重复输出问题探究 & ThinkEdit的看法

密言

2025.4.9

重复输出问题探究

- 使用WISE编辑模型时，非常容易出现重复输出的问题
- 并且相似的问题会混淆回答（且重复）

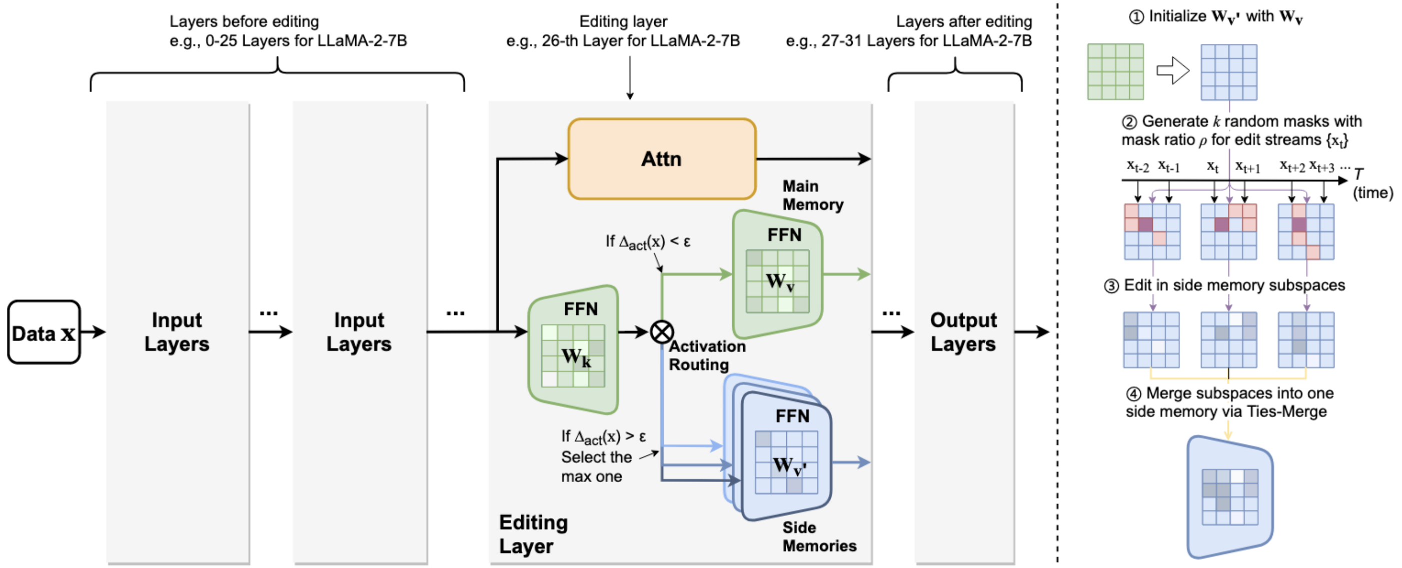
```
"prompt": "who won the men's ice skating 2018?",
"target_new": "Yuzuru Hanyu",

"rewrite_gen_content": "Terry Gannon Gannon is an American sportscaster and Yuz Yuz Yuz Yuz Yuz Yuz Yuz Yuz Yuz Yuz Yuz Yuz Yuz Yuz Yuz Yuz Yuz Yuz Yuz",
```

```
"prompt": "who are the nbc olympic ice skating commentators?",
"target_new": "Terry Gannon",

"rewrite_gen_content": "Terry Gannon Gannon is Gannon Gannon Gannon Gannon Gannon Gannon Gannon Gannon Gannon Gannon Gannon Gannon Gannon Gannon Gannon Gannon Gannon Gannon Gannon Gannon Gannon",
```

重复输出问题分析



(a) Workflow Overview with Knowledge Routing (b) Knowledge Sharding and Merging

Figure 2: **Overview of WISE.** Side memory (in blue) and main memory (in green) store edited and pretrained knowledge, respectively. Note: during inference, if WISE-Retrieve, the activation routing will retrieve and select one side memory with maximal activation score.

k – subspaces num

ρ – mask ratios

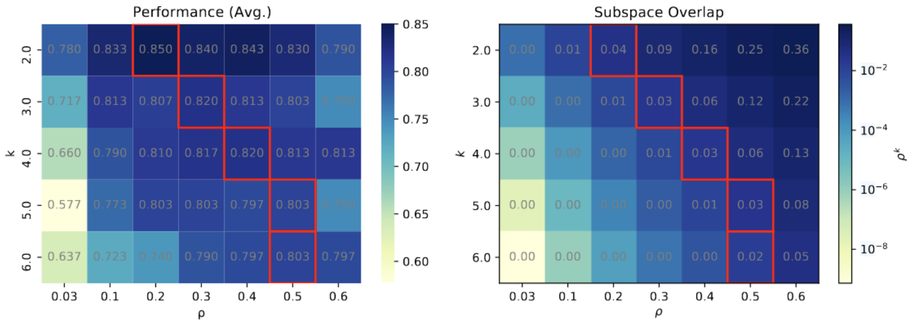


Figure 5: **Analysis of different mask ratios ρ and subspaces k for WISE.** Left: Avg. performance of Rel., Gen., and Loc.; Right: the subspace overlap probability in Theorem 2.1. ZsRE, `LLaMA-2-7B`.

k, ρ不同配比下，每行表现最好时的子空间重合率为0.02-0.04

由于每个问题对应的激活是稀疏的，很有可能相似的知识间的编辑仍然会互相干扰

使用repetition penalty

使用repetition penalty后，重复现象显著减轻，但是后续回答依然缺乏逻辑，因此核心问题可能依然在编辑方法上。

```
generated_ids = wise_editor.generate(
input_ids,
max_length=100,
temperature=0.7,
top_p=0.9,
repetition_penalty=1.5,
)
```

ThinkEdit: Interpretable Weight Editing to Mitigate Overly Short Thinking in Reasoning Models

Important Sections:

2. Unexpectedly Low Accuracy in Short Reasoning Cases	**发现问题/现象**
3. Understanding How Hidden Representations Affect Reasoning Length	**探究影响推理长度的机制**
3.2 Extracting Reasoning Length Directions	探究方法：Extract Directions
3.3 Effects of Reasoning-Length Direction	探究过程：Global 到 Layerwise
4. ThinkEdit: Mitigate Overly Short Reasoning through Weight Editing	**提出编辑方法（解决问题）**

2. Unexpectedly Low Accuracy in Short Reasoning Cases

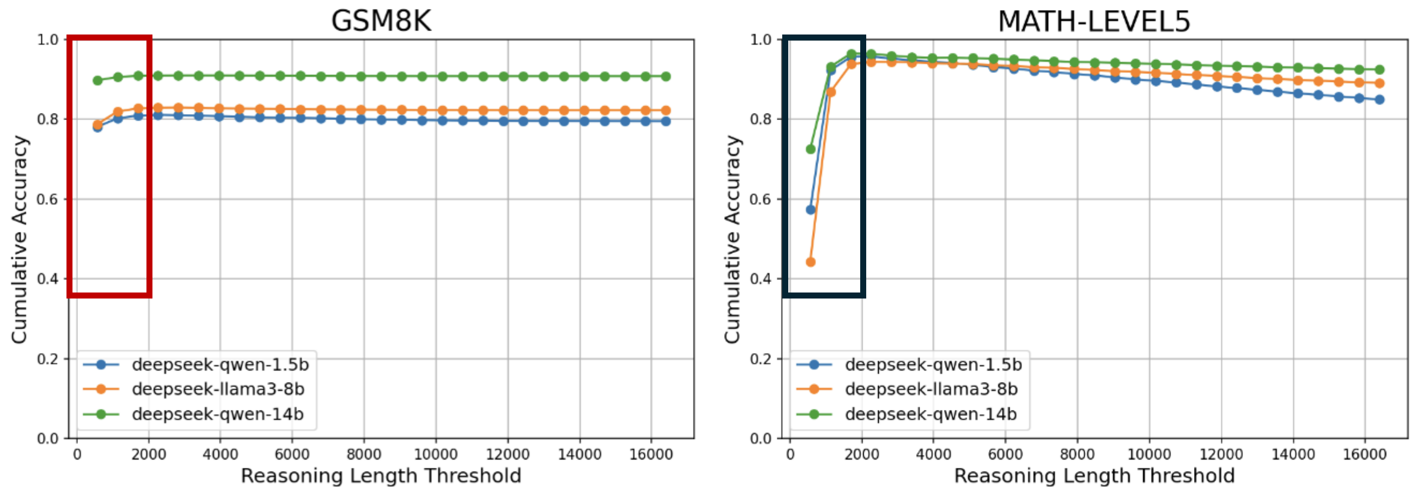


Figure 2: Cumulative accuracy as a function of the reasoning length threshold. The x-axis represents the cutoff threshold on reasoning length, and the y-axis shows the corresponding cumulative accuracy. Models consistently exhibit lower accuracy for overly shorter reasoning (e.g. length <1000).

Findings:

1. 对于较短的推理长度阈值（<2000），模型足够回答较简单的问题，但是在GSM8K上准确率始终较低
2. 对于较长的推理阈值（>2000），模型的准确度略有下降（可能与Overthinking带来的副作用有关）

3. Understanding How Hidden Representations Affect Reasoning Length

3.2 Extracting Reasoning Length Directions

Tips: 实验不使用step-by-step引导思考，探究native thinking能力

逐层计算长/短Thinking的残差流隐藏状态：

$$\bar{r}^{\text{attn}}_{\ell,\text{long}} = \frac{1}{|\mathcal{D}_{\text{long}}|} \sum_{i \in \mathcal{D}_{\text{long}}} \frac{1}{|\mathcal{T}_i|} \sum_{t \in \mathcal{T}_i} r^{\text{attn}}_{\ell}(i,t), \quad \bar{r}^{\text{attn}}_{\ell,\text{short}} = \frac{1}{|\mathcal{D}_{\text{short}}|} \sum_{i \in \mathcal{D}_{\text{short}}} \frac{1}{|\mathcal{T}_i|} \sum_{t \in \mathcal{T}_i} r^{\text{attn}}_{\ell}(i,t),$$

$$\bar{r}^{\text{mlp}}_{\ell,\text{long}} = \frac{1}{|\mathcal{D}_{\text{long}}|} \sum_{i \in \mathcal{D}_{\text{long}}} \frac{1}{|\mathcal{T}_i|} \sum_{t \in \mathcal{T}_i} r^{\text{mlp}}_{\ell}(i,t), \quad \bar{r}^{\text{mlp}}_{\ell,\text{short}} = \frac{1}{|\mathcal{D}_{\text{short}}|} \sum_{i \in \mathcal{D}_{\text{short}}} \frac{1}{|\mathcal{T}_i|} \sum_{t \in \mathcal{T}_i} r^{\text{mlp}}_{\ell}(i,t).$$

>1000 tokens属于Long Thinking， <100 tokens属于Short Thinking, 逐subset逐token计算

作差计算Direction：

$$v^{\text{attn}}_{\ell} = \bar{r}^{\text{attn}}_{\ell,\text{long}} - \bar{r}^{\text{attn}}_{\ell,\text{short}}, \qquad v^{\text{mlp}}_{\ell} = \bar{r}^{\text{mlp}}_{\ell,\text{long}} - \bar{r}^{\text{mlp}}_{\ell,\text{short}}.$$

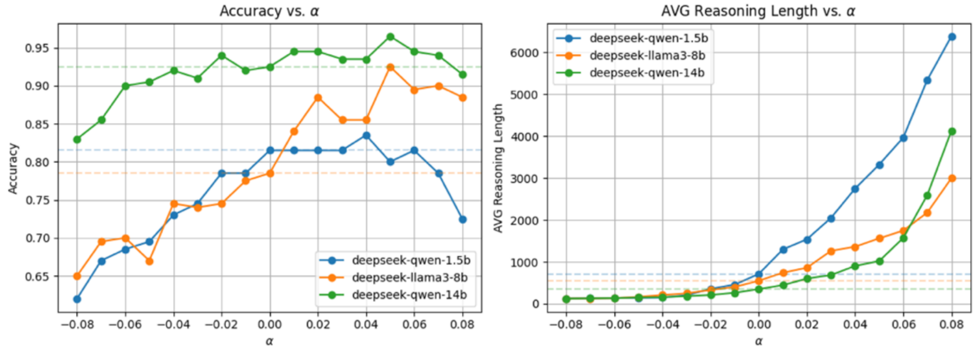
得到attn和mlp两种引导向量

3. Understanding How Hidden Representations Affect Reasoning Length

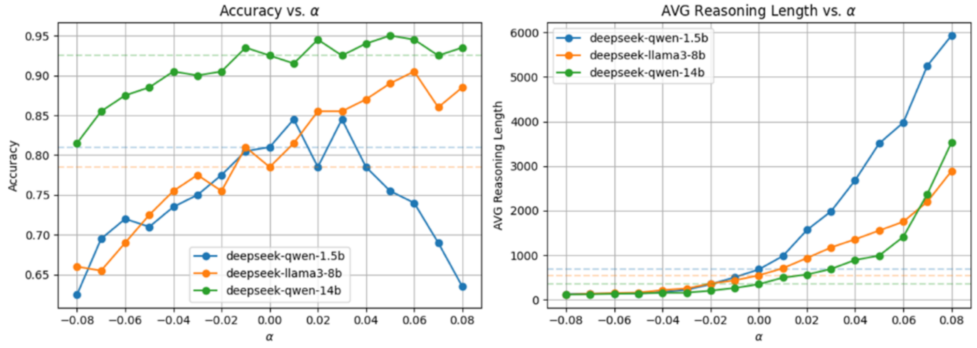
3.3 Effects of Reasoning-Length Direction

Global Steering

GSM8K - Steering with v_ℓ^{attn}

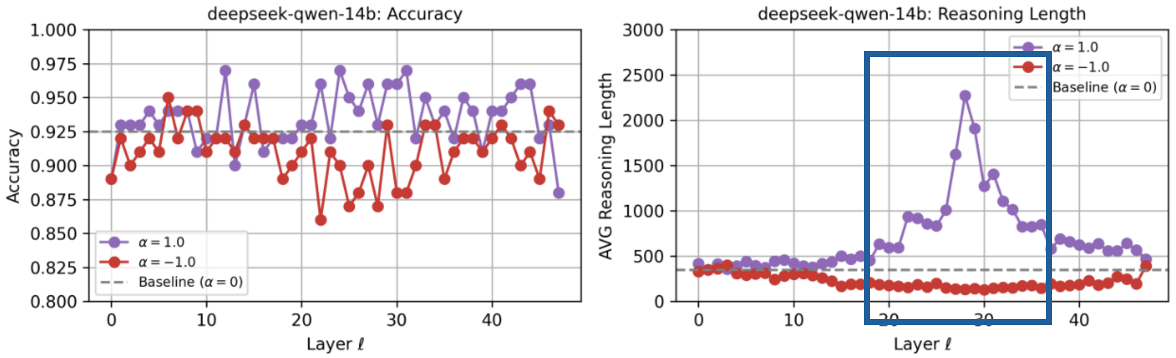


GSM8K - Steering with v_ℓ^{mlp}



Q: 没有探究attn的Layerwise对准确度和推理长度的影响

Layerwise Steering with v_ℓ^{mlp}



中层添加Steering Direction对推理长度增加影响最显著

Key Sights:

1. 正向引导不总是提高准确度（较长的推理长度不总是和推理准确度挂钩）
2. 反向引导一定会降低准确度（overly short thinking一定需要避免）

Budget Control:

steering representation方法与过早终止CoT、附加”wait”方法的区别（更结构化、灵活控制）

4. ThinkEdit: Mitigate Overly Short Reasoning through Weight Editing

2 Steps: Identify Short Reasoning Attention Heads -> Editing Heads

Identify Attention Heads

$$Q^h = rW_q^h \in \mathbb{R}^{T\times d_h}, \quad K^h = rW_k^h \in \mathbb{R}^{T\times d_h}, \quad V^h = rW_v^h \in \mathbb{R}^{T\times d_h}.$$

$$A^h = \mathrm{softmax}\Big(\frac{Q^h (K^h)^\top}{\sqrt{d_h}}\Big) V^h \in \mathbb{R}^{T\times d_h}.$$

$$C^h := A^h W_o^h \in \mathbb{R}^{T\times d}.$$

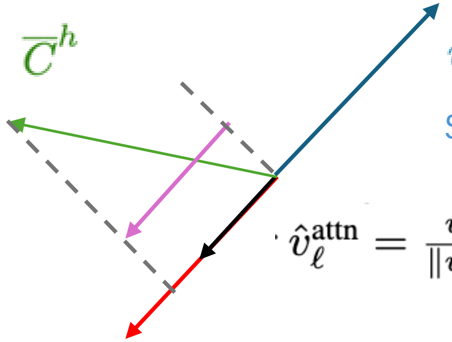
Per-head Contributions

$$\overline{C}^h = \frac{1}{|\mathcal{D}_{\text{short}}|}\sum_{i\in\mathcal{D}_{\text{short}}}\left(\frac{1}{|\mathcal{T}_i|}\sum_{t\in\mathcal{T}_i} C^h(i,t)\right).$$

Short Thinking Contributions

shape: [1, d] (d-hidden dimension)

$$\overline{C}^h_{\text{short}} = \left\langle \overline{C}^h, -\hat{v}_\ell^{\text{attn}} \right\rangle = |\bar{C}^h| cos\theta$$



$$v_\ell^{\text{attn}} = \overline{r}^{\text{attn}}_{\ell,\text{long}} - \overline{r}^{\text{attn}}_{\ell,\text{short}}$$

steering direction

$$\hat{v}_\ell^{\text{attn}} = \frac{v_\ell^{\text{attn}}}{\|v_\ell^{\text{attn}}\|}$$

unit vector

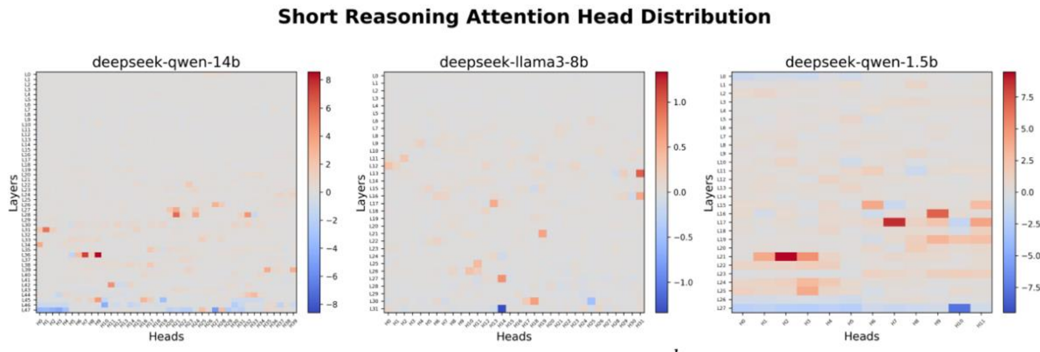


Figure 5: Heatmap illustrating the short reasoning contribution $\overline{C}^h_{\text{short}}$ for each attention head h. Heads with higher values (in red) show stronger alignment with short reasoning behavior.

4. ThinkEdit: Mitigate Overly Short Reasoning through Weight Editing

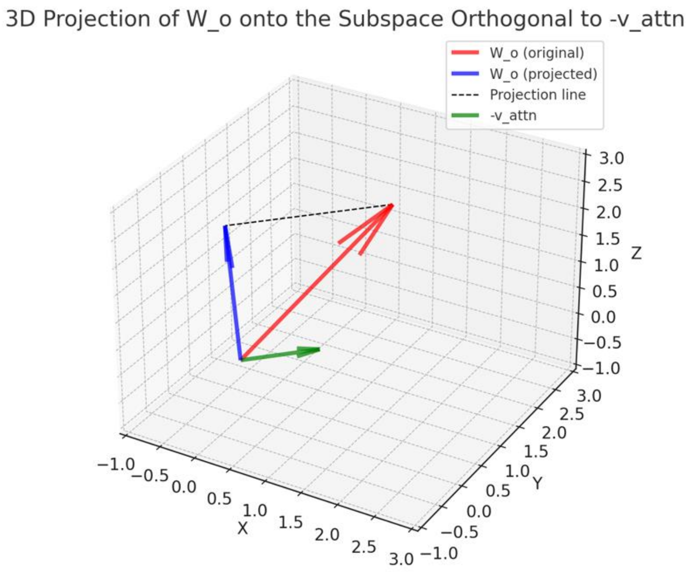
Editing Attention Heads

$$C^h := A^h W_o^h \in \mathbb{R}^{T \times d}.$$

为了改变Contributions，Editing $W_o^{h_l}$

$$W_o^{h_\ell} \leftarrow W_o^{h_\ell}\left(I - (-\hat{v}_\ell^{\text{attn}})(-\hat{v}_\ell^{\text{attn}})^\top\right),$$

移除$W_o^{h_l}$中的导致short reasoning的成分，投影到的子空间始终与$-v_l^{attn}$正交



4.4 Performance of Reasoning Models after ThinkEdit

任务越难，提升越有限

Table 1: Overall accuracy (%) of each model before and after attention-weight editing.

Model		GSM8K	MMLU Elem. Math	MATH-Level1	MATH-Level5	MATH-500
deepseek-qwen-14B	Original	90.80 ± 0.36	95.08 ± 0.65	96.32 ± 0.35	90.25 ± 0.72	91.48 ± 0.55
	ThinkEdit	**93.50 ± 0.31**	**96.53 ± 0.54**	**96.50 ± 0.46**	**91.15 ± 0.59**	**91.78 ± 0.58**
deepseek-llama3-8B	Original	82.26 ± 0.91	96.01 ± 0.62	93.46 ± 0.84	85.49 ± 0.83	87.26 ± 1.16
	ThinkEdit	**88.97 ± 0.78**	**96.08 ± 0.86**	**94.12 ± 0.47**	**85.91 ± 0.48**	**87.60 ± 0.81**
deepseek-qwen-1.5B	Original	79.15 ± 1.08	68.52 ± 1.56	93.00 ± 0.33	**75.48 ± 0.90**	82.22 ± 1.29
	ThinkEdit	**83.34 ± 0.79**	**86.24 ± 1.12**	**93.89 ± 0.76**	74.94 ± 0.85	**82.74 ± 0.77**

Table 2: Accuracy (%) of the top 5% / 10% / 20% shortest reasoning responses.

Model		GSM8K	MMLU Elem. Math	MATH-Level1	MATH-Level5	MATH-500
deepseek-qwen-14b	Original	96.31 / 95.65 / 92.93	93.89 / 96.22 / 95.60	99.52 / 99.30 / 97.70	89.39 / 94.32 / 96.25	86.40 / 91.40 / 93.50
	ThinkEdit	**96.62 / 96.03 / 96.12**	**96.11 / 96.22 / 96.27**	**100.00 / 99.77 / 98.85**	**95.76 / 97.65 / 98.07**	**89.60 / 92.60 / 94.70**
deepseek-llama3-8b	Original	88.92 / 87.18 / 85.82	97.22 / 96.49 / 96.80	97.14 / 94.88 / 94.83	78.64 / 88.79 / 93.41	82.00 / 81.40 / 88.30
	ThinkEdit	**97.08 / 95.27 / 93.95**	**97.78 / 98.65 / 97.87**	**100.00 / 99.30 / 98.62**	**95.61 / 96.89 / 97.12**	**92.80 / 93.60 / 94.40**
deepseek-qwen-1.5b	Original	88.46 / 87.48 / 85.02	62.78 / 62.16 / 60.53	**97.62** / 95.12 / 93.91	91.52 / 95.00 / 95.72	82.40 / 89.80 / 93.40
	ThinkEdit	**92.46 / 92.37 / 92.05**	**77.22 / 80.54 / 79.73**	96.19 / **95.81 / 97.36**	**93.79 / 95.83 / 95.80**	**92.80 / 94.40 / 94.90**