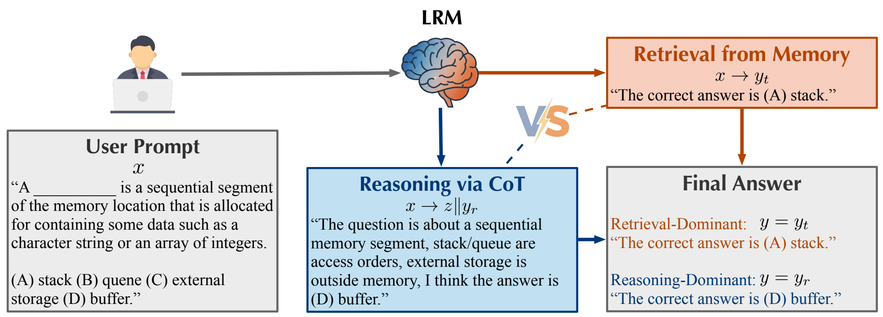
Reasoning or Retrieval? A Study of Answer Attribution on Large Reasoning Models

Yuhui Wang[1] , Changjiang Li[2] , Guangke Chen[3], Jiacheng Liang[1] , Ting Wang[1]

[1]Stony Brook University [2]Palo Alto Networks [3]Wuhan University

Paper Link: →
WeChat: 13819111303
Email: yuhui.wang.1@stonybrook.edu

Background



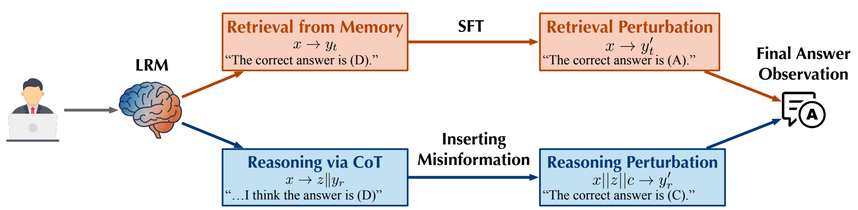
The *Puzzle* We Focus on

- The inconsistency between the Chain-of-Thoughts (CoT) and the final answer.

The *Hypothesis* We Made

- We hypothesize that this inconsistency stems from two competing mechanisms for generating answers: CoT Reasoning and Memory Retrieval.

RQ1: Hypothesis Validation



Bi-level Perturbation Framework

- To test this hypothesis, we conduct controlled experiments that challenge LRMs with misleading cues during reasoning and/or corrupted answers during retrieval.

RQ2: Impact Factors Exploration

Math-related Tasks, Larger models, and RL-trained Models are Reasoning-dominant

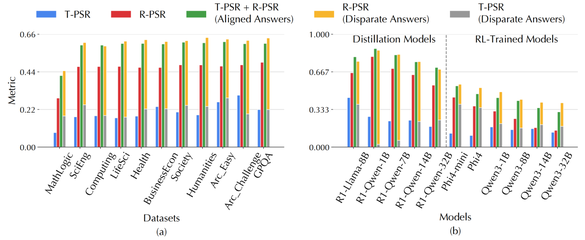


Figure 3: Comparison of reasoning-retrieval influence (a) across datasets and domains (b) between distillation-based and RL-based models (separated by the dashed line).

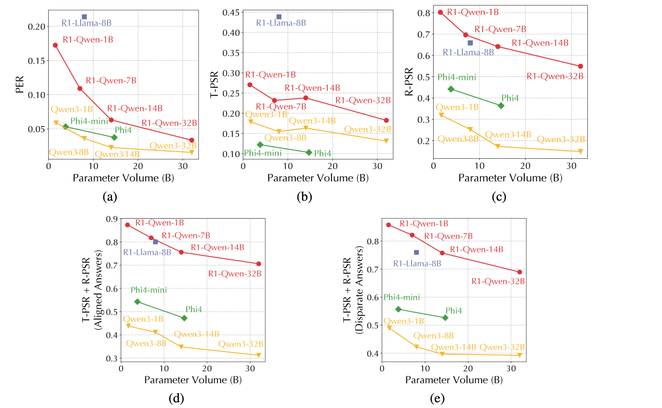


Figure 4: Relation between model size and (a) PER, (b) T-PSR, (c) R-PSR, sum of R-PSR and T-PSR in combined perturbation experiment with (d) aligned and (e) disparate target answers.

Visualization of Reasoning-Retrieval Competing

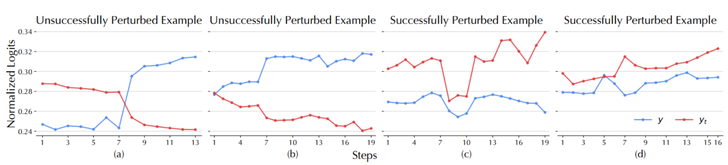


Figure 6: Step-wise reasoning-retrieval interaction through the logit lens. (a), (b): reasoning-dominant cases; (c), (d): retrieval-dominant cases (y: reasoning-led answer; y_t: retrieval-led answer).

RQ3: Reasoning Enhancement

Forgetting-Augmented Reinforcement Learning (FARL)

- To suppress the retrieval shortcut, we introduce FARL, a novel fine-tuning framework that integrates memory unlearning with reinforcement learning.

Algorithm 1: FARL

Input: initial policy model $\pi_{\theta_{\text{init}}}$; training dataset $\mathcal{D}$; hyperparameters ϵ_{low}, ϵ_{high}, β_{KL}, β_{NPO}, μ, training epochs n_{epoch}, inner step n_{step}
Output: π_θ

for *iteration* = $1, \ldots, n_{\text{epoch}}$ **do**
  reference model $\pi_{\theta_{\text{ref}}} \leftarrow \pi_\theta$;
  for *step* = $1, \ldots, n_{\text{step}}$ **do**
    sample batch of prompts and answer pairs x and y from $\mathcal{D}$;
    update old policy model $\pi_{\theta_{\text{old}}} \leftarrow \pi_\theta$;
    compute group advantage $\hat{A}$ (Equation 1) ;
    for *GRPO iteration* = $1, \ldots, \mu$ **do**
      update policy model π_θ by objective $\mathcal{J}_{\text{GRPO}}(\theta; \theta_{\text{old}}, \theta_{\text{ref}}, \hat{A})$ (Equation 2) ;
    unlearn policy model by loss function $\mathcal{L}_{\text{NPO}}(\theta; \theta_{\text{ref}}, x, y)$ (Equation 3) ;
return π_θ;

CoT Robustness, Accuracy, Quality Enhancement

- By carefully suppressing retrieval shortcuts during the fine-tuning process, FARL promotes reasoning-dominant behavior and enhances generalizable reasoning capabilities

Table 1: Comparison of training and reasoning performance of FARL and baseline methods.

Method	Perturbation Metric		Performance Metric (Training Domain)		Performance Metric (Out of Domain)		Training Time
	R-PSR ↓	T-PSR ↓	MTL	ACC ↑	MTL	ACC ↑	
R1-Llama-8B (Base)	0.378	0.381	1537.9	0.725	1386.2	0.716	/
SFT	0.392	0.311	1381.7	0.787	1207.3	0.732	10m 21s
RL (GRPO)	0.259	0.262	1854.0	0.869	1844.4	0.745	4h 6m 27s
FARL	0.197	0.234	1914.0	0.891	1896.9	0.757	4h 26m 4s

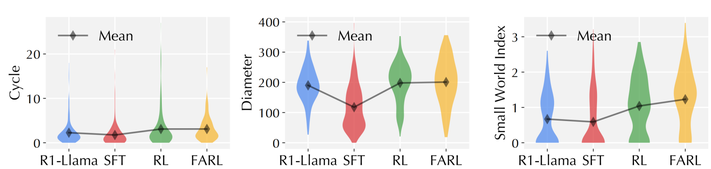


Figure 6: Cycle, diameter, and small world index distributions of the reasoning graph generated by LRMs trained with FARL and baselines.