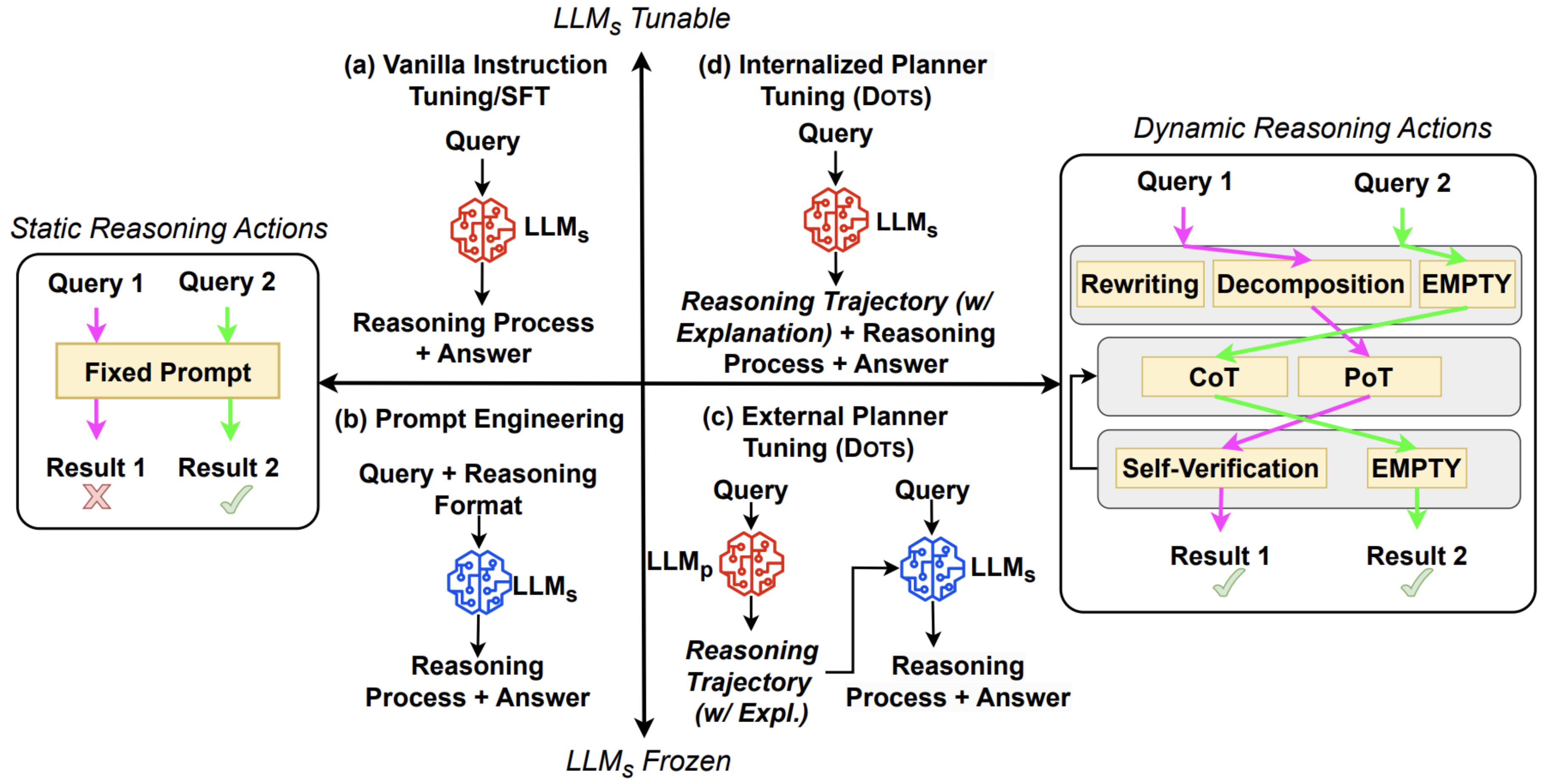
DOTS: Learning to Reason Dynamically in LLMs via Optimal Reasoning Trajectories Search

Murong Yue[1], Wenlin Yao[2], Haitao Mi[2], Dian Yu[2], Ziyu Yao[1], Dong Yu[2]

[1]George Mason University, [2]Tencent AI Lab, Bellevue

Motivation


LLM_s Tunable
(a) Vanilla Instruction Tuning/SFT
Query
LLM_s
Reasoning Process + Answer
(d) Internalized Planner Tuning (DOTS)
Query
LLM_s
Reasoning Trajectory (w/ Explanation) + Reasoning Process + Answer
Static Reasoning Actions
Query 1
Query 2
Fixed Prompt
Result 1
Result 2
(b) Prompt Engineering
Query + Reasoning Format
LLM_s
Reasoning Process + Answer
(c) External Planner Tuning (DOTS)
Query
LLM_p
Reasoning Trajectory (w/ Expl.)
Query
LLM_s
Reasoning Process + Answer
LLM_s Frozen
Dynamic Reasoning Actions
Query 1
Query 2
Rewriting
Decomposition
EMPTY
CoT
PoT
Self-Verification
EMPTY
Result 1
Result 2

Static and dynamic paradigms of LLM reasoning

Question II:
How many r in word 'strawberry'?

Answer from GPT-4o-2024-08-06 (Incorrect):
The word "strawberry" contains 2 letter "r"s.

Answer from our internal planner tuned Llama-8b-Instruct:
It can efficiently count the occurrences of a specific character (in this case, 'r') in a word. Programming_solver provides a quick and reliable method for counting characters in text, ensuring accuracy in the result.
Required Action: programming

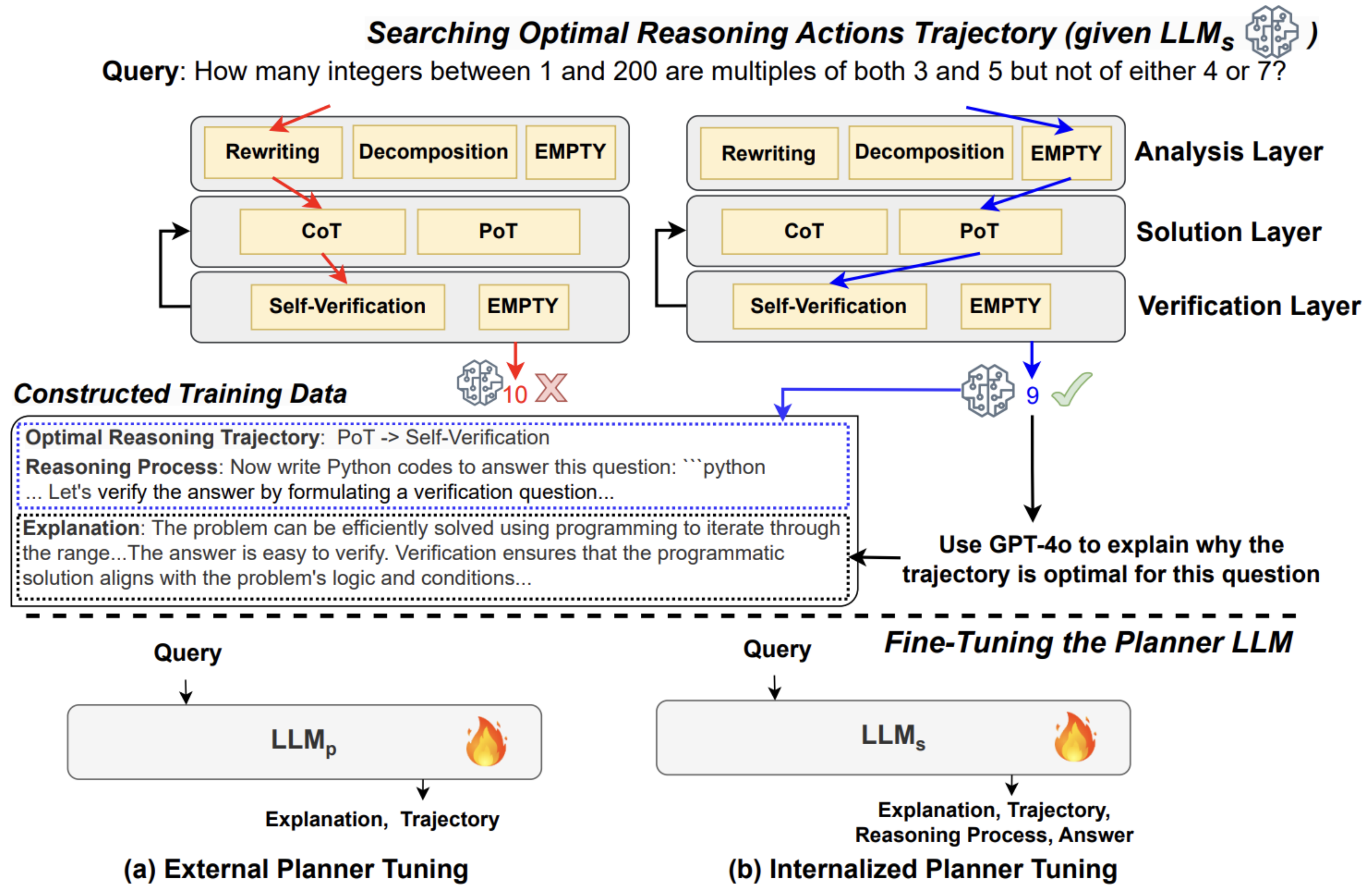
```
word = 'strawberry'
count=word.lower().count('r')
print(count)
```

After execution, we get: 3
Answer: 3

Case Study

Motivation: The current LLM fails to **dynamically** decide the best reasoning strategies, e.g., when to use program.

Method


Searching Optimal Reasoning Actions Trajectory (given LLM_s)
Query: How many integers between 1 and 200 are multiples of both 3 and 5 but not of either 4 or 7?
Rewriting
Decomposition
EMPTY
CoT
PoT
Self-Verification
EMPTY
Analysis Layer
Solution Layer
Verification Layer
10
9
Constructed Training Data
Use GPT-4o to explain why the trajectory is optimal for this question
Fine-Tuning the Planner LLM
Query
LLM_p
Explanation, Trajectory
(a) External Planner Tuning
Query
LLM_s
Explanation, Trajectory, Reasoning Process, Answer
(b) Internalized Planner Tuning

Optimal Reasoning Trajectory: PoT -> Self-Verification
Reasoning Process: Now write Python codes to answer this question: ```python ... Let's verify the answer by formulating a verification question...
Explanation: The problem can be efficiently solved using programming to iterate through the range...The answer is easy to verify. Verification ensures that the programmatic solution aligns with the problem's logic and conditions...

1. **Module Design**: We manually design atomic reasoning action modules based on prior prompting works, e.g., programming, verification;
2. **Searching Best Trajectory**: We perform multiple trials on all possible modules permutations and select the best one with the highest accuracy;
3. **Training a planner**: We use GPT-4o to generate the reason why the selected trajectory is the best based on the nature of the question, then either finetune a planner LLM as an external planning module (DOTS: External) or directly finetune the solver LLM and internalize the planning ability (DOTS: Internalized).

Experiments

Method	Tuning	Reasoning Format	MATH	BBH	Game of 24	TheoremQA	Average
External Planner: Llama-3-8B-Instruct; Solver: Llama-3-70B-Instruct							
CoT	✗	$\mathcal{L}$	50.4	72.7	27.5	27.4	44.5
LTM	✗	$\mathcal{L}$	50.1	73.8	24.9	28.8	44.4
PA	✓	$\mathcal{L}$	52.5	72.9	26.8	28.8	45.3
PoT	✗	$\mathcal{P}$	54.7	65.8	63.9	31.1	53.9
Self-refine	✗	$\mathcal{L}, \mathcal{P}$	55.9	71.4	**68.3**	30.8	56.6
DOTS: External	✓	$\mathcal{L}, \mathcal{P}$	**57.7**	**77.3**	67.7	**31.2**	**58.5**
External Planner: Llama-3-8B-Instruct; Solver: GPT4o-mini							
CoT	✗	$\mathcal{L}$	70.2	80.3	27.7	38.9	54.2
LTM	✗	$\mathcal{L}$	72.2	79.4	25.5	36.4	53.3
PA	✓	$\mathcal{L}$	73.5	81.1	26.7	38.9	55.1
PoT	✗	$\mathcal{P}$	67.2	73.9	61.4	35.8	59.6
Self-refine	✗	$\mathcal{L}, \mathcal{P}$	73.7	74.8	**68.7**	34.6	63.0
DOTS: External	✓	$\mathcal{L}, \mathcal{P}$	**75.4**	**84.2**	65.2	**41.4**	**66.5**

Method	Tuning	Reasoning format	MATH	BBH	Game of 24	TheoremQA	Average
Solver: Llama-3-8B-Instruct							
CoT	✗	$\mathcal{L}$	29.6	48.9	12.7	14.8	26.5
LTM	✗	$\mathcal{L}$	29.5	50.3	14.4	15.2	27.4
PA	✓	$\mathcal{L}$	31.0	47.2	11.8	15.1	26.3
PoT	✗	$\mathcal{P}$	25.3	44.6	16.8	**16.7**	25.9
Self-refine	✗	$\mathcal{L}, \mathcal{P}$	28.7	46.6	17.0	15.3	30.1
Vanilla SFT	✓	$\mathcal{L}$	33.9	61.0	18.5	14.8	33.6
DOTS: Internalized	✓	$\mathcal{L}, \mathcal{P}$	**34.4**	**69.7**	**21.9**	16.1	**35.5**

- Both the external planner and internal planner are better than the baselines;
- Further analysis show that our method can both adapt to the characteristics of specific questions and the capability of specific task-solving LLMs.