TaD: A Plug-and-Play Task-Aware Decoding Method to Better Adapt LLMs on Downstream Tasks

Xinhao Xu, Hui Chen, Zijia Lin, Jungong Han, Lixing Gong, Guoxin Wang, Yongjun Bao, Guiguang Ding

IJCAI JEJU 2024

University of Sheffield

JD.COM

Introduction

- **Fine-tuning as a common strategy to enhance the pre-trained LLMs in downstream tasks:**
  - Algorithmic side: better fine-tuning methods, e.g. PEFT.
  - Data side: more effective datasets.
  - Inherent knowledge acquisition of fine-tuned LLMs **rarely** investigated in the existing works.
- **Motivation:** The outputs of pre-trained LLMs do not always accurately reflect the knowledge they possess.

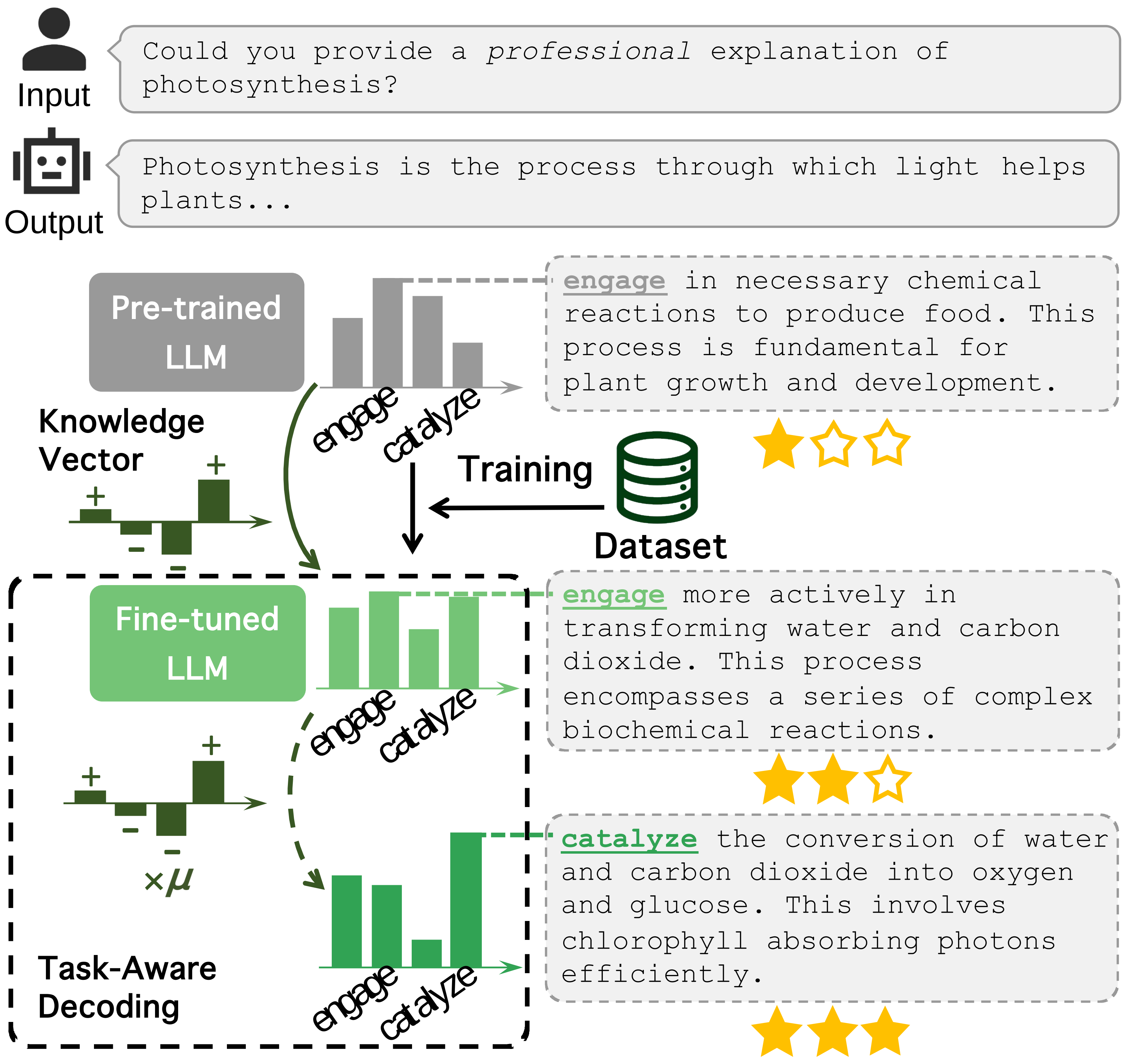
Research Problem

How can we leverage such inherent knowledge in the fine-tuned LLMs to enhance their performance in downstream tasks?

- **Intuitive Ideas:**
  - Token-predicting behavior alterations during the fine-tuning process reflect the the inherent knowledge.
  - Such alterations indicate an adaptive shift from common knowledge to specific knowledge for downstream tasks.
  - Manually mining and leveraging such inherent knowledge can improve the adaptation of LLMs on downstream tasks.

Method

- **A demonstration of the proposed knowledge vector and TaD:**



Input: Could you provide a *professional* explanation of photosynthesis?

Output: Photosynthesis is the process through which light helps plants...

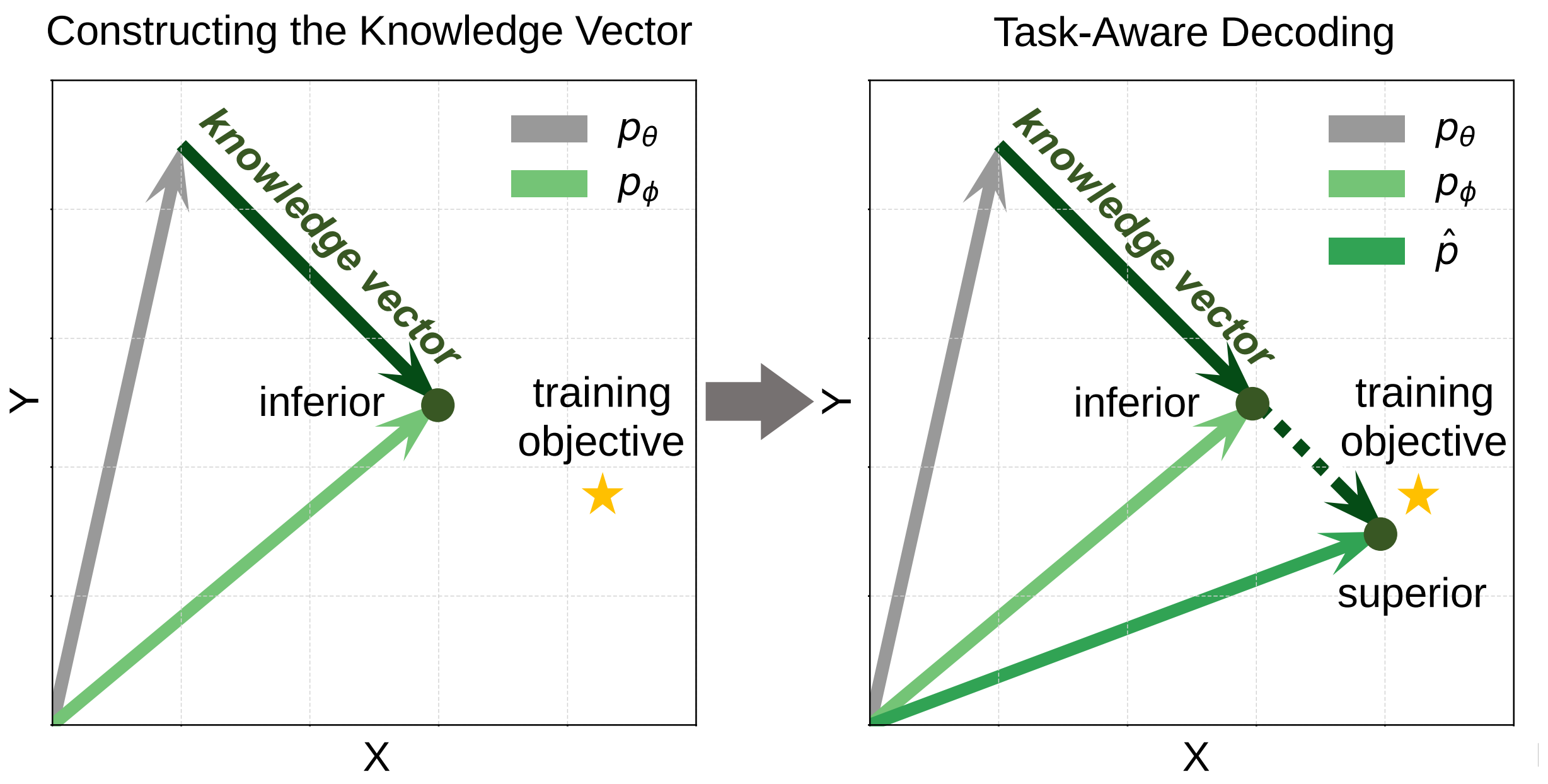
Pre-trained LLM: engage in necessary chemical reactions to produce food. This process is fundamental for plant growth and development. ★☆☆

Fine-tuned LLM: engage more actively in transforming water and carbon dioxide. This process encompasses a series of complex biochemical reactions. ★★☆

Task-Aware Decoding: catalyze the conversion of water and carbon dioxide into oxygen and glucose. This involves chlorophyll absorbing photons efficiently. ★★★

- **Knowledge Vector:**
  - Formulating the knowledge difference.
  - Explicitly denoting the direction of knowledge adaptation learned by a pre-trained LLM during fine-tuning.
  - Naturally possessing semantic information.
- **Task-Aware Decoding:**
  - Enhancing the fine-tuned LLM's output probability distribution with the knowledge vector.
  - Reinforcing the model's knowledge adaptation to downstream tasks for better performance.

- **A simplified illustration of our work:**
  - p_θ and p_ϕ: predicted token distribution of pre-trained and fine-tuned LLMs



Experiments

- **Results on multiple-choice and CBQA tasks:**

Model	Method	Multiple Choices			CBQA
		MC1	MC2	MC3	True*Info
GPT-J-6b	LoRA	30.6	51.3	25.6	35.7
	+*TaD*	**33.0**	**52.5**	**27.1**	**37.0**
	AdapterP	34.9	54.3	28.0	51.5
	+*TaD*	**38.2**	**55.5**	**29.5**	**51.7**
	AdapterH	36.4	55.0	28.5	53.0
	+*TaD*	**38.3**	**55.8**	**28.7**	**55.3**
	Parallel	34.3	54.0	27.7	47.2
	+*TaD*	**37.5**	**55.1**	**28.9**	**47.4**
BLOOMz-7b	LoRA	30.8	51.4	25.7	17.4
	+*TaD*	**32.8**	**52.3**	**27.2**	**17.5**
	AdapterP	35.3	53.8	**28.5**	20.6
	+*TaD*	**35.7**	**54.8**	28.4	**20.7**
	AdapterH	36.8	54.5	28.9	50.3
	+*TaD*	**37.9**	**55.2**	**29.2**	**50.8**
	Parallel	34.5	53.6	28.2	21.8
	+*TaD*	**36.5**	**54.4**	**28.5**	**22.7**

Model	Method	Multiple Choices			CBQA
		MC1	MC2	MC3	True*Info
LLaMa-7b	LoRA	32.9	55.0	28.5	49.1
	+*TaD*	**34.2**	**55.7**	**29.0**	**51.2**
	AdapterP	38.1	57.4	30.8	61.4
	+*TaD*	**40.6**	**58.5**	**32.1**	**61.8**
	AdapterH	37.8	57.6	30.3	60.3
	+*TaD*	**39.8**	**59.0**	**32.0**	**61.0**
	Parallel	37.0	56.3	29.5	54.3
	+*TaD*	**39.5**	**57.0**	**30.4**	**55.2**
LLaMa-13b	LoRA	33.4	55.7	29.0	54.1
	+*TaD*	**35.1**	**56.7**	**29.7**	**54.7**
	AdapterP	40.6	58.8	32.4	58.6
	+*TaD*	**42.6**	**60.0**	**33.1**	**60.0**
	AdapterH	38.2	57.0	30.4	61.8
	+*TaD*	**39.5**	**57.8**	**31.2**	**63.3**
	Parallel	39.8	58.2	31.7	60.0
	+*TaD*	**42.0**	**60.2**	**33.8**	**61.6**

- **Results on reasoning tasks:**

Model	Method	Math Reasoning		CS Reasoning	
		GSM8K	MultiArith	BoolQ	PIQA
GPT-J-6b	LoRA	21.9	92.5	61.8	63.4
	+*TaD*	**22.8**	**94.2**	**62.7**	**64.6**
	AdapterP	19.0	92.2	63.9	71.0
	+*TaD*	**19.5**	**92.5**	**64.2**	**71.2**
BLOOMz-7b	LoRA	18.9	91.7	66.8	73.6
	+*TaD*	**19.3**	**94.2**	**66.9**	**73.9**
	AdapterP	16.3	90.7	66.2	74.4
	+*TaD*	**17.1**	**93.0**	66.2	**75.0**
LLaMa-7b	LoRA	26.6	90.5	68.7	78.9
	+*TaD*	**27.7**	**91.0**	**69.3**	**79.5**
	AdapterP	31.5	93.5	65.4	76.3
	+*TaD*	**32.0**	**93.7**	**66.3**	76.3
LLaMa-13b	LoRA	35.9	91.5	70.1	82.5
	+*TaD*	**38.1**	**92.0**	**70.8**	**83.1**
	AdapterP	36.8	91.5	69.4	78.1
	+*TaD*	**37.5**	**94.0**	69.4	**79.2**

- **Ablation study of the knowledge vector:**

$\mathcal{M}$	$p_{\mathcal{S}} \to p_{\mathcal{E}}$	G / M
7b	/	10.8/37.5
7b*	/	26.6/90.5
13b	/	16.7/53.2
13b*	/	35.9/91.5

(a) Comparison results on pre-trained and fine-tuned models.

$\mathcal{M}$	$p_{\mathcal{S}} \to p_{\mathcal{E}}$	G / M
7b*	/	26.6/90.5
	7b → 7b*	27.7/91.0
13b*	/	35.9/91.5
	13b →13b*	38.1/92.0

(b) *TaD*'s effectiveness on the fine-tuned models.

$\mathcal{M}$	$p_{\mathcal{S}} \to p_{\mathcal{E}}$	G / M
7b*	/	26.6/90.5
	7b*→ 7b	23.7/79.0

(c) The effect of the opposite direction of the proposed *knowledge vector* (from the fine-tuned to the pre-trained model).

$\mathcal{M}$	$p_{\mathcal{S}} \to p_{\mathcal{E}}$	G / M
13b	/	16.7/53.2
	7b →13b	17.2/51.8
13b*	/	35.9/91.5
	7b*→13b*	36.2/91.8

(d) The effect of the direction of the model size difference (from the smaller to the larger model).

$\mathcal{M}$	$p_{\mathcal{S}} \to p_{\mathcal{E}}$	G / M
7b	/	10.8/37.5
	7b → 7b*	11.9/38.1
	7b →13b	11.2/37.6

(e) Comparison results on the direction of the knowledge and model size difference.

$\mathcal{M}$	$p_{\mathcal{S}} \to p_{\mathcal{E}}$	G / M
13b*	/	35.9/91.5
	7b*→13b*	36.2/91.8
	13b →13b*	38.1/92.0
	7b →13b*	38.2/92.0

(f) The cumulative effect of the direction of the knowledge and model size difference.

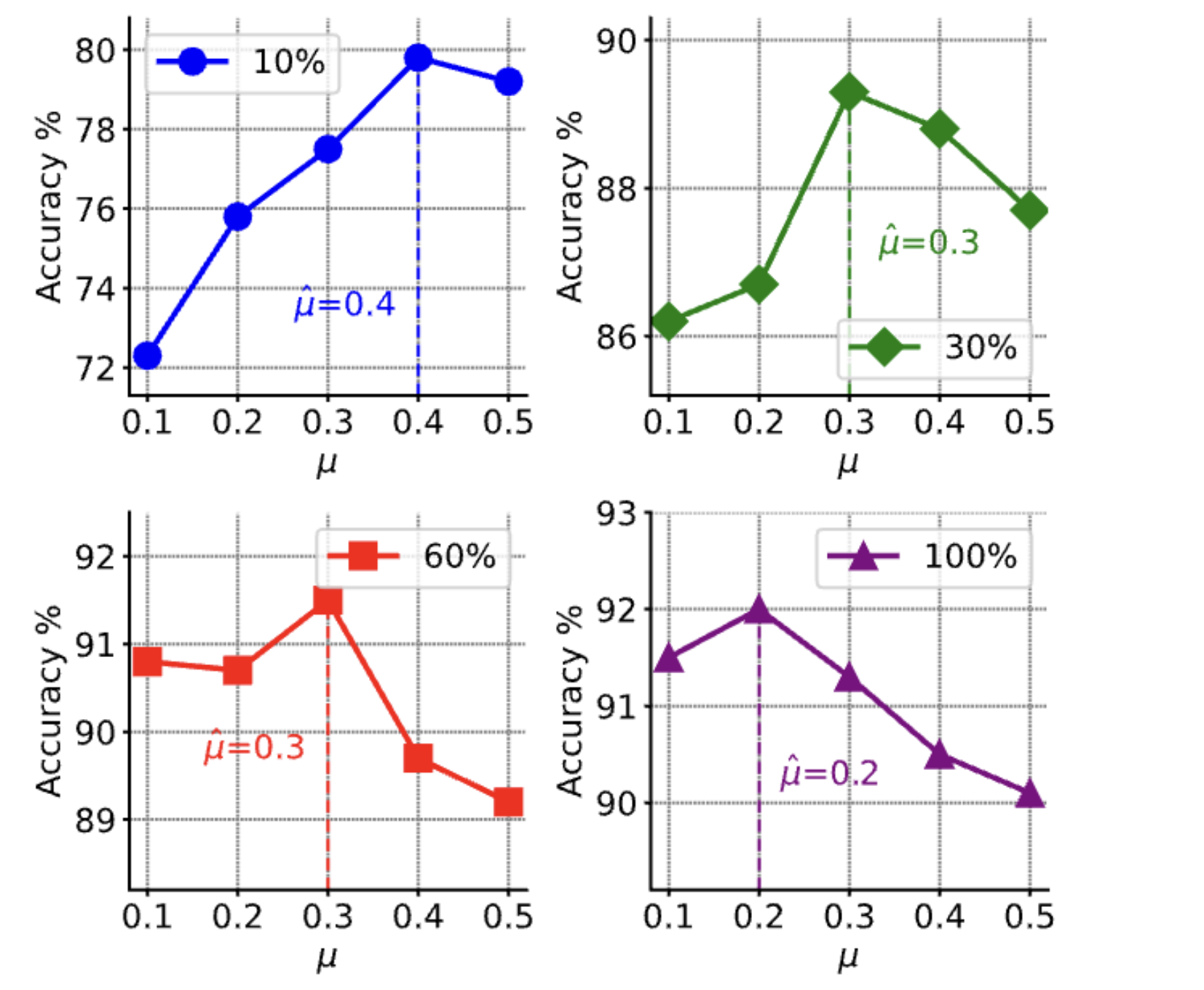
- **Comparison with other decoding strategies:**

Model	Method	Multiple Choices			Math Reasoning	
		MC1	MC2	MC3	GSM8K	MultiArith
LLaMa-7b	LoRA	32.9	55.0	28.5	26.6	90.5
	+*DoLa*	31.6	48.6	22.7	26.6	89.7
	+*TaD*	**34.2**	**55.7**	**29.0**	**27.7**	**91.0**
	AdapterP	38.1	57.4	30.8	31.5	93.5
	+*DoLa*	39.7	54.9	25.5	31.5	93.3
	+*TaD*	**40.6**	**58.5**	**32.1**	**32.0**	**93.7**
LLaMa-13b	LoRA	33.4	55.7	29.0	35.9	91.5
	+*CD*	**36.2**	55.4	26.5	19.0	70.3
	+*DoLa*	34.9	51.2	24.8	38.0	**94.2**
	+*TaD*	35.1	**56.7**	**29.7**	**38.1**	92.0
	AdapterP	40.6	58.8	32.4	36.8	91.5
	+*CD*	41.1	56.0	26.2	17.8	72.5
	+*DoLa*	41.3	56.5	27.5	35.9	93.5
	+*TaD*	**42.6**	**60.0**	**33.1**	**37.5**	**94.0**

- **Integrated with different basic decoding strategies:**

Model	Method	G / M	Model	Method	G / M
LLaMa-7b	Greedy	26.6/90.5	LLaMa-13b	Greedy	35.9/91.5
	+*TaD*	**27.7/91.0**		+*TaD*	**38.1/92.0**
	Beam-4	30.5/91.3		Beam-4	43.6/93.3
	+*TaD*	**30.9/91.8**		+*TaD*	**43.7/94.3**
	Top-p	26.7/90.7		Top-p	36.7/91.7
	+*TaD*	**27.4/91.3**		+*TaD*	**37.1/93.0**
	Top-k	27.0/90.3		Top-k	36.8/91.7
	+*TaD*	**27.7/91.6**		+*TaD*	**37.2/93.0**

- **Different ratios of training data and the selection of step:**



School of Software, BNRist, Tsinghua University, China

Mail: xxh22@mails.tsinghua.edu.cn