IJCAI 2024
The 33rd International Joint Conference on Artificial Intelligence

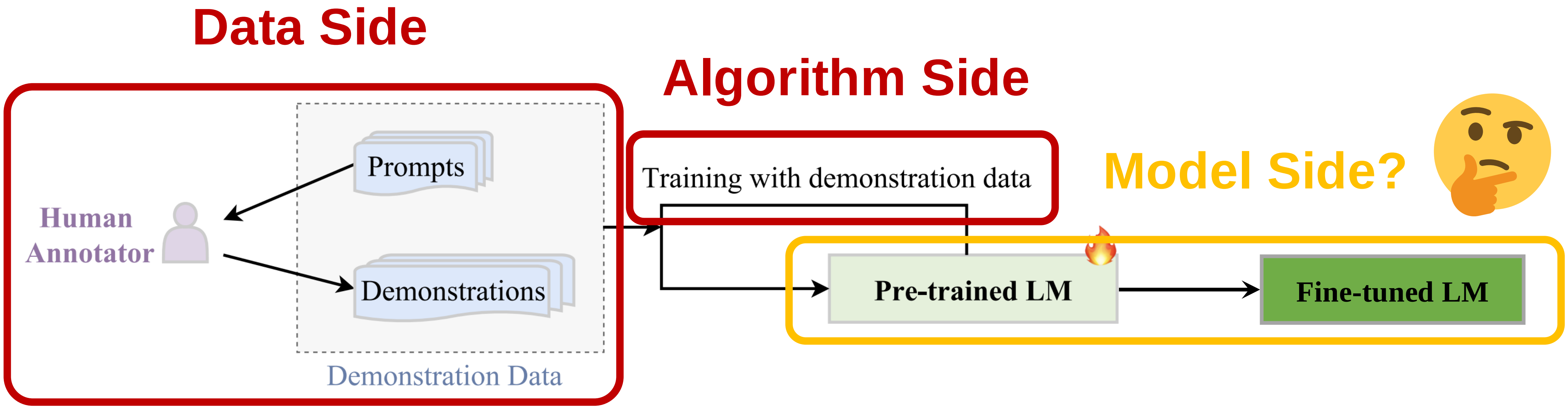
# TaD: A Plug-and-Play Task-Aware Decoding Method to Better Adapt LLMs on Downstream Tasks

**Xinhao Xu**[1,2], **Hui Chen**[2*], **Zijia Lin**[1], **Jungong Han**[3], **Lixing Gong**[4], **Guoxin Wang**[4], **Yongjun Bao**[4] and **Guiguang Ding**[1,2]
[1]School of Software, Tsinghua University
[2]Beijing National Research Center for Information Science and Technology (BNRist)
[3]Department of Computer Science, University of Sheffield
[4]JD.com

# Large Language Models in Downstream Tasks

**A common strategy:**

Fine-tuning pre-trained LLMs with downstream demonstration data.




Zhao W X, Zhou K, Li J, et al. A survey of large language models, 2023.

# Inherent Knowledge Acquisition of Fine-tuned LLMs

**LLM Outputs ≠ Knowledge:**

LLMs can possess correct knowledge even if their outputs are incorrect.

```
Question: Who was the third president of the United States?
Here are some brainstormed ideas: James Monroe\n Thomas Jefferson\n Jefferson\n
    Thomas Jefferson\n George Washington
Possible Answer: James Monroe
Is the possible answer:
 (A) True
 (B) False
The possible answer is: (B)
```

**Incorrect Output** (for "Possible Answer: James Monroe")

**Correct Evaluation** (for "The possible answer is: (B)")

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

Kadavath S, Conerly T, Askell A, et al. Language models (mostly) know what they know, 2022.

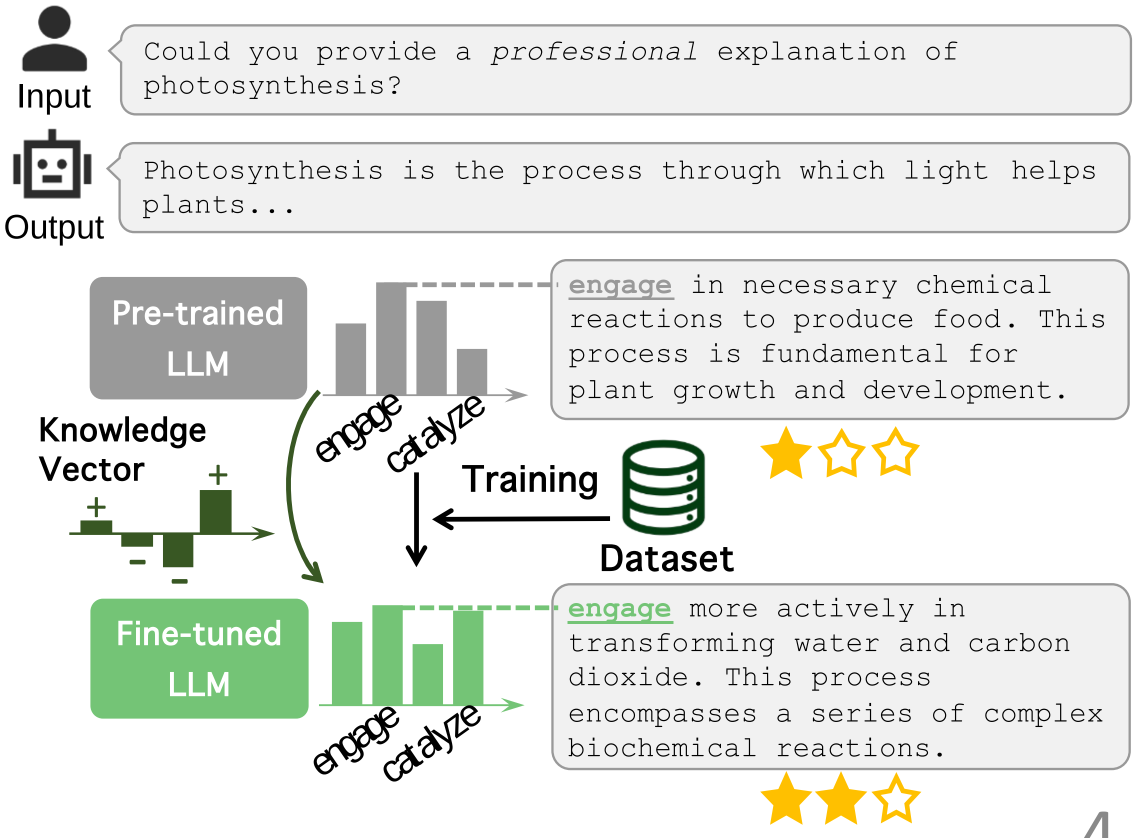
# Our work: Plug and Play Task-Aware Decoding

**Intuitive Ideas:**

- Token-predicting **alterations** during fine-tuning reflect the inherent knowledge.
- Such alterations indicate an adaptive shift from **common** knowledge to **task-specific** knowledge.

## Knowledge Vector

Explicitly denoting the direction of knowledge adaptation learned during fine-tuning, naturally with semantic information.



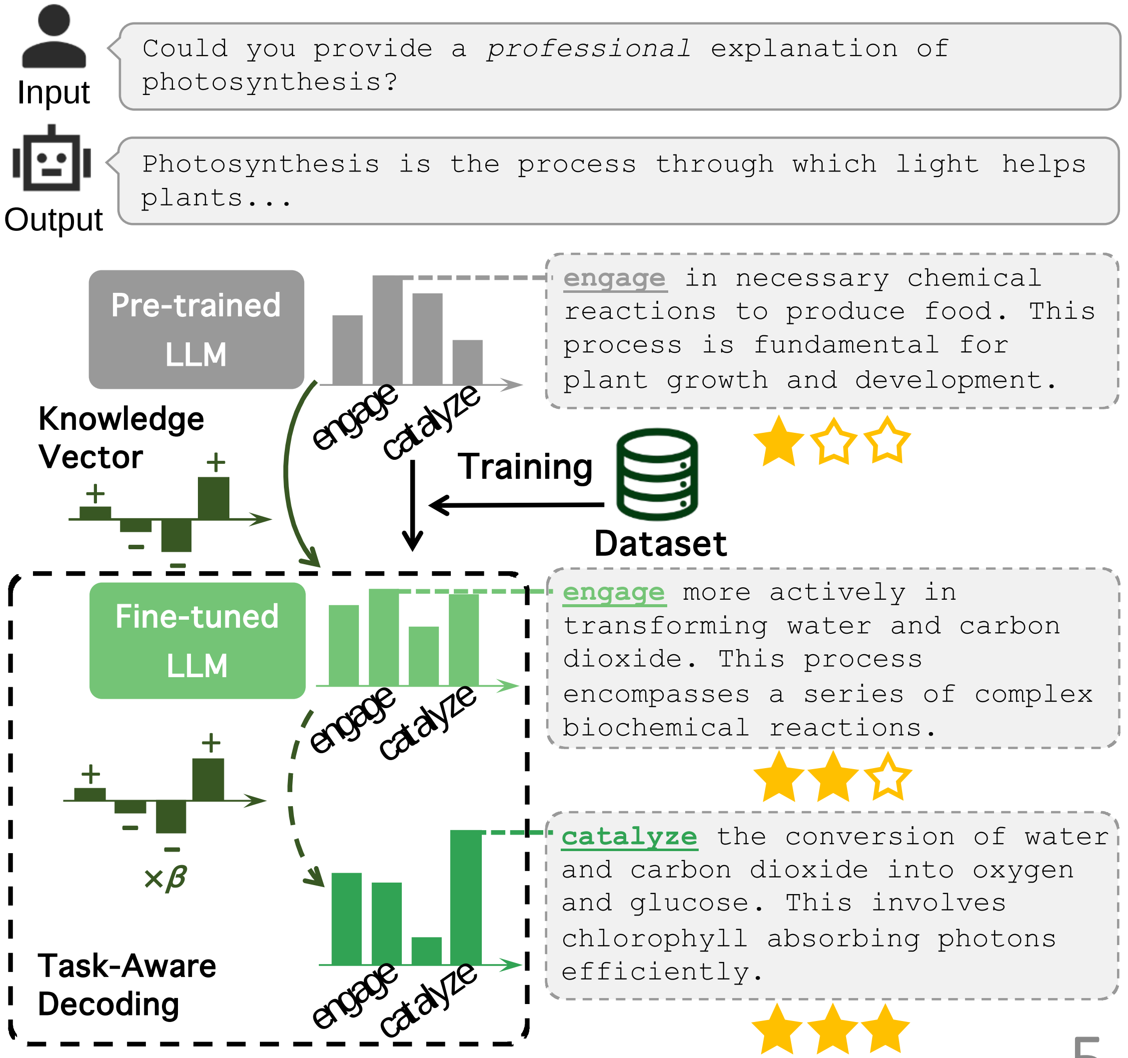
# Our work: Plug and Play Task-Aware Decoding

## Task-Aware Decoding (TaD):

- 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.

## Features

- A plug and play method
- Promising potential in data-scarce scenarios.



# Task-Aware Decoding: An Implementation

## Constructing the Knowledge Vector

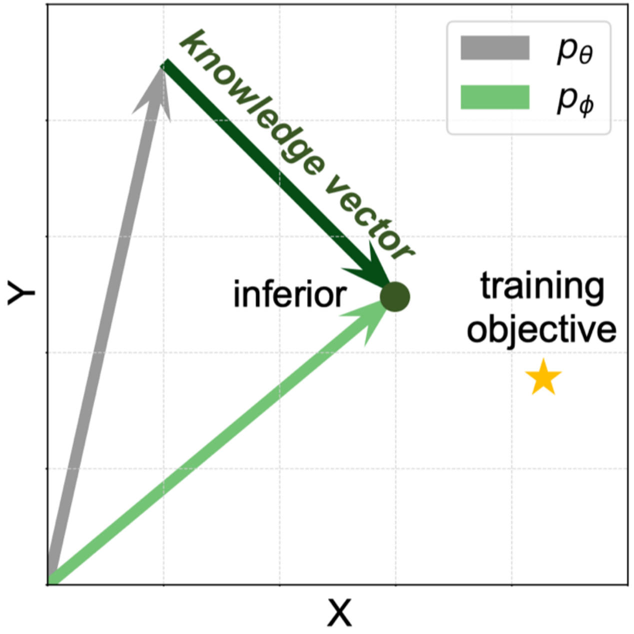
Pre-trained LLM: $p_\theta(x_t|x_{<t}),\ x_t \in \mathcal{V}$

**Fine-tuning** $\phi = \Phi(\theta, \mathcal{D})$

conditional probability distribution for tokens

($\mathcal{V}$ denotes the vocabulary)

Fine-tuned LLM: $p_\phi(x_t|x_{<t}),\ x_t \in \mathcal{V}$




$$\mathcal{V}_K = p_\mathcal{E} - p_\mathcal{S}$$
$$= \log p_\phi(x_t|x_{<t}) - \log p_\theta(x_t|x_{<t})$$

**$|\mathcal{V}|$-dimensional vector**

### Constraint Function to Avoid False Positive Cases

$$\mathcal{C}_t = \{x_t \in \mathcal{V} : p_\phi(x_t|x_{<t}) \geq \alpha \max_{x_t' \in \mathcal{V}} p_\phi(x_t'|x_{<t})\}$$

$$\mathbb{I}(x_t) = \begin{cases} 1 & \text{if } x_t \in \mathcal{C}_t \\ 0 & \text{otherwise} \end{cases}$$

### Knowledge Vector w/ penalty coefficient

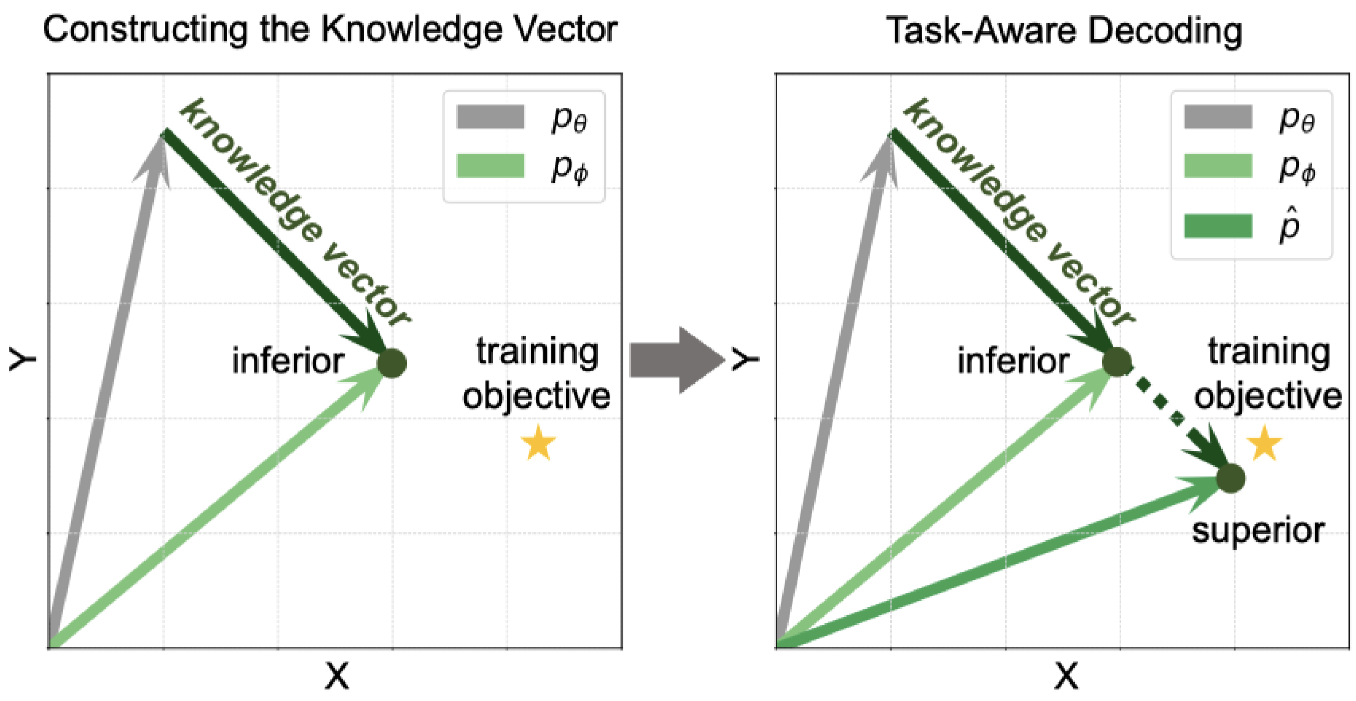
$$\hat{\mathcal{V}}_K = \mathbb{I}(x_t) \cdot \mathcal{V}_K + (1 - \mathbb{I}(x_t)) \cdot \lambda$$

# Task-Aware Decoding: An Implementation

## Task-Aware Decoding (TaD):

Convert the knowledge vector into a probability distribution

$$p_K(x_t|x_{<t}) = \mathrm{softmax}(\hat{\mathcal{V}}_K)$$




**TaD's Final Output Token Distribution**

$$\hat{p}(x_t|x_{<t}) = (1-\mu)\cdot p_\phi(x_t|x_{<t}) + \mu \cdot p_K(x_t|x_{<t})$$

# Main Results: Multiple-Choice & Generation 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** |

Multiple-choice and CBQA 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** |

Reasoning tasks

**TaD yields considerable improvements in nearly all cases.**

# Other Results: Superiority and Integration Capability

| 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** |

**TaD outperforms the baselines in most cases.**

| 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** |

**TaD consistently improves upon the fine-tuned LLMs' performance across different basic decoding strategies.**

# Analysis: Ablation and Data-Scarce Scenarios Study

## Ablation study of the knowledge vector

| $\mathcal{M}$ | $p_S \to p_{\mathcal{E}}$ | 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.

| $\mathcal{M}$ | $p_S \to p_{\mathcal{E}}$ | 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.

| $\mathcal{M}$ | $p_S \to p_{\mathcal{E}}$ | 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).

| $\mathcal{M}$ | $p_S \to p_{\mathcal{E}}$ | 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).

| $\mathcal{M}$ | $p_S \to p_{\mathcal{E}}$ | 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.

| $\mathcal{M}$ | $p_S \to p_{\mathcal{E}}$ | 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.

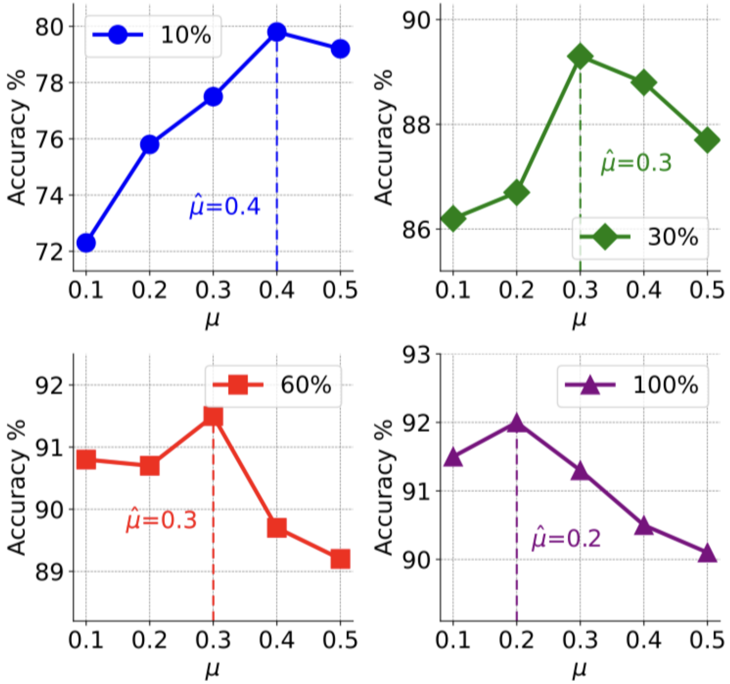
## Different ratios of training datasets and the selection of $\mu$

**Training data ↓**

**TaD's gains ↑**

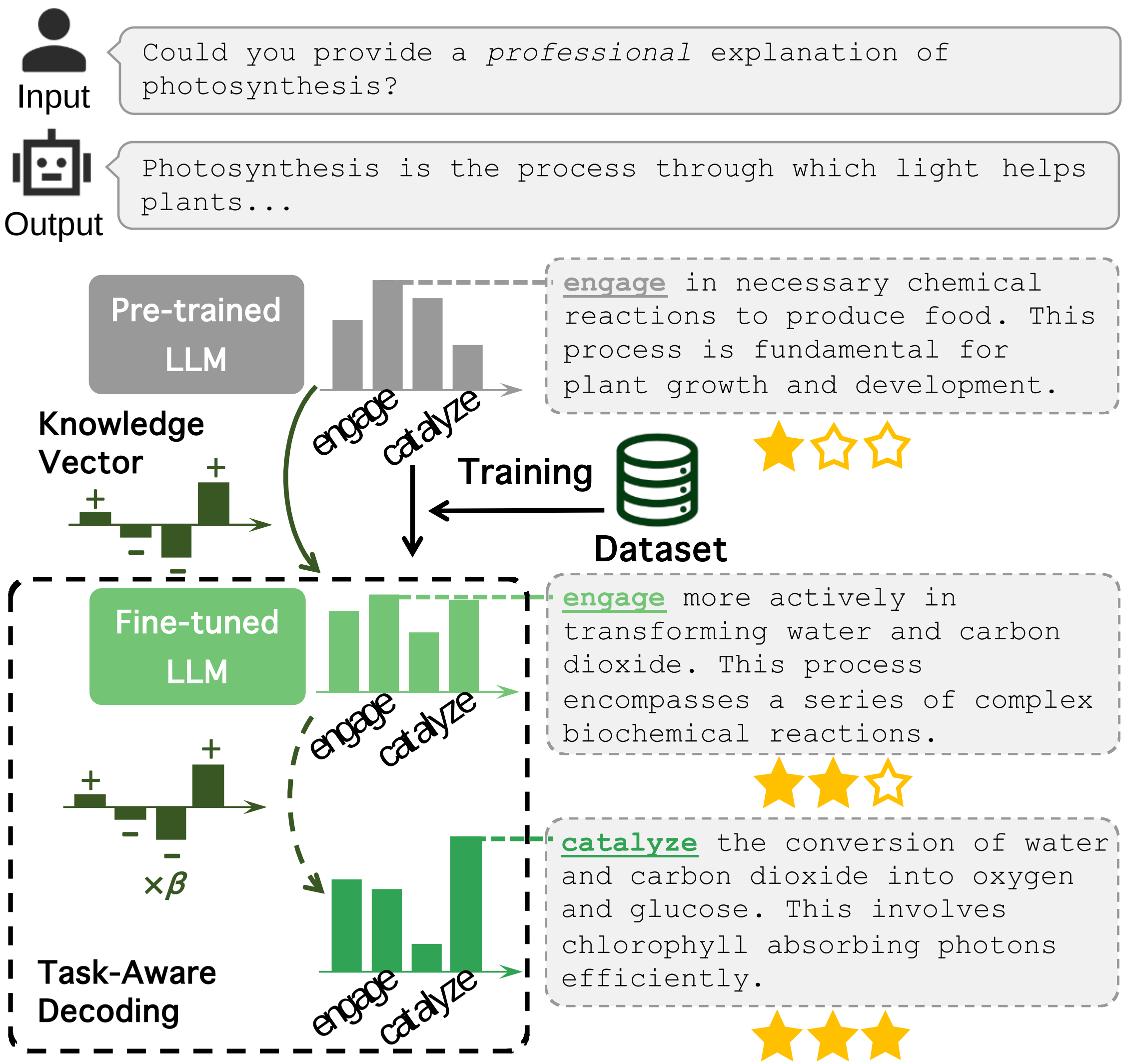
**Optimal $\mu$ ↑**

| Model | Method | 10% | 30% | 60% | 100% |
|---|---|---|---|---|---|
| LLaMa-7b | LoRA | 58.8 | 80.5 | 86.2 | 90.5 |
| | +*TaD* | **62.2** | **83.2** | **88.0** | **91.0** |
| | Δ | **+3.4** | **+2.7** | **+1.8** | **+0.5** |
| LLaMa-13b | LoRA | 70.8 | 86.5 | 90.2 | 91.5 |
| | +*TaD* | **79.8** | **89.3** | **91.5** | **92.0** |
| | Δ | **+9.0** | **+3.2** | **+1.3** | **+0.5** |



# Summary

- A concept of **knowledge vector**, explicitly denoting the knowledge adaptation learned by LLMs during fine-tuning.
- TaD enhancing fine-tuned LLMs' **output probability distribution** with the knowledge vector.
- Effective across various tasks, models, and finetuning methods, superior to baselines and showing promising potential in **data-scarce scenarios**.



# Thank You!

Contact: xxh22@mails.tsinghua.edu.cn

Multimedia Intelligence Group, School of Software, Tsinghua University

http://ise.thss.tsinghua.edu.cn/MIG/index.html