Reasoning Model

Correctness of intermediate answers during reasoning

Reasoning Models Know When They're Right: Probing Hidden States for Self-Verification

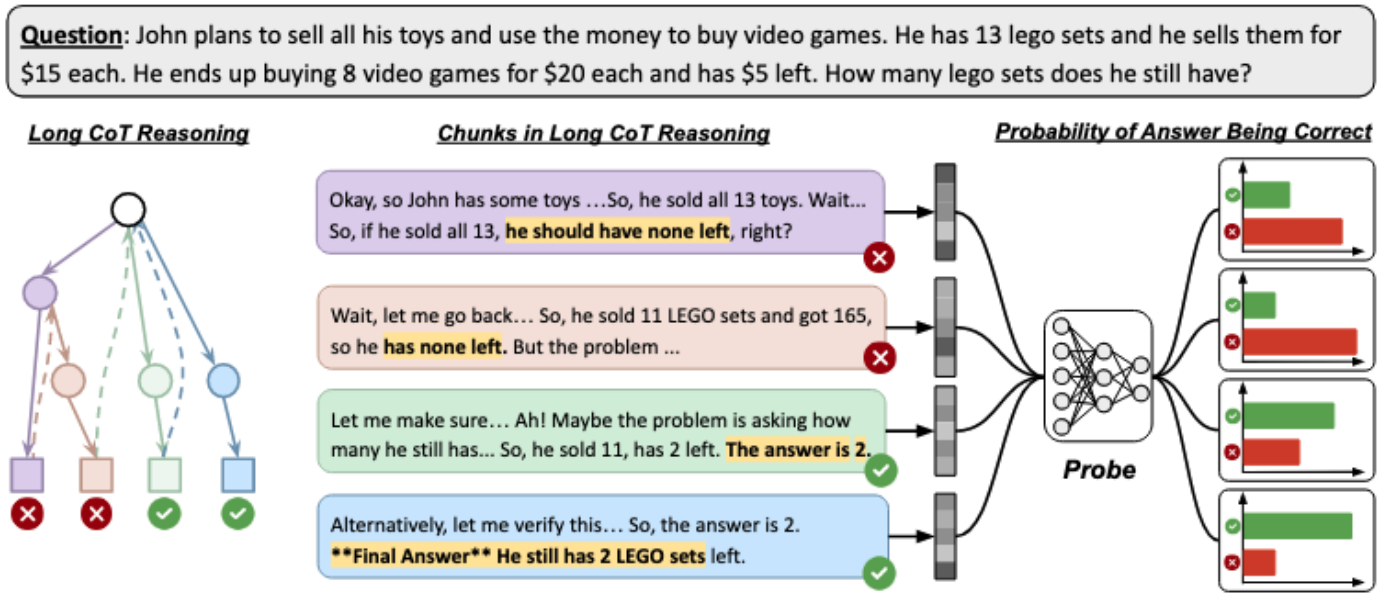


Figure 1: An illustration of the probing method. On the left side, long CoT is parsed into multiple chunks, each corresponding to a reasoning path and contains an intermediate answer as termination. On the right side, representations for each chunk are obtained and probe is used to predict the probability of answer being correct.

http://arxiv.org/abs/2504.05419

Intro

Key Sight:

A key advantage of reasoning models lies in their ability to search: they often explore multiple reasoning paths leading to different intermediate answers to the original problem before arriving at a final solution

Motivation:

Reasoning models tend to *overthink* by exploring additional reasoning paths even after reaching a correct answer.

Question:

To what extent can models evaluate the correctness of their intermediate answers during reasoning?

Method:

We investigate this question by probing the model's hidden states for answer correctness.

Probing for intermediate answer correctness

Step 1: Data collection

使用Gemini 2.0 Flash提取 (c_1, y_1) – (intermediate answer, correctness)

question

<think>

intermediate answer 1

Wait, double-check…

intermediate answer 2

…

</think>

answer

Step 2: Training the probe

2 layers MLP

由于训练数据不均衡（大多是正确的中间答案），使用加权交叉熵损失

$$p_i = \sigma(\mathrm{ReLU}(e_i \mathbf{W}_1 + \mathbf{b}_1)\mathbf{W}_2 + b_2)$$

$$\mathcal{L}(\mathbf{W}, \mathbf{b}) = -\frac{1}{N}\sum_{i=1}^{N}(w\alpha y_i \log p_i + (1 - y_i)\log(1 - p_i))$$

ω – 负样本与正样本比例
α – 用来缩放不平衡权重的超参

Experiments

4.1 基础实验设置

4.2 探索答案正确性的信息是否被编码在推理模型中

4.3 是否跨数据集具有泛化性

4.4 这种信息是否与长链推理能力有关

4.5 这种信息在明确回答形成之前是否已经被良好编码

4.2 Reasoning models encode answer correctness

test **in-distribution** performance

Metrics:

ROC-AUC Score	ROC曲线：在各种分类阈值中，真正率和假正率的关系 AUC：曲线下面积 越接近1，说明模型区分正负样本能力越强
ECE (Expected Calibration Error)	校准误差
$\mathrm{ECE} = \sum_{m=1}^{M} \frac{\lvert B_m \rvert}{n} \lvert acc(B_m) - conf(B_m) \rvert$	输出概率分成若干区间（0-0.1，0.1-0.2 …） 对于每个区间，计算该区间样本实际准确率 和模型预测置信度
Brier Score	模型预测概率与实际标签的误差平方和的均值 不仅考虑是否正确，还考虑置信度

4.2 Reasoning models encode answer correctness

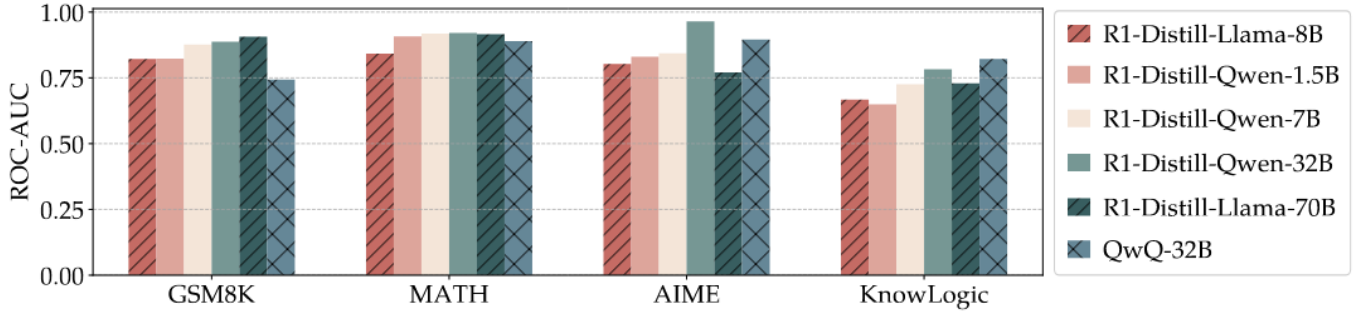


Figure 2: ROC-AUC scores for each probe trained on hidden states from different reasoning models and datasets. We train a separate probe on each probing dataset and evaluate it on in-distribution test set.

Reasoning Model	GSM8K		MATH		AIME		KnowLogic	
	ECE ↓	Brier ↓	ECE ↓	Brier ↓	ECE ↓	Brier ↓	ECE ↓	Brier ↓
R1-Distill-Llama-8B	0.05	0.17	0.03	0.14	0.10	0.11	0.07	0.23
R1-Distill-Llama-70B	0.03	0.07	0.07	0.10	0.10	0.18	0.03	0.19
R1-Distill-Qwen-1.5B	0.04	0.16	0.04	0.12	0.14	0.12	0.09	0.20
R1-Distill-Qwen-7B	0.02	0.11	0.03	0.10	0.09	0.15	0.06	0.21
R1-Distill-Qwen-32B	0.01	0.08	0.06	0.09	0.13	0.10	0.10	0.19
QwQ-32B	0.03	0.13	0.13	0.10	0.08	0.13	0.03	0.15

Table 1: Expected Calibration Error (ECE) and Brier score for the in-distribution performance of each probe trained on each probing dataset.

ROC-AUC 均超过0.7，且ECE低于0.1

在数学任务上表现优于逻辑任务

在某些任务中最优的probe的hidden dim为0，说明正确性信息可以线形编码

4.3 Probes generalize to some out-of-distribution datasets

Training Data	GSM8K		MATH		AIME		KnowLogic	
	AUC ↑	ECE ↓	AUC ↑	ECE ↓	AUC ↑	ECE ↓	AUC ↑	ECE ↓
GSM8K	0.82	0.05	0.80 (-0.04)	0.08 (+0.05)	0.69 (-0.11)	0.25 (+0.15)	0.56 (-0.11)	0.10 (+0.03)
MATH	0.83 (+0.01)	0.04 (-0.01)	0.84	0.03	0.76 (-0.04)	0.28 (+0.18)	0.63 (-0.04)	0.08 (+0.01)
KnowLogic	0.77 (-0.05)	0.17 (+0.12)	0.74 (-0.10)	0.19 (+0.16)	0.81 (+0.01)	0.31 (+0.21)	0.67	0.07

Table 2: ROC-AUC scores and ECE of trained probes on out-of-distribution test set. The numbers in red and green denote performance decrease and increase relative to the probe trained on in-distribution training set, respectively. R1-Distill-Llama-8B is used as the reasoning model.

4.4 Encoding of correctness is related to long CoT reasoning abilities

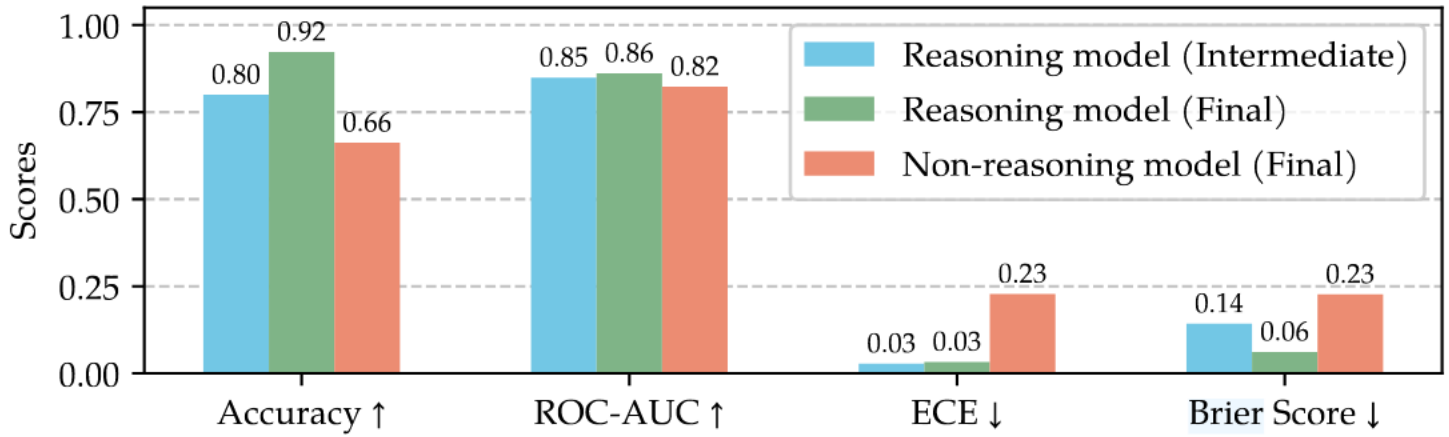
探究这种正确性的编码是否与长CoT reasoning能力有关

训练了一个非推理模型的probe

question

answer

(c, y) – (final answer, correctness)



该probe的表现比reasoning model上训练的probe差很多，可能表明reasoning model在长时间的链式思维监督训练中，self-validation的能力得到了增强

4.5 Correctness can be detected before the answer is generated

question

<think>

intermediate answer 1

Wait

intermediate answer 2

…

</think>

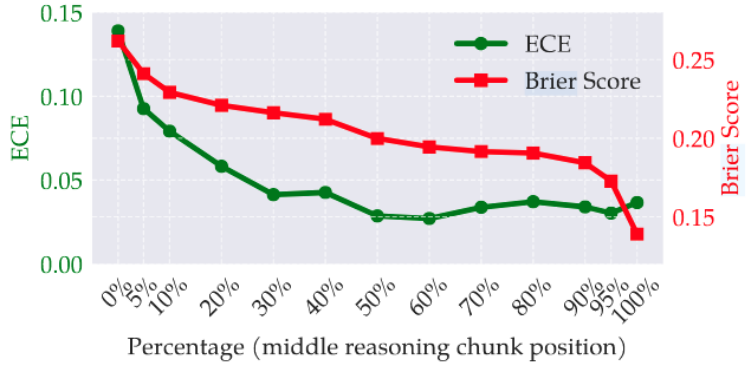
answer

intermediate answer 2

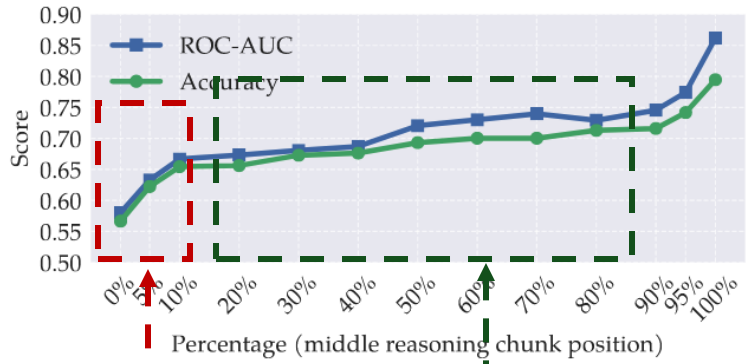
0%

50%

100%



(a) ECE and Brier Score decrease as the paragraph position approaches the answer at the end of the reasoning chunk



(b) Accuracy and ROC-AUC increase as the paragraph position approaches the generated answer at the end of the reasoning chunk

- probe performance与中间答案的接近程度成正相关
- Accuracy在0-10%区间内陡峭增加，中间位置平缓，

  说明早期位置也包含正确性信息

Probe as a verifier for early-exit

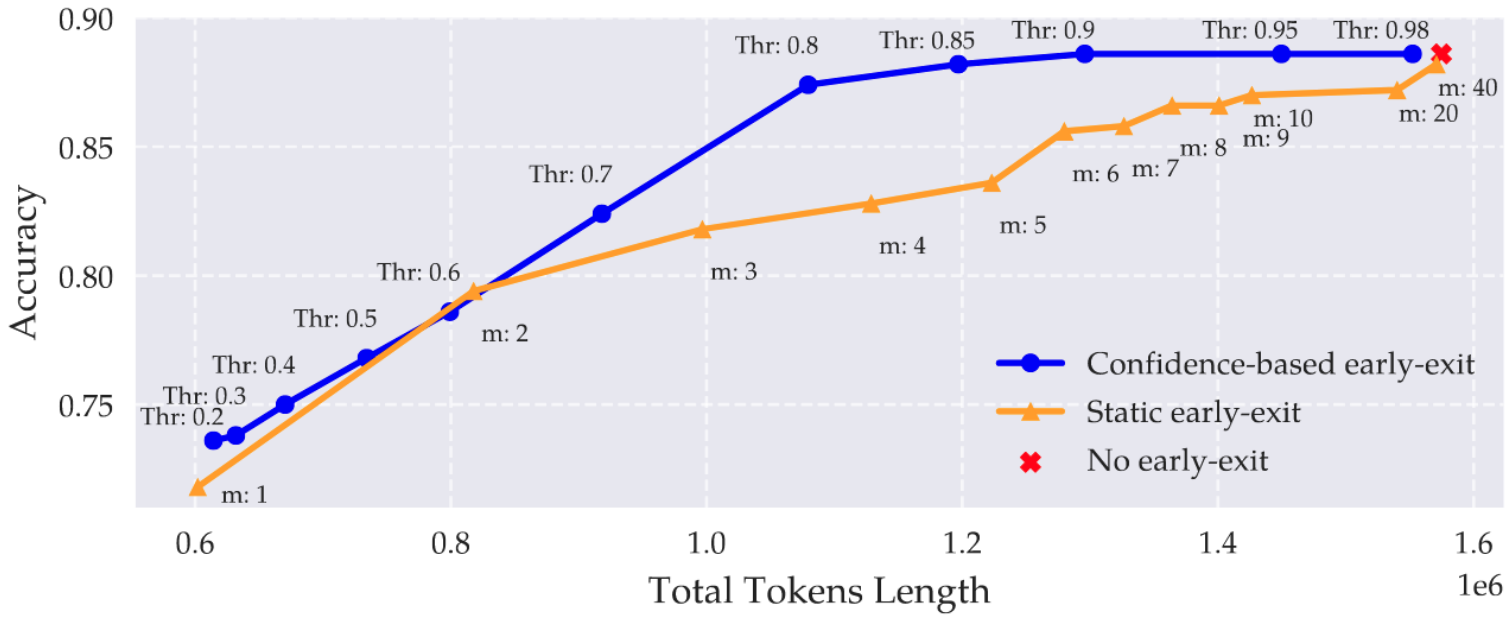


Figure 5: Final answer accuracy versus inference token cost with different early-exit strategies. For confidence-based early-exit, the curve is obtained by varying the confidence threshold for answer correctness. For static early-exit, the curve is generated by varying the chunk number m.