answers.

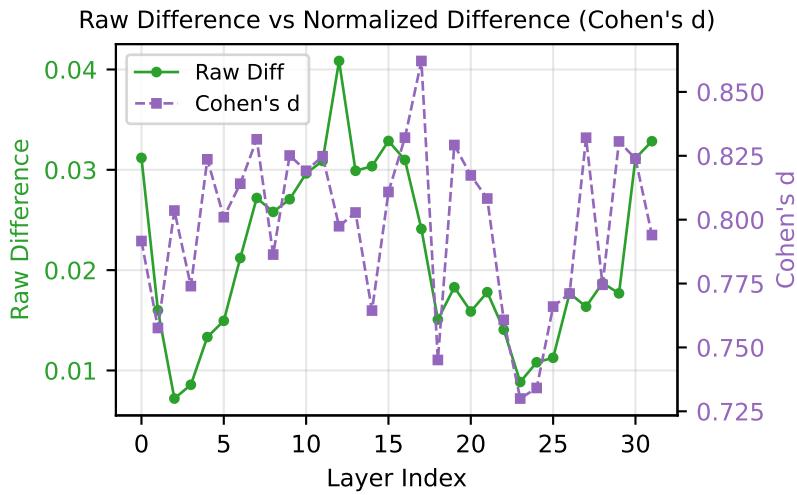
### 744 A.6 VERIFICATION OF ATTENTION NORMALIZATION

745 To ensure that the observed differences in attention weights between correct and incorrect traces are  
 746 not artifacts of absolute weight fluctuations across layers, we performed a normalization analysis.  
 747 We computed the standardized mean difference (Cohen’s  $d$ ) and Z-score differences for each layer.

756  
 757 Table 8: Ablation study on fixed prefix lengths for *answer*→*answer-prefix* attention. Attention  
 758 weights are averaged across all layers. The positive difference (Correct – Wrong) persists for all  
 759 values of  $K$ .

Prefix Length ( $K$ )	Correct (%)	Wrong (%)	Difference (%)
32 tokens	10.61	8.42	2.19
64 tokens	16.78	13.94	2.83
128 tokens	20.51	19.43	1.08

760  
 761  
 762  
 763  
 764  
 765  
 766 Figure 5 presents a dual-axis comparison of the raw attention difference (Correct – Wrong) and the  
 767 normalized Cohen’s  $d$  effect size for the *answer*→*answer-prefix* metric. The two curves track each  
 768 other closely, with the normalized effect size consistently exceeding 0.75 in the critical middle layers  
 769 (7–18) and peaking at 0.86. This confirms that the attention refocusing signal is robust to layer-  
 770 specific magnitude variations and represents a statistically significant difference in model behavior.



790 Figure 5: Comparison of raw attention difference (Correct – Wrong) and normalized effect size  
 791 (Cohen’s  $d$ ) across layers for the *answer*→*answer-prefix* metric. The strong alignment between the  
 792 raw and normalized metrics confirms that the mid-layer refocusing peak is a robust phenomenon.

### A.7 LINKING $\Delta\mathcal{L}$ TO CORRECTNESS: A LIKELIHOOD DECOMPOSITION

798 Let  $\pi_\theta(y | x)$  be the base model and  $\tau_\theta(y | x) = \pi_\theta(y | x) \mathbf{1}_{y \in \mathcal{Y}_{\text{trim}}} / Z_x$  the trimmed distribution  
 799 with  $Z_x > 0$ . For a raw trace  $y_{\text{raw}} = [e, s]$  and its trimmed counterpart  $y_{\text{trim}} = s$ , define the  
 800 per-token log-likelihood  $\mathcal{L}_\pi(y | x) = \frac{1}{|y|} \sum_t \log \pi_\theta(y_t | x, y_{<t})$ , and the Echo Likelihood Gap  
 801  $\Delta\mathcal{L} = \mathcal{L}_\pi(y_{\text{raw}} | x) - \mathcal{L}_\pi(y_{\text{trim}} | x)$ .

802 Because  $\log \tau_\theta(y_{\text{trim}} | x) = \log \pi_\theta(y_{\text{trim}} | x) - \log Z_x$ , we have, for  $n = |y_{\text{trim}}|$ ,

$$\mathcal{L}_\tau(y_{\text{trim}} | x) = \mathcal{L}_\pi(y_{\text{trim}} | x) - c(x, n) = \mathcal{L}_\pi(y_{\text{raw}} | x) - \Delta\mathcal{L} - c(x, n),$$

806 where the “constant” shift is  $c(x, n) = \frac{1}{n} \log Z_x$ . Taking conditional expectations with respect to  
 807 the correctness label  $G \in \{\text{C}, \text{W}\}$  yields

$$\mathbb{E}[\mathcal{L}_\tau | G] = \mathbb{E}[\mathcal{L}_\pi(y_{\text{raw}} | x) | G] - \mathbb{E}[\Delta\mathcal{L} | G] - \mathbb{E}[c(x, n) | G].$$

810 Therefore,

$$\begin{aligned}
 & \mathbb{E}[\mathcal{L}_\tau \mid C] - \mathbb{E}[\mathcal{L}_\tau \mid W] \\
 &= \underbrace{\left( \mathbb{E}[\mathcal{L}_\pi(y_{\text{raw}} \mid x) \mid C] - \mathbb{E}[\mathcal{L}_\pi(y_{\text{raw}} \mid x) \mid W] \right)}_{\text{controlled by length/suffix stratification}} \\
 &\quad - \left( \mathbb{E}[\Delta\mathcal{L} \mid C] - \mathbb{E}[\Delta\mathcal{L} \mid W] \right) \\
 &\quad - \underbrace{\left( \mathbb{E}[c(x, n) \mid C] - \mathbb{E}[c(x, n) \mid W] \right)}_{\approx 0 \text{ if } x \text{ and } n \text{ are matched}}.
 \end{aligned}$$

821 Under matched prompts  $x$  and matched (or stratified) lengths  $n$ , the first and last terms are negligible,  
 822 so the group difference in  $\mathcal{L}_\tau$  is approximately the *negative* of the group difference in  $\Delta\mathcal{L}$ .

#### 823 A.8 COT DISTILLATION PIPELINE

825 Our Chain-of-Thought (CoT) distillation pipeline, used to create the datasets for Echo-Distilled SFT  
 826 (ED-SFT), follows a teacher-student approach grounded in a *single* shared pool of teacher traces.  
 827 We first use a highly capable teacher model (gpt-oss-120B) to generate reasoning traces for the  
 828 training questions (e.g., from the GSM8K training set) with a standard CoT prompt that wraps the  
 829 reasoning in a `<think>` block and enforces an exact-match final answer. Any trace whose final  
 830 answer does not match the gold label is discarded, yielding a pool of verified (question, CoT,  
 831 answer) triples.

832 From this pool we derive two closely matched SFT datasets. For the **ED-SFT** dataset, we encourage  
 833 an explicit echo-then-reason pattern: we train an MLP probe to detect whether an early Echo of  
 834 Prompt segment is present, and for traces predicted to be echo-free we ask gpt-oss-120B to  
 835 minimally insert a short echo-prefix that restates the question while preserving the existing reasoning  
 836 and answer. Traces that already contain an echo are kept as-is. For the **normal-SFT** baseline, we  
 837 again start from the same verified traces but, when the probe predicts EOP presence, we prompt the  
 838 teacher to delete the echo-prefix under a “do not change the reasoning or final answer” instruction.  
 839 In both directions we re-run answer checking and drop any edited example whose final answer  
 840 changes.

841 This procedure yields paired ED-SFT and normal-SFT corpora that are nearly identical token-wise  
 842 and differ primarily in the presence or absence of the initial echo. As reported in the main text, on  
 843 GSM8K the inclusion of the echo prefix results in longer average sequences for ED-SFT compared  
 844 to normal-SFT (175 vs. 136 tokens).

#### 845 A.9 TOKEN-WISE ATTENTION SIGNIFICANCE

847 To verify that the attention refocusing effect is not driven by positional bias or a few outlier tokens,  
 848 we performed a token-wise analysis of the attention weights. Figure 6 shows the average attention  
 849 weights for the first 32 answer tokens towards the answer-prefix (left) and the question (right),  
 850 comparing Correct and Wrong groups.

852 We conducted a Welch’s t-test at each token position. For *answer*→*answer-prefix*, the Correct group  
 853 shows consistently higher attention, with significant differences ( $p < 0.05$ ) at 10 out of 32 positions.  
 854 Conversely, for *answer*→*question*, the Wrong group attends significantly more to the question at 22  
 855 out of 32 positions. This confirms that successful reasoning involves a systematic shift of attention  
 856 from the original question to the model’s own echoed representation.

#### 857 A.10 WORD-LEVEL ATTENTION CASE STUDY

859 To visualize which specific parts of the echo are attended to, we aggregated token-level attention into  
 860 word-level scores. Figure 7 shows a heatmap of attention from the reasoning trace to the echo prefix  
 861 for a representative correct solution to a GSM8K problem (“Janet’s ducks”). The model focuses  
 862 most intensely on the key numerical entities and constraints (e.g., “16”, “eggs”, “3”, “13”) within the  
 863 echo, rather than on function words. This supports the hypothesis that the echo serves as a semantic  
 864 anchor for critical problem details.

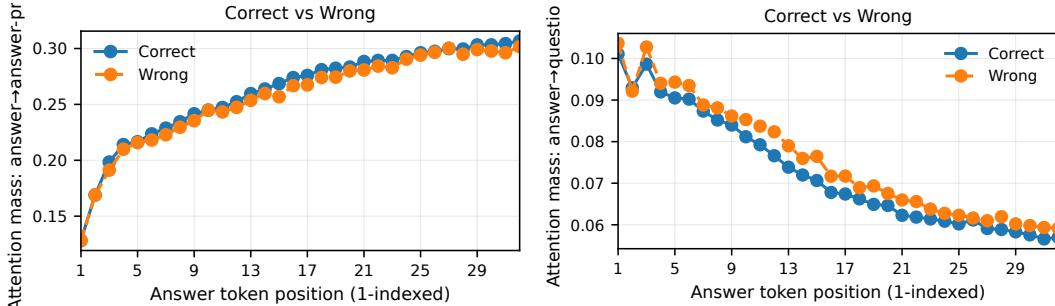


Figure 6: Token-wise average attention weights for the first 32 answer tokens. Left: Attention to Answer-Prefix. Right: Attention to Question. Shaded regions indicate standard error. The Correct group (blue) consistently attends more to the prefix, while the Wrong group (orange) attends more to the question.



Figure 7: Word-level attention heatmap from the reasoning trace to the echo prefix (mid-layers 7-18). Darker red indicates higher attention. The model selectively attends to key quantities (numbers of eggs, dollars) in the echoed prompt.

### A.11 INFORMATION FLOW ANALYSIS

We further investigated how information propagates through the model layers using an information flow route analysis. Figure 8 visualizes the attention-based routing of information for a single answer token (orange triangle at the top right). In the visualization, blue nodes represent question tokens, and green nodes represent echo/prefix tokens. In Correct traces, backward attribution paths from answer tokens repeatedly route through the echo-prefix tokens before reaching the rest of the prompt, whereas in Wrong traces these paths more often terminate in the question region or fail to reach the key numerals (e.g., idx 1151/1001 vs. 594). This suggests that the echo acts as a recurrent internal hub that integrates and refines information before it is used in the final generation.

### A.12 LOGISTIC REGRESSION ANALYSIS OF $\Delta\mathcal{L}$

To quantify the predictive power of the Echo Likelihood Gap ( $\Delta\mathcal{L}$ ) on reasoning correctness, we fitted a logistic regression model on the 1,319 GSM8K samples used in our analysis. We predicted the binary correctness outcome  $Y \in \{0, 1\}$  using  $\Delta\mathcal{L}$  and the length of the trimmed echo ( $L_{\text{echo}}$ ) as predictors:

$$\text{logit}(P(Y = 1)) = \beta_0 + \beta_1 \Delta\mathcal{L} + \beta_2 L_{\text{echo}} \quad (8)$$

The regression results (Table 9) indicate that  $\Delta\mathcal{L}$  is a statistically significant positive predictor of correctness ( $p \approx 0.022$ ), with a coefficient  $\beta_1 \approx 0.24$ . This implies that for every 1.0 nat/token increase in the likelihood gap, the odds of a correct answer increase by a factor of  $\exp(0.24) \approx 1.27$ , confirming that the model's probabilistic preference for the echo is meaningfully associated with task success.

### A.13 ANALYSIS OF EOP-PRESENT VS. EOP-ABSENT TRACES

To disentangle the effect of the Echo of Prompt (EOP) from general model capabilities, we compared traces where the model spontaneously produced an echo (EOP-present) versus those where it did not (EOP-absent). As shown in Table 10, the EOP-present group has a higher overall accuracy (63.8% vs 57.2%). Furthermore, even when controlling for the final outcome (Correct or Wrong), traces

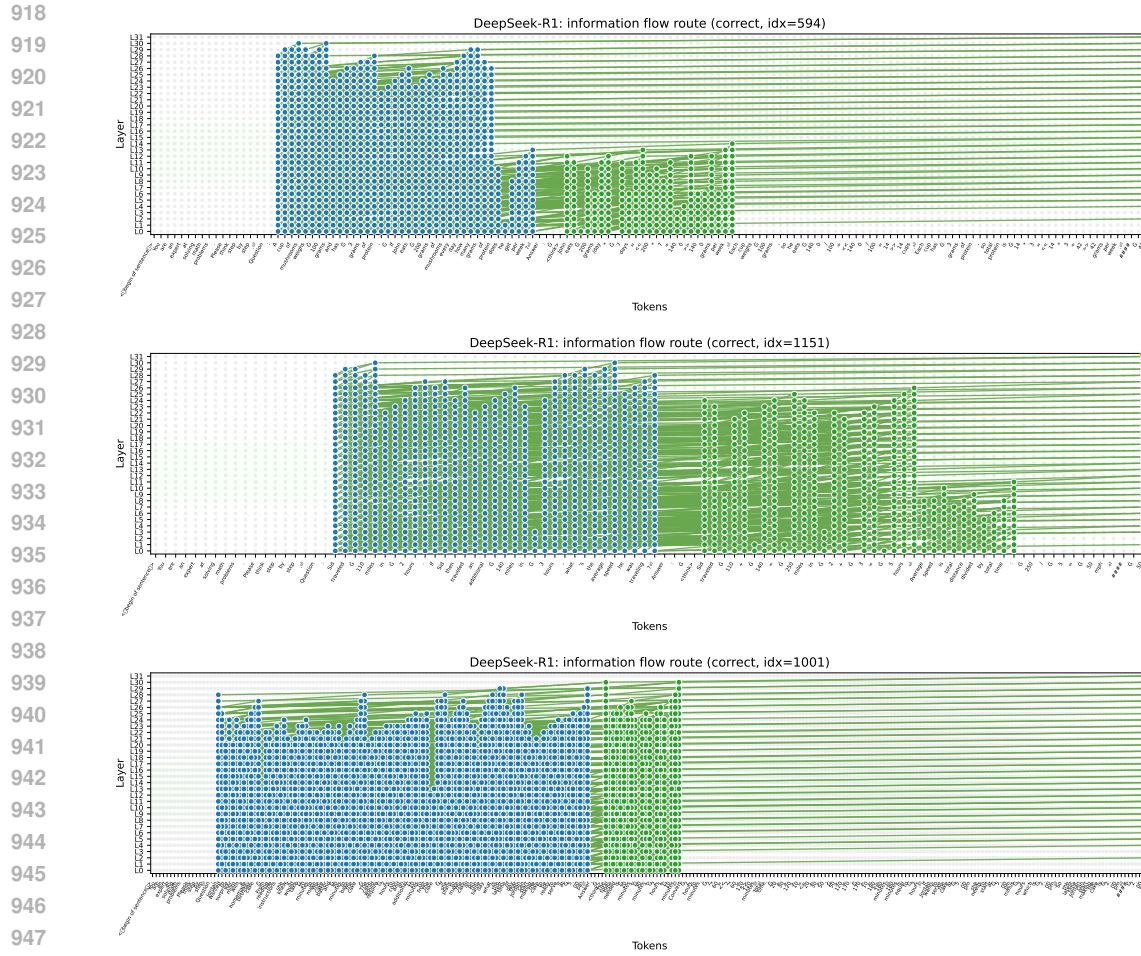


Figure 8: Information flow visualization for generated answer tokens across three different examples. The graphs show the primary attention routes through the model layers. In correct traces, information flows significantly through the Echo of Prompt (EOP) tokens in the middle layers, acting as a bridge between the question and the answer.

Table 9: Logistic regression results predicting correctness on GSM8K.

Predictor	Coefficient ( $\beta$ )	Std. Error	P-value
Intercept ( $\beta_0$ )	0.12	0.15	0.423
Echo Likelihood Gap ( $\beta_1$ )	0.24	0.11	<b>0.022</b>
Echo Length ( $\beta_2$ )	0.001	0.0005	0.089

with an EOP exhibit stronger attention refocusing (Answer  $\rightarrow$  Answer-Prefix attention) than those without. This suggests that the presence of an echo actively facilitates the attention mechanism that supports correct reasoning.

#### A.14 PROMPT TEMPLATES FOR TEACHER CoT AND EDITING

For reproducibility, we list the main prompt templates used to generate and edit teacher traces.

**Standard CoT generation prompt.** To obtain the initial pool of verified CoTs, we query `gpt-oss-120B` with a simple reasoning prompt that separates internal thinking from the final answer:

972  
 973 Table 10: Comparison of EOP-present and EOP-absent traces on GSM8K (DeepSeek-R1-Distill-  
 974 Llama-8B). EOP presence is associated with higher accuracy and stronger attention refocusing.

975 <b>Group</b>	976 <b>N</b>	977 <b>Accuracy (%)</b>	978 <b>Last-layer Ans→Ans-prefix Attn</b>
979 EOP-present	985	63.8	980 –
981 EOP-absent	334	57.2	982 –
<i>Conditioned on Outcome:</i>			
983 Correct, EOP-present	984 628	985 –	986 0.1374
987 Correct, EOP-absent	988 191	989 –	990 0.1328
991 Wrong, EOP-present	992 357	993 –	994 0.1046
995 Wrong, EOP-absent	996 143	997 –	998 0.0987

985 You are solving math problems. Structure your entire  
 986 thought process within a single pair of `<think>` and `</think>`  
 987 tags. After you've finished thinking, provide the final,  
 988 concise answer on a new line. The final answer should be a  
 989 plain value.

990 **Echo-insertion prompt for ED-SFT.** When the MLP probe predicts that a trace lacks an early  
 991 Echo of Prompt, we ask the teacher to minimally insert an echo-style opening that repeatedly brings  
 992 the question back into focus. The high-level instruction is:

993 You are solving math problems. Think out loud naturally.  
 994 To ensure you fully understand the problem, you must repeat  
 995 the question or key parts of it multiple times throughout  
 996 your reasoning process before you start solving. For  
 997 instance, you might re-read it to confirm details or after a  
 998 few steps of calculation to ensure you are on track. Start  
 999 by repeating the problem, then reason step by step. Wrap  
 1000 the entire internal thinking process with a single pair of  
 1001 `<think>` and `</think>` tags, and put the final answer after  
 1002 the thinking. The final answer should be a concise plain  
 1003 value (number if applicable). At the very beginning of  
 1004 your `<think>`, start with the following opening line and then  
 1005 continue the original reasoning. Do not change the final  
 1006 answer.

1007 The opening line is randomly chosen from a small set of naturalistic variants, e.g.:

- 1008 • Okay, let me see. The problem is asking: [QUESTION]
- 1009 • Alright, so the question is: [QUESTION]
- 1010 • Let me understand this problem. We have: [QUESTION]
- 1011 • So the problem states: [QUESTION]

1012 where [QUESTION] is replaced with the original GSM8K question text.

1013 **Echo-removal prompt for normal-SFT.** For traces where the MLP probe predicts the presence  
 1014 of an early EOP, we construct the normal-SFT counterpart by asking the teacher to remove the  
 1015 echo-prefix while preserving all later reasoning steps and the final answer:

1016 You are given a math question and a chain-of-thought  
 1017 solution that begins by repeating or paraphrasing the  
 1018 question. Rewrite the reasoning so that it no longer  
 1019 repeats the question at the beginning. Keep all subsequent  
 1020 reasoning steps and the final answer exactly the same. Do  
 1021 not change the logic or the final answer; only remove the  
 1022 initial echo segment. Wrap the internal thinking in `<think>`  
 1023 and `</think>` as before.

	Benchmark	Size / Format	Domain / Source	Description (from original papers)
1026	Hendrycks-MATH	12,500 problems; open-ended; step-by-step solutions	Competition mathematics (algebra, geometry, number theory, probability, etc.)	“A new dataset of 12,500 challenging competition mathematics problems. Each problem in MATH has a full step-by-step solution which can be used to teach models to generate answer derivations and explanations.” The evaluation follows the <code>hendrycks_math</code> group in the LM Evaluation Harness, which decomposes MATH into diverse subsets ( <code>_algebra</code> , <code>_counting_and_prob</code> , <code>_geometry</code> , <code>_intermediate_algebra</code> , <code>_num_theory</code> , <code>_prealgebra</code> , <code>_precalc</code> ), covering a wide range of mathematical skills beyond our training distribution.
1027	GSM8K	8.5K problems; free-form answers; word problems	Grade-school math word problems	“A dataset of 8.5K high quality linguistically diverse grade school math word problems. We find that even the largest transformer models fail to achieve high test performance, despite the conceptual simplicity of this problem distribution.”
1028	MathQA	37K problems; multiple choice	Multiple math domains; derived from AQuA operation programs	“A large-scale dataset of 37k English multiple-choice math word problems covering multiple math domain categories by modeling operation programs corresponding to word problems in the AQuA dataset.”

Table 11: Benchmarks used for our evaluation.

### A.15 ILLUSTRATIVE EXAMPLES OF ECHO-THEN-REASON PATTERN

We provide two illustrative examples as case studies to demonstrate the Echo-then-Reason pattern, see Figure 9

In these examples, we compare outputs from models with and without the initial echo phase. The echo-prefixed response (green) correctly identifies that the yellow marbles are identical and thus treats pairs of yellow marbles as a single unique group, leading to the correct count of 7. In contrast, the direct response (red) incorrectly treats the yellow marbles as distinguishable, overcounting the combinations and arriving at 12.

### A.16 SFT EVALUATION BENCHMARKS

In this section, we clarify our terminology regarding generalization. While we initially referred to performance on MathQA and Hendrycks-MATH as “out-of-domain generalization”, we acknowledge that all tested datasets fall within the broader domain of mathematics. However, they represent significant **distributional shifts** in terms of difficulty, topic coverage, and problem format compared to the GSM8K training set. Therefore, we adopt the terms **distributional generalization** or **robustness to distribution shift** to more precisely describe these experiments. Table 11 summarizes the characteristics of each benchmark, highlighting these differences.

1080  
1081  
1082  
1083  
1084  
1085  
1086  
1087  
1088  
1089  
1090

Question

*Tom has a red marble, a green marble, a blue marble, and three identical yellow marbles. How many different groups of two marbles can Tom choose?*

1091  
1092  
1093  
1094  
1095  
1096  
1097  
1098  
1099  
1100  
1101  
1102  
1103  
1104  
1105  
1106  
1107  
1108  
1109  
1110  
1111  
1112  
1113  
1114  
1115  
1116  
1117  
1118  
1119  
1120  
1121  
1122  
1123  
1124  
1125  
1126  
1127  
1128  
1129  
1130  
1131  
1132  
1133

Model output with Echo-then-Reason (Correct)

*<think> The problem asks: "Tom has a red marble, a green marble, a blue marble, and three identical yellow marbles. How many different groups of two marbles can Tom choose?" First, repeat the key details: Tom has one red marble, one green marble, one blue marble, and three yellow marbles that are identical. He wants to form groups of two marbles. To avoid counting identical groups multiple times, consider the distinct colors: red, green, blue, and yellow. There are four distinct colors. Now, let's count the pairs systematically to avoid overcounting: - Red can be paired with green, blue, or yellow, giving 3 pairs. - Green can be paired with blue or yellow (red-green is already counted), giving 2 new pairs. - Blue can be paired with yellow (red-blue and green-blue are already counted), giving 1 new pair. - Two yellow marbles can be paired. Since they are identical, this is 1 more pair. Adding these up:  $3 + 2 + 1 + 1 = 7$  pairs. Thus, the number of different groups of two marbles Tom can choose is 7. </think> The answer is 7.*

Model output without Echo (Incorrect)

*To determine the number of different groups of two marbles that Tom can choose, we need to consider all possible combinations of the marbles he has. Tom has a red marble (R), a green marble (G), a blue marble (B), and three identical yellow marbles (Y). The total number of marbles is 6, but since the yellow marbles are identical, we need to account for this in our combinations. We can use the combination formula  $\binom{n}{k}$  to find the number of ways to choose 2 marbles out of 6, but we need to adjust for the identical yellow marbles. Let's break it down: 1. Choose 2 marbles from the 3 yellow marbles: This is not possible since we can't choose 2 identical marbles from a set of identical items. 2. Choose 1 yellow marble and 1 non-yellow marble: We can choose 1 yellow marble in  $\binom{3}{1} = 3$  ways, and we can choose 1 non-yellow marble in  $\binom{3}{1} = 3$  ways. So, the total number of ways to choose 1 yellow and 1 non-yellow marble is  $3 \times 3 = 9$ . 3. Choose 2 non-yellow marbles: We can choose 2 non-yellow marbles from the 3 non-yellow marbles (R, G, B) in  $\binom{3}{2} = 3$  ways. Adding these up, the total number of different groups of two marbles that Tom can choose is  $0 + 9 + 3 = 12$ . The number of different groups of two marbles that Tom can choose is 12.*

Figure 9: Qwen3-8B-Base model outputs illustrating the Echo-then-Reason pattern. Top: the math question from Hendrycks-MATH dataset. Middle (green): Qwen3-8B-ED-SFT's correct answer. Bottom (red): Qwen3-8B-Base that immediately jumps to calculation without echoing, resulting in the wrong answer.