

A Implement Details

In this section, we will introduce the implementation details of our approach, focusing on the data selection process for computing the steering direction to mitigate overthinking in each model and the specific configurations of the baseline methods used for comparison.

A.1 Data Selection for Direction Computation

To construct representative datasets for computing the steering direction to mitigate overthinking, we begin by randomly sampling 20k questions from the OpenMathInstruct-2 training set [34]. For each model, we generate five independent responses per question using the official sampling configuration: a temperature of 0.6, top-p of 0.95, and a maximum length of 16k tokens. These responses form the basis for constructing two model-specific datasets, the **Redundant set** ($D_{\text{redundant}}$) and the **Concise set** (D_{concise}), as described below:

- **Redundant set** ($D_{\text{redundant}}$): This dataset includes questions where all five responses exceed 16k tokens without terminating and contain more than 20 times of hesitation keywords (e.g., “wait”, “alternatively”, etc.). To capture meaningful overthinking behavior, we process the responses using the following template, truncating the response at the occurrence of the hesitation keyword:

```
<|begin_of_sentence|><|User|>{instruction}<|Assistant|><think>\n
{partial_response}{hesitation keyword}
```

The truncation at a hesitation keyword is reasonable because overthinking typically emerges after a certain point in the response, rather than immediately upon encountering the question. Moreover, through activation visualization, we observe no significant differences in the activation patterns of different hesitation keywords. Thus we choose “wait” as a consistent marker here.

- **Concise set** (D_{concise}): This dataset includes questions where all five responses are under 1k tokens and contain none of the hesitation keywords. The template for these responses includes only the instruction without the response, as they inherently represent concise and focused outputs:

```
<|begin_of_sentence|><|User|>{instruction}<|Assistant|><think>\n
```

These selection criteria ensure that $D_{\text{redundant}}$ captures responses exhibiting excessive verbosity and hesitation, while D_{concise} represents efficient and direct responses, providing a clear contrast for computing the steering direction representing overthinking. To ensure high-quality data, we retain only 500 samples for each dataset after applying the selection criteria and double checking. For computing the steering direction, we follow [47] to sample 100 samples from each dataset and employ the IsolationForest algorithm to filter out outliers. For manifold subspace estimation, we utilize the entire set of 500 samples from each dataset to capture the full representational structure.

A.2 Baseline Methods

As stated in Sec. 5.1, we select two latest baselines, Dynasor [15] and SEAL [8], for their ability to preserve the original accuracy in reasoning tasks. Below, we detail the specific settings for them:

General Setting. All large reasoning models adopt the official recommended settings with a temperature of 0.6, top-p of 0.95, and a maximum length of 16k tokens.

Dynasor. We adopt the official settings for Dynasor. The configuration probes the model every 32 tokens with a “Probe-In-The-Middle” technique and injects a “Final Answer” prompt at each iteration to ensure complete solutions upon early termination. Generation stops when the Certindex metric (\tilde{H}) exceeds a predefined confidence threshold. To be aware, Dynasor’s early stopping often omits the problem-solving process in the final answer, which is impractical for real-world applications. Thus, we require the model to provide a complete solution in the final answer upon stopping.

SEAL. We adopt the official settings for SEAL [8], using 1k training samples from the Math dataset [18] to extract the reasoning steering vector. Reasoning processes are segmented into thoughts using “\n\n” delimiters, classified as execution, reflection, or transition via keyword-based rules

(e.g., “Alternatively” for transition, “Wait” for reflection). The steering vector is computed at layer 20 as $S = \bar{H}_E - \bar{H}_{RT}$, where \bar{H}_E and \bar{H}_{RT} are average representations of execution and reflection/transition thoughts, respectively. During greedy decoding, hidden states of “\n\n” tokens at layer 20 are adjusted as $\bar{H} = H + 1.0 \cdot S$.

B Proofs

B.1 Proof of Theorem 4.1

Proof. We derive the expected noise norm of the interference component \mathbf{r}_{other} , the part of the overthinking direction $\mathbf{r}^{(l*)}$ in the orthogonal complement \mathcal{M}^\perp of the low-dimensional manifold \mathcal{M} . The theorem states:

$$\mathbb{E}[\|\mathbf{r}_{other}\|_2^2] = \text{tr} \left((\mathbf{I} - \mathbf{P}_{\mathcal{M}}) \boldsymbol{\Sigma}_{\text{noise}}^{(l)} \right), \quad \boldsymbol{\Sigma}_{\text{noise}}^{(l)} = \frac{\mathbf{C}^{(l)}}{|D_{\text{redundant}}|} + \frac{\mathbf{C}^{(l)}}{|D_{\text{concise}}|},$$

where $\mathbf{P}_{\mathcal{M}} = \mathbf{U}^{(l)}[:, 1:k] (\mathbf{U}^{(l)}[:, 1:k])^\top$, and $\mathbf{U}^{(l)}[:, 1:k]$ are the top- k principal components of the activation covariance $\mathbf{C}^{(l)}$. We build on prior findings that \mathcal{M} is low-dimensional, identified via PCA on $D_{\text{reasoning}} = D_{\text{redundant}} \cup D_{\text{concise}}$, with $k = 10$ capturing over 70% of the variance, validating the linear manifold assumption.

Step 1: Define the overthinking direction $\mathbf{r}^{(l*)}$. Per Eq. (2), $\mathbf{r}^{(l*)} = \mathbf{r}_{overthinking} + \mathbf{r}_{other}$, where $\mathbf{r}_{overthinking} \in \mathcal{M}$ captures the shift between redundant and concise reasoning, and $\mathbf{r}_{other} \in \mathcal{M}^\perp$ is interference. We model:

$$\mathbf{r}^{(l*)} = \frac{1}{|D_{\text{redundant}}|} \sum_{x_i \in D_{\text{redundant}}} \mathbf{h}^{(l)}(x_i) - \frac{1}{|D_{\text{concise}}|} \sum_{x_i \in D_{\text{concise}}} \mathbf{h}^{(l)}(x_i).$$

Assume activations $\mathbf{h}^{(l)}(x_i) \sim \mathcal{N}(\boldsymbol{\mu}_{\text{set}}, \mathbf{C}^{(l)})$, with $\boldsymbol{\mu}_{\text{redundant}}$ or $\boldsymbol{\mu}_{\text{concise}}$ for each dataset, and $\mathbf{C}^{(l)}$ estimated over $D_{\text{reasoning}}$. The covariance is:

$$\mathbb{E}[\mathbf{r}^{(l*)} \mathbf{r}^{(l*)\top}] = \frac{\mathbf{C}^{(l)}}{|D_{\text{redundant}}|} + \frac{\mathbf{C}^{(l)}}{|D_{\text{concise}}|}.$$

Step 2: Define \mathcal{M} and derive $\mathbf{I} - \mathbf{P}_{\mathcal{M}}$. The manifold \mathcal{M} is spanned by the top- k eigenvectors of $\mathbf{C}^{(l)} = \frac{1}{N-1} \mathbf{A}^{(l)} (\mathbf{A}^{(l)} - \bar{\mathbf{A}}^{(l)})^\top$, where $\mathbf{A}^{(l)} = [\mathbf{h}^{(l)}(x_1), \dots, \mathbf{h}^{(l)}(x_N)]$, and $\bar{\mathbf{A}}^{(l)} = \frac{1}{N} \sum_{i=1}^N \mathbf{h}^{(l)}(x_i)$. The eigendecomposition $\mathbf{C}^{(l)} = \mathbf{U}^{(l)} \boldsymbol{\Lambda}^{(l)} (\mathbf{U}^{(l)})^\top$ yields $\mathbf{U}^{(l)}[:, 1:k]$, and:

$$\mathbf{P}_{\mathcal{M}} = \mathbf{U}^{(l)}[:, 1:k] (\mathbf{U}^{(l)}[:, 1:k])^\top.$$

The projection onto \mathcal{M}^\perp is $\mathbf{I} - \mathbf{P}_{\mathcal{M}}$, as it removes the \mathcal{M} -component. Since $\mathbf{U}^{(l)}[:, 1:k]$ is orthonormal, $\mathbf{P}_{\mathcal{M}}$ is idempotent and symmetric, so:

$$(\mathbf{I} - \mathbf{P}_{\mathcal{M}})^2 = \mathbf{I} - \mathbf{P}_{\mathcal{M}}, \quad (\mathbf{I} - \mathbf{P}_{\mathcal{M}})^\top = \mathbf{I} - \mathbf{P}_{\mathcal{M}}.$$

PCA’s linear basis ensures \mathcal{M}^\perp captures the $d - k$ dimensions of noise, critical when $d \gg k$.

Step 3: Define \mathbf{r}_{other} . Since $\mathbf{r}_{overthinking} \in \mathcal{M}$, the interference is:

$$\mathbf{r}_{other} = (\mathbf{I} - \mathbf{P}_{\mathcal{M}}) \mathbf{r}^{(l*)}.$$

This isolates noise in \mathcal{M}^\perp , which disrupts normal abilities due to high-dimensional computation.

Step 4: Compute the squared norm. Calculate:

$$\|\mathbf{r}_{other}\|_2^2 = [(\mathbf{I} - \mathbf{P}_{\mathcal{M}}) \mathbf{r}^{(l*)}]^\top (\mathbf{I} - \mathbf{P}_{\mathcal{M}}) \mathbf{r}^{(l*)} = \mathbf{r}^{(l*)\top} (\mathbf{I} - \mathbf{P}_{\mathcal{M}}) \mathbf{r}^{(l*)},$$

using the idempotence of $\mathbf{I} - \mathbf{P}_{\mathcal{M}}$.

Step 5: Take the expectation. Compute:

$$\mathbb{E}[\|\mathbf{r}_{other}\|_2^2] = \mathbb{E}[\mathbf{r}^{(l*)\top} (\mathbf{I} - \mathbf{P}_{\mathcal{M}}) \mathbf{r}^{(l*)}] = \text{tr}((\mathbf{I} - \mathbf{P}_{\mathcal{M}}) \mathbb{E}[\mathbf{r}^{(l*)} \mathbf{r}^{(l*)\top}]).$$

Substitute:

$$\mathbb{E}[\mathbf{r}^{(l*)} \mathbf{r}^{(l*)\top}] = \boldsymbol{\Sigma}_{\text{noise}}^{(l)} = \frac{\mathbf{C}^{(l)}}{|D_{\text{redundant}}|} + \frac{\mathbf{C}^{(l)}}{|D_{\text{concise}}|}.$$

Thus:

$$\mathbb{E}[\|\mathbf{r}_{other}\|_2^2] = \text{tr} \left((\mathbf{I} - \mathbf{P}_{\mathcal{M}}) \boldsymbol{\Sigma}_{\text{noise}}^{(l)} \right).$$

□

B.2 Proof of Theorem 4.2

Proof. We derive the mean activation shift $\Delta\boldsymbol{\mu}^{(l)}$ at layer l due to the intervention (applied as in Eq. (4).) along the overthinking direction $\mathbf{r}^{(l*)}$, showing its norm is proportional to $\alpha\|\mathbf{r}_{other}\|_2$, and establish the layer-wise amplification of the shift at layer $l+1$. The theorem builds on Theorem 4.1.

Step 1: Derive the mean activation shift. The intervention at layer l is:

$$\mathbf{h}^{(l)'}(x_i) = \mathbf{h}^{(l)}(x_i) - \alpha[(\mathbf{r}^{(l*)})^\top \mathbf{h}^{(l)}(x_i)]\mathbf{r}^{(l*)},$$

with $\alpha > 0$. The mean activation before intervention is:

$$\boldsymbol{\mu}^{(l)} = \frac{1}{N} \sum_{i=1}^N \mathbf{h}^{(l)}(x_i),$$

and post-intervention:

$$\boldsymbol{\mu}^{(l)'} = \frac{1}{N} \sum_{i=1}^N \mathbf{h}^{(l)'}(x_i) = \frac{1}{N} \sum_{i=1}^N \left(\mathbf{h}^{(l)}(x_i) - \alpha[(\mathbf{r}^{(l*)})^\top \mathbf{h}^{(l)}(x_i)]\mathbf{r}^{(l*)} \right).$$

Compute:

$$\boldsymbol{\mu}^{(l)'} = \boldsymbol{\mu}^{(l)} - \alpha \frac{1}{N} \sum_{i=1}^N [(\mathbf{r}^{(l*)})^\top \mathbf{h}^{(l)}(x_i)]\mathbf{r}^{(l*)}.$$

The mean shift is:

$$\Delta\boldsymbol{\mu}^{(l)} = \boldsymbol{\mu}^{(l)'} - \boldsymbol{\mu}^{(l)} = -\alpha \frac{1}{N} \sum_{i=1}^N [(\mathbf{r}^{(l*)})^\top \mathbf{h}^{(l)}(x_i)]\mathbf{r}^{(l*)},$$

matching the first part of Eq. (7).

Step 2: Decompose the shift and isolate \mathbf{r}_{other} contribution. Since $\mathbf{r}^{(l*)} = \mathbf{r}_{\mathcal{M}} + \mathbf{r}_{other}$ with $\mathbf{r}_{\mathcal{M}} \perp \mathbf{r}_{other}$, we can decompose:

$$(\mathbf{r}^{(l*)})^\top \mathbf{h}^{(l)}(x_i) = (\mathbf{r}_{\mathcal{M}})^\top \mathbf{h}^{(l)}(x_i) + (\mathbf{r}_{other})^\top \mathbf{h}^{(l)}(x_i).$$

Thus:

$$\begin{aligned} \Delta\boldsymbol{\mu}^{(l)} &= -\alpha \frac{1}{N} \sum_{i=1}^N [(\mathbf{r}_{\mathcal{M}})^\top \mathbf{h}^{(l)}(x_i) + (\mathbf{r}_{other})^\top \mathbf{h}^{(l)}(x_i)](\mathbf{r}_{\mathcal{M}} + \mathbf{r}_{other}) \\ &= -\alpha \frac{1}{N} \sum_{i=1}^N [(\mathbf{r}_{\mathcal{M}})^\top \mathbf{h}^{(l)}(x_i)]\mathbf{r}_{\mathcal{M}} - \alpha \frac{1}{N} \sum_{i=1}^N [(\mathbf{r}_{\mathcal{M}})^\top \mathbf{h}^{(l)}(x_i)]\mathbf{r}_{other} \\ &\quad - \alpha \frac{1}{N} \sum_{i=1}^N [(\mathbf{r}_{other})^\top \mathbf{h}^{(l)}(x_i)]\mathbf{r}_{\mathcal{M}} - \alpha \frac{1}{N} \sum_{i=1}^N [(\mathbf{r}_{other})^\top \mathbf{h}^{(l)}(x_i)]\mathbf{r}_{other}. \end{aligned}$$

Let:

$$s_{\mathcal{M}} = \frac{1}{N} \sum_{i=1}^N [(\mathbf{r}_{\mathcal{M}})^\top \mathbf{h}^{(l)}(x_i)], \quad s_{other} = \frac{1}{N} \sum_{i=1}^N [(\mathbf{r}_{other})^\top \mathbf{h}^{(l)}(x_i)].$$

Then:

$$\Delta\boldsymbol{\mu}^{(l)} = -\alpha s_{\mathcal{M}}\mathbf{r}_{\mathcal{M}} - \alpha s_{\mathcal{M}}\mathbf{r}_{other} - \alpha s_{other}\mathbf{r}_{\mathcal{M}} - \alpha s_{other}\mathbf{r}_{other}.$$

Since $\mathbf{r}_{\mathcal{M}} \perp \mathbf{r}_{other}$:

$$\begin{aligned} \|\Delta\boldsymbol{\mu}^{(l)}\|_2^2 &= \alpha^2 \|s_{\mathcal{M}}\mathbf{r}_{\mathcal{M}} + s_{other}\mathbf{r}_{\mathcal{M}}\|_2^2 + \alpha^2 \|s_{\mathcal{M}}\mathbf{r}_{other} + s_{other}\mathbf{r}_{other}\|_2^2 \\ &= \alpha^2 (s_{\mathcal{M}} + s_{other})^2 \|\mathbf{r}_{\mathcal{M}}\|_2^2 + \alpha^2 (s_{\mathcal{M}} + s_{other})^2 \|\mathbf{r}_{other}\|_2^2. \end{aligned}$$

Let $s = s_{\mathcal{M}} + s_{other}$. Then:

$$\|\Delta\boldsymbol{\mu}^{(l)}\|_2 = \alpha |s| \sqrt{\|\mathbf{r}_{\mathcal{M}}\|_2^2 + \|\mathbf{r}_{other}\|_2^2}.$$

By Theorem 4.1, when the error component \mathbf{r}_{other} is present (i.e., $\|\mathbf{r}_{other}\|_2 > 0$), it contributes to the total norm. The dominant term depends on the relative magnitudes of $\|\mathbf{r}_{\mathcal{M}}\|_2$ and $\|\mathbf{r}_{other}\|_2$. Assuming $\mathbf{h}^{(l)}(x_i) \sim \mathcal{N}(\boldsymbol{\mu}_{\text{set}}, \mathbf{C}^{(l)})$ and $|s|$ is a positive constant, we obtain:

$$\|\Delta\boldsymbol{\mu}^{(l)}\|_2 \propto \alpha\|\mathbf{r}_{other}\|_2,$$

when $\|\mathbf{r}_{other}\|_2$ dominates, completing Eq. (7).

Step 3: Derive the layer-wise amplification from \mathbf{r}_{other} . For layer $l + 1$, the activation is:

$$\mathbf{h}^{(l+1)}(x_i) = \sigma\left(\mathbf{W}^{(l+1)} \text{Attn}(\mathbf{h}^{(l)}(x_i))\right),$$

and post-intervention:

$$\mathbf{h}^{(l+1)'}(x_i) = \sigma\left(\mathbf{W}^{(l+1)} \text{Attn}(\mathbf{h}^{(l)'}(x_i))\right),$$

where $\mathbf{W}^{(l+1)}$ combines MLP and attention weights, Attn is the attention mechanism, and σ is GeLU. The mean shift is:

$$\Delta\boldsymbol{\mu}^{(l+1)} = \boldsymbol{\mu}^{(l+1)'} - \boldsymbol{\mu}^{(l+1)}, \quad \boldsymbol{\mu}^{(l+1)'} = \frac{1}{N} \sum_{i=1}^N \mathbf{h}^{(l+1)'}(x_i), \quad \boldsymbol{\mu}^{(l+1)} = \frac{1}{N} \sum_{i=1}^N \mathbf{h}^{(l+1)}(x_i).$$

To isolate the \mathbf{r}_{other} contribution, decompose the single-input shift at layer l :

$$\begin{aligned} \Delta\mathbf{h}^{(l)}(x_i) &= \mathbf{h}^{(l)'}(x_i) - \mathbf{h}^{(l)}(x_i) \\ &= -\alpha[(\mathbf{r}^{(l*)})^\top \mathbf{h}^{(l)}(x_i)]\mathbf{r}^{(l*)} \\ &= -\alpha[(\mathbf{r}_{\mathcal{M}})^\top \mathbf{h}^{(l)}(x_i) + (\mathbf{r}_{other})^\top \mathbf{h}^{(l)}(x_i)](\mathbf{r}_{\mathcal{M}} + \mathbf{r}_{other}). \end{aligned}$$

The norm is:

$$\begin{aligned} \|\Delta\mathbf{h}^{(l)}(x_i)\|_2 &= \alpha|(\mathbf{r}^{(l*)})^\top \mathbf{h}^{(l)}(x_i)|\|\mathbf{r}^{(l*)}\|_2 \\ &= \alpha|(\mathbf{r}_{\mathcal{M}})^\top \mathbf{h}^{(l)}(x_i) + (\mathbf{r}_{other})^\top \mathbf{h}^{(l)}(x_i)|\sqrt{\|\mathbf{r}_{\mathcal{M}}\|_2^2 + \|\mathbf{r}_{other}\|_2^2}. \end{aligned}$$

The component from \mathbf{r}_{other} can be isolated by considering its contribution:

$$\|\Delta\mathbf{h}^{(l)}(x_i)\|_2 \geq \alpha|(\mathbf{r}_{other})^\top \mathbf{h}^{(l)}(x_i)|\|\mathbf{r}_{other}\|_2,$$

when the \mathbf{r}_{other} term is significant. Propagate to layer $l + 1$:

$$\Delta\mathbf{h}^{(l+1)}(x_i) = \mathbf{h}^{(l+1)'}(x_i) - \mathbf{h}^{(l+1)}(x_i) \approx \sigma'\left(\mathbf{W}^{(l+1)} \text{Attn}'(\mathbf{h}^{(l)}(x_i))\Delta\mathbf{h}^{(l)}(x_i)\right),$$

where Attn' and σ' are the Jacobians of attention and GeLU. The attention softmax and GeLU have minimum amplification factors $\gamma_{\text{attn}}, \gamma_\sigma > 0$, and the linear transformation by $\mathbf{W}^{(l+1)}$ satisfies:

$$\|\mathbf{W}^{(l+1)}\mathbf{x}\|_2 \geq \sigma_{\min}(\mathbf{W}^{(l+1)})\|\mathbf{x}\|_2.$$

Thus:

$$\|\Delta\mathbf{h}^{(l+1)}(x_i)\|_2 \geq \gamma_{\text{attn}}\gamma_\sigma\sigma_{\min}(\mathbf{W}^{(l+1)})\|\Delta\mathbf{h}^{(l)}(x_i)\|_2.$$

Focusing on the \mathbf{r}_{other} contribution:

$$\|\Delta\mathbf{h}^{(l+1)}(x_i)\|_2 \geq \gamma_{\text{attn}}\gamma_\sigma\sigma_{\min}(\mathbf{W}^{(l+1)})\alpha|(\mathbf{r}_{other})^\top \mathbf{h}^{(l)}(x_i)|\|\mathbf{r}_{other}\|_2.$$

The mean shift norm is:

$$\|\Delta\boldsymbol{\mu}^{(l+1)}\|_2 = \left\| \frac{1}{N} \sum_{i=1}^N \Delta\mathbf{h}^{(l+1)}(x_i) \right\|_2.$$

Assume the layer-wise propagation amplifies the previous shift by $\gamma > 1$, reflecting attention and non-linear effects across layers. Combining the amplification of the existing shift and the new \mathbf{r}_{other} contribution:

$$\|\Delta\boldsymbol{\mu}^{(l+1)}\|_2 \geq \gamma\|\Delta\boldsymbol{\mu}^{(l)}\|_2 + \alpha\gamma_{\text{attn}}\gamma_\sigma\sigma_{\min}(\mathbf{W}^{(l+1)})|(\mathbf{r}_{other})^\top \mathbf{h}^{(l)}(x_i)|\|\mathbf{r}_{other}\|_2,$$

matching Eq. (8). This shows that the \mathbf{r}_{other} component causes layer-wise amplification through both the accumulated shift (first term) and the direct contribution at each layer (second term).

Step 4: Analyze the amplification mechanism. The amplification factors $\gamma > 1$, $\gamma_{\text{attn}}, \gamma_\sigma > 0$, and non-zero $\sigma_{\min}(\mathbf{W}^{(l+1)})$ ensure that perturbations from \mathbf{r}_{other} grow across layers. The first term $\gamma\|\Delta\boldsymbol{\mu}^{(l)}\|_2$ represents the propagation of accumulated shift, while the second term represents the fresh perturbation introduced at layer $l + 1$ due to \mathbf{r}_{other} . This dual mechanism ensures the shift grows across layers, disrupting the model's normal abilities. \square

C Hyperparameter Tuning

In this section, we present the results of tuning the intervention strength α across four models: DeepSeek-R1-Distill-Qwen-1.5B, DeepSeek-R1-Distill-Qwen-7B, DeepSeek-R1-Distill-Llama-8B, and DeepSeek-R1-Distill-Qwen-14B on MATH500 [18]. As shown in Fig. 7, to achieve an optimal balance between efficiency and accuracy, we ultimately select $\alpha = 0.7$ for R1-1.5B, $\alpha = 0.3$ for R1-7B, $\alpha = 0.5$ for R1-8B, and $\alpha = 0.3$ for R1-14B.

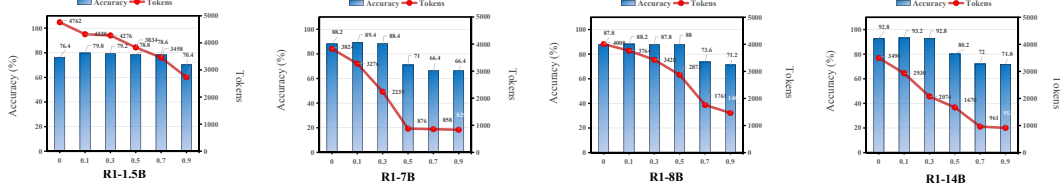


Figure 7: Impact of intervention strength α on the token reduction and accuracy of R1-1.5B, R1-7B, R1-8B, and R1-14B on the MATH500 dataset

D Layer Selection for Manifold Steering

The selection of intervention layers is critical for the effectiveness of Manifold Steering. We conduct a layer-wise analysis across multiple model sizes to determine the optimal intervention points. As shown in the tables below, we evaluate the performance across different layers by measuring accuracy and tokens on the MATH500. The results demonstrate that later layers consistently achieve better performance: Layer 27 for R1-1.5B and R1-7B, Layer 31 for R1-8B, and Layer 47 for R1-14B.

Table 2: Layer-wise performance analysis for R1-1.5B on MATH500.

	Vanilla	Layer 1	Layer 5	Layer 10	Layer 15	Layer 20	Layer 25	Layer 27
Accuracy (%)	76.4	76.6	77.0	76.4	57.6	74.8	67.4	78.6
# Tokens	4762	4472	4434	4223	1469	3930	1179	3458

Table 3: Layer-wise performance analysis for R1-7B on MATH500.

	Vanilla	Layer 1	Layer 5	Layer 10	Layer 15	Layer 20	Layer 25	Layer 27
Accuracy (%)	88.2	88.4	88.0	88.2	84.4	80.6	72.2	88.4
# Tokens	3824	3685	3665	3701	2713	1906	1070	2239

Table 4: Layer-wise performance analysis for R1-8B on MATH500.

	Vanilla	Layer 1	Layer 5	Layer 10	Layer 15	Layer 20	Layer 25	Layer 30	Layer 31
Accuracy (%)	87.8	87.2	87.8	88.2	75.6	86.4	71.8	87.6	88.0
# Tokens	4009	3896	3820	3654	2950	3280	1856	2975	2873

Table 5: Layer-wise performance analysis for R1-14B on MATH500.

	Vanilla	Layer 1	Layer 5	Layer 10	Layer 15	Layer 20	Layer 25	Layer 30	Layer 35	Layer 40	Layer 45	Layer 47
Accuracy (%)	92.8	92.4	92.4	92.0	92.2	92.6	89.8	80.4	87.4	84.6	82.4	92.8
# Tokens	3496	3384	3420	3095	2958	2857	2398	1814	2207	1836	1625	2074

E Time Latency Analysis

In this section, we analyze the time latency for the DeepSeek-R1-Distill-Qwen-7B model on the Math500 dataset [18], comparing our approach with Dynasor [15] and SEAL [8]. All experiments are conducted on an Ubuntu 22.04 system with A800 GPUs. We find that Dynasor exhibits the

significantly longest time latency, which is reasonable due to its frequent probing of intermediate states and its unsuitability for parallel processing of large reasoning models. For SEAL, although both SEAL and our method introduce negligible additional computational cost, SEAL’s token reduction is less effective than ours, resulting in higher time latency.

Table 6: Average Time Latency on Math500 for different overthinking-mitigation methods in R1-7B.

Methods	Original	Dynasor	SEAL	Ours
Time Latency (s)	1.74	39.89	1.37	1.05