Contents

两篇关于高效 reasoning 的综述：

1 reasoning economy

Harnessing the Reasoning Economy A Survey of Efficient Reasoning for Large Language Models

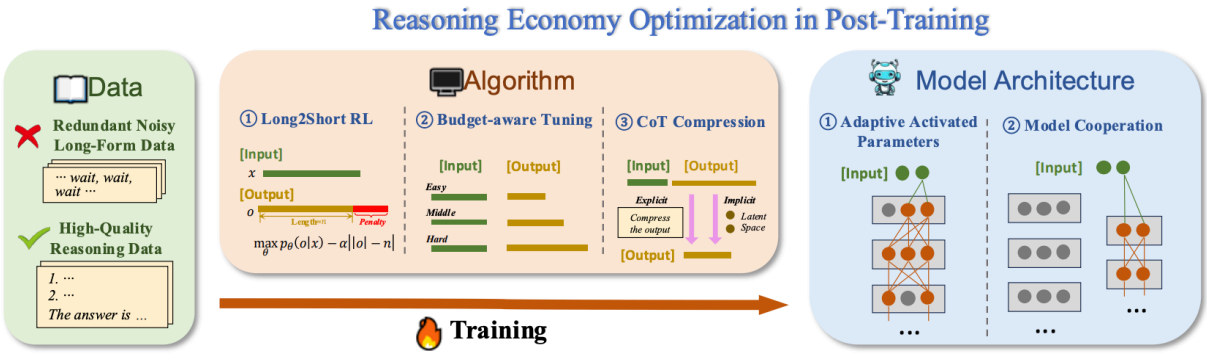


Figure 4: Post-training Methods Optimization for Reasoning Economy.

- 数据：
  - 干掉冗余、噪声、太长的数据
  - 保留高质量数据
- 算法：
  - Long2Short RL：惩罚过长的输出
  - budget-aware tuning：简单问题输出短，复杂问题输出长
  - CoT Compression：显式/隐式压缩 CoT
- 模型：
  - adaptive Activated Parameters：类似稀疏激活
  - Model Cooperation：搞 2 个模型，简单问题走简单模型，复杂问题走复杂模型

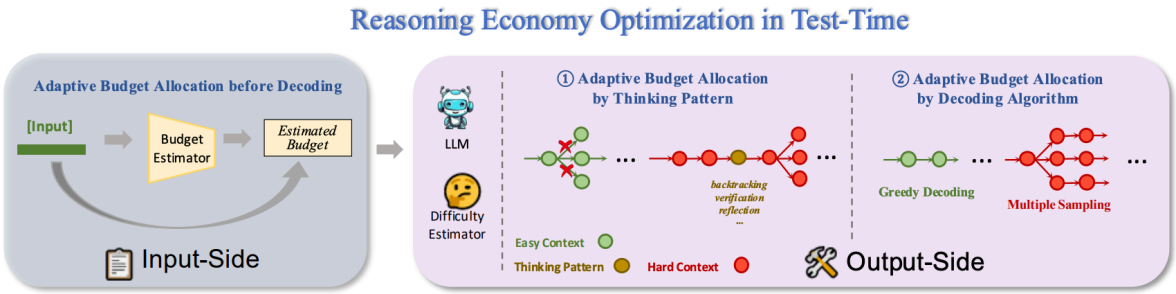


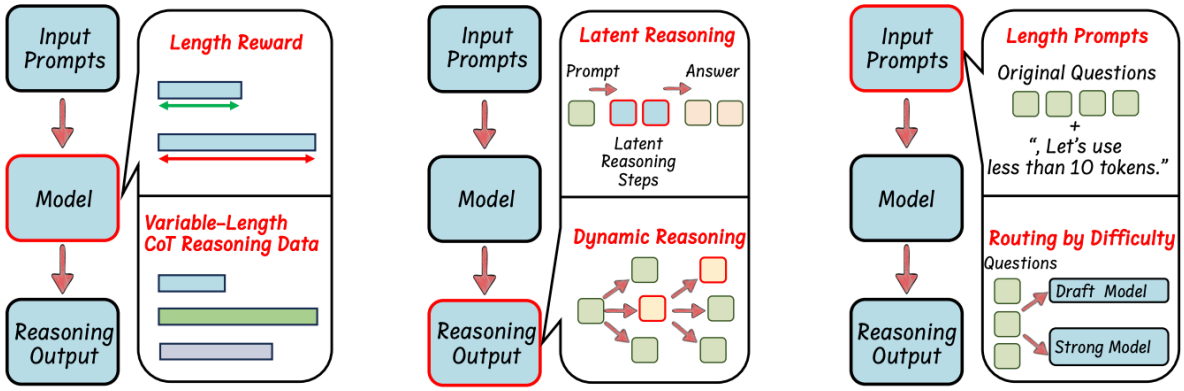
Figure 5: Test-time Methods Optimization for Reasoning Economy. The methods are divided into adding optimal computation constraint in the input-side, and selecting the best-performing decoding algorithm and controlling computation usage during decoding in the output-side.

- 输入侧：解码之前进行自适应预算分配

  - 对输入进行成本预估
- 输出侧：
  - thinking pattern 的自适应预算分配：简单问题直接剪枝，复杂问题需要回溯 + 验证 + 反思
  - 解码算法的自适应预算分配：简单问题贪心解码，复杂问题多一些采样

2 Stop Overthinking

Stop Overthinking: A Survey on Efficient Reasoning for Large Language Models



- model：
  - length reward：RL 时加上新的 reward，鼓励答案正确且 cot 短的
  - variable-length cot reasoning data：构造不同长度 cot 的数据集，然后 sft
- reasoning-output：
  - latent reasoning：将 reasoning steps 压缩成 latent 表示
  - dynamic reasoning：reasoning 的过程改成投机采样、拒绝采样、tree-of-thoughts 等方式
- input prompt：
  - length prompts：let’s use less than k tokens
  - routing by difficulty：训一个小模型，决定简单问题不思考，困难问题再思考