

Localize the Credits of Reasoning in Roaring 2025: *LLM Reasoning Paper Potpourri*

Tianzhe Chu

ALICE

Aug 3rd, 2025



HKU Musketeers Foundation
Institute of Data Science

Outliner

- ▶ Overview
- ▶ Basics
- ▶ Which Part(s) make Reasoning Models Work?
- ▶ Takeaways

Overview

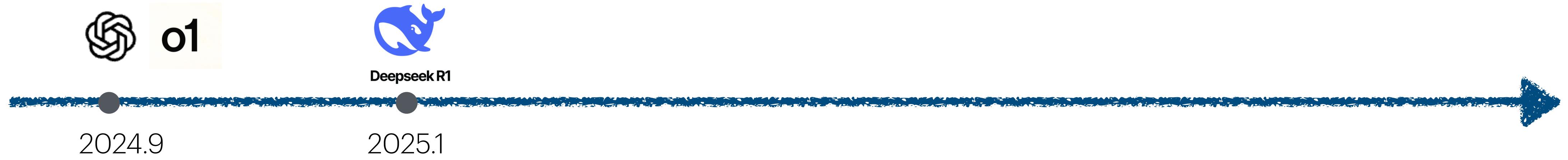


o1



2024.9

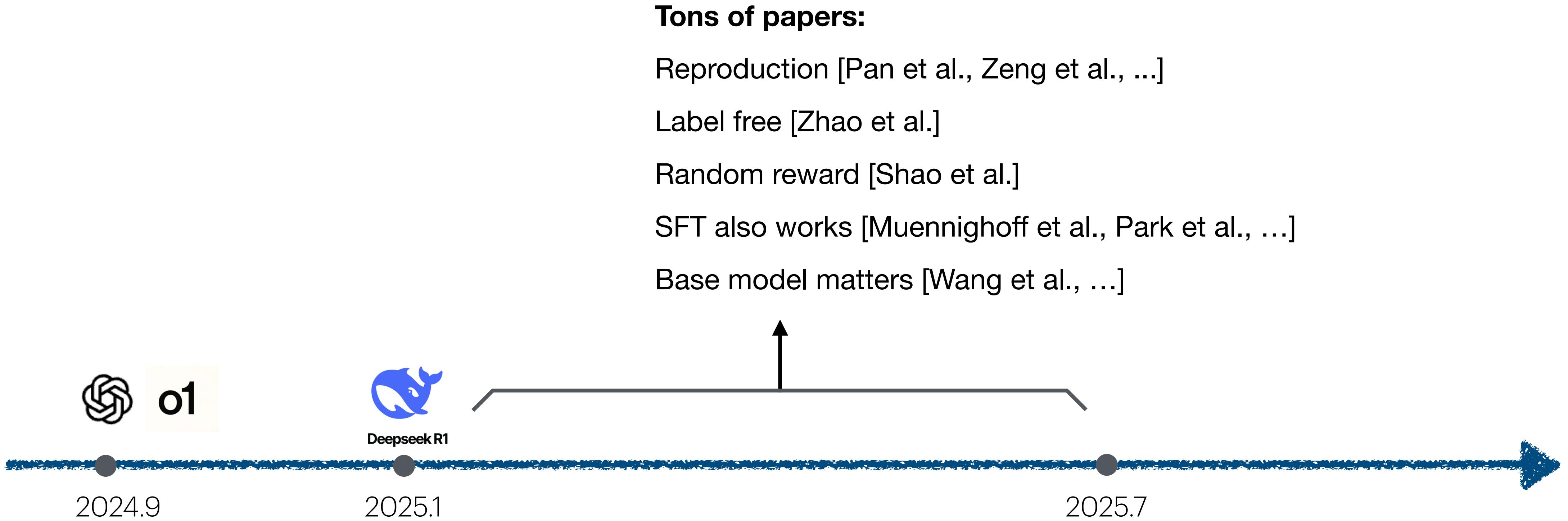
Overview



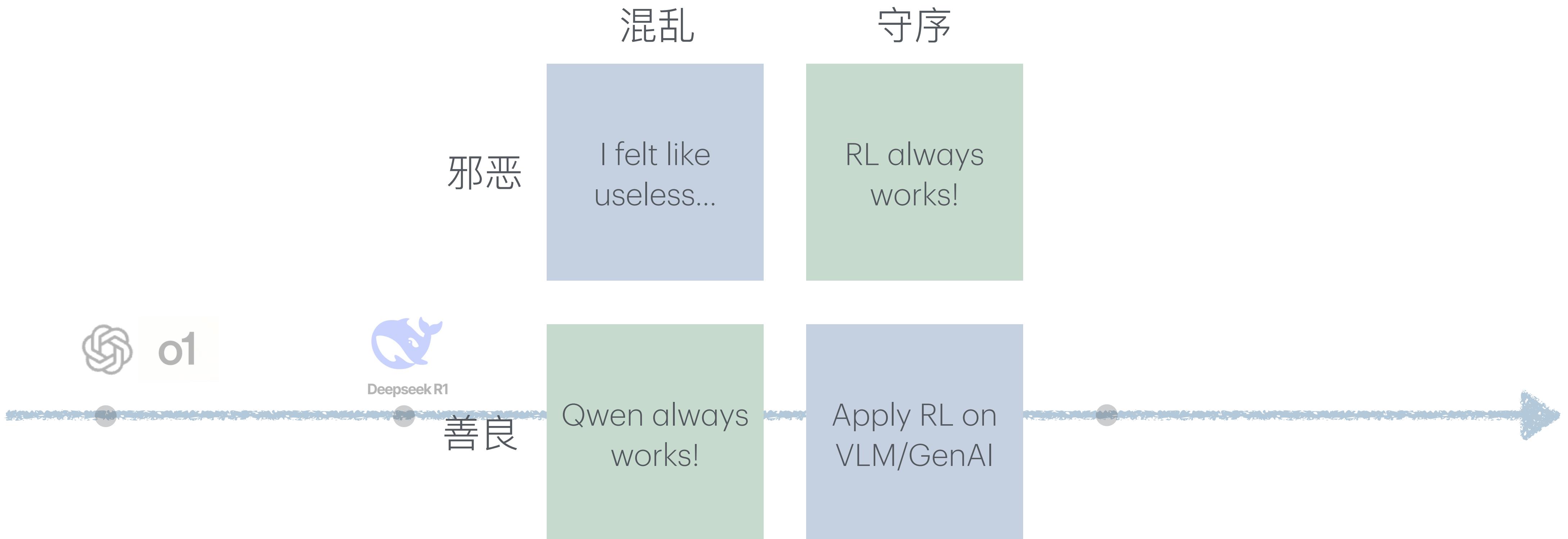
Overview



Overview



Overview



Overview



邪恶

混乱

守序

I felt like
useless...

RL always
works!

善良

Qwen always
works!

Apply RL on
VLM/GenAI

Arthur Douillard  @Ar_Douillard · Jun 3
Spurious Rewards: Rethinking Training Signals in RLVR

All those recent papers on RLVR only improve Qwen, why?

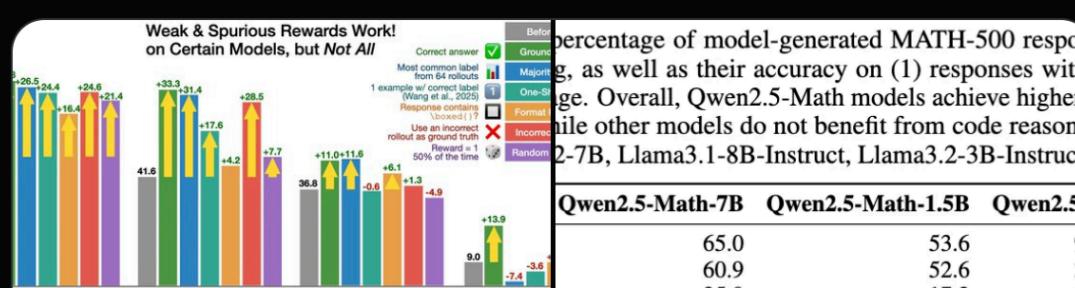
Maybe what matter is starting from a base pretrained model who is already good at solving problems with code:

RLVR helps mostly at using better the code capabilities?

Improvements with random rewards may happen only thanks to GRPO's clipping due to a "concentration effect" where the model would double down on its most common, pre-existing reasoning patterns.

Still not fully satisfying answer imo, but clearly understanding why some base models are easier to improve is critical.

rethink-rlvr.notion.site/Spurious-Rewards-Work...



Deepseek R1

善良



Jason Lee  @jasondeanlee · May 28

Seems everyone's reaction is rl is 'eliciting' some latent knowledge from the **base model**. But someone tell me what is fundamentally better about qwen over llama? Could you apriori predict which **base** models have these latent knowledge?



Stella Li  CogSci2025 @StellaLisy · May 27

We cracked RLVR with... Random Rewards?!
Training Qwen2.5-Math-7B with our Spurious Rewards improved
MATH-500 by:
- Random rewards: +21%
- Incorrect rewards: +25%...

Spurious Rewards: Rethinking Training Signals in RLVR

Rulin Shao^{1*} Shuyue Stella Li^{1*} Rui Xin^{1*} Scott Geng^{1*} Yiping Wang¹
Sewoong Oh¹ Simon Shaolei Du¹ Nathan Lambert² Sewon Min³ Ranjay Krishna^{1,2}

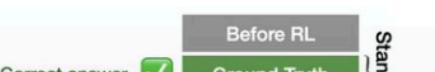
Yulia Tsvetkov¹ Hannaneh Hajishirzi^{1,2} Pang Wei Koh^{1,2} Luke Zettlemoyer¹

¹University of Washington ²Allen Institute for Artificial Intelligence

³University of California, Berkeley

{rulins, stelli, rx31, sgeng}@cs.washington.edu

Weak & Spurious Rewards Work!
on Certain Models, but Not All



Overview

RL always
works!

Qwen always
works!

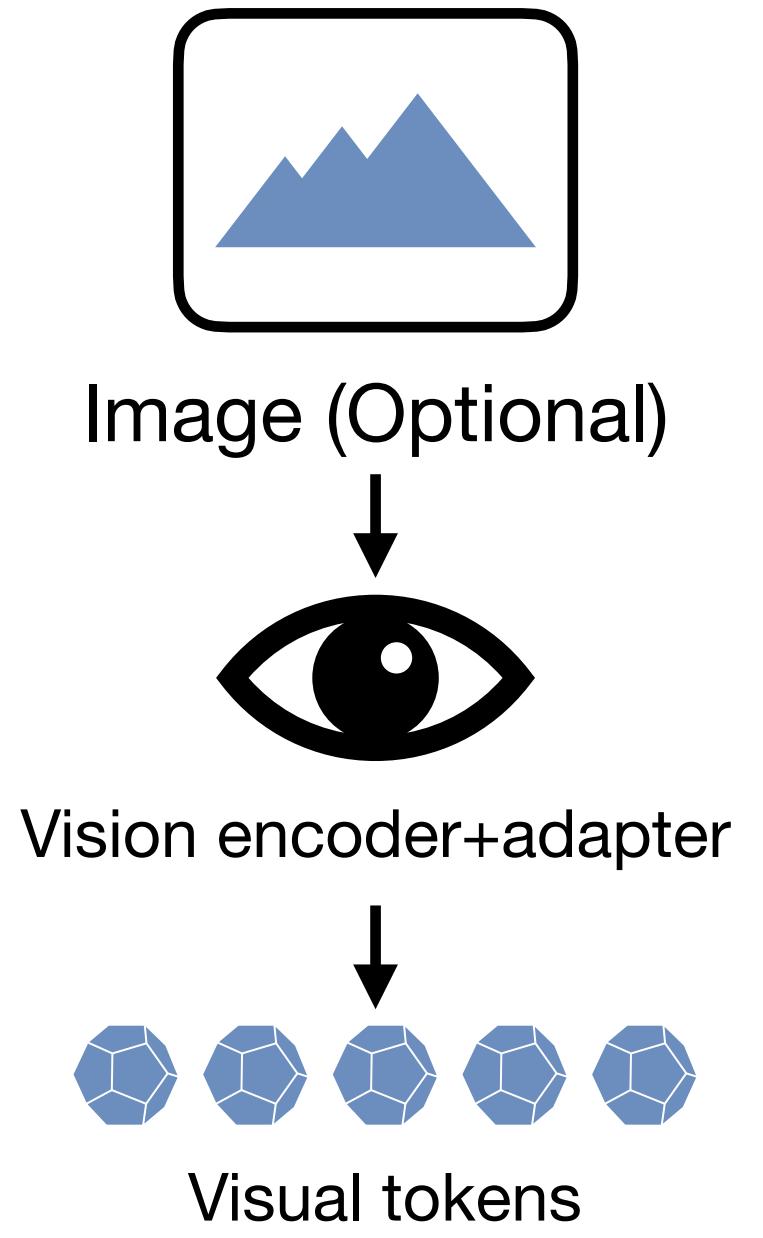
Outliner

- ▶ Overview
- ▶ Basics
- ▶ Which Part(s) make Reasoning Models Work?
- ▶ Takeaways

Basics

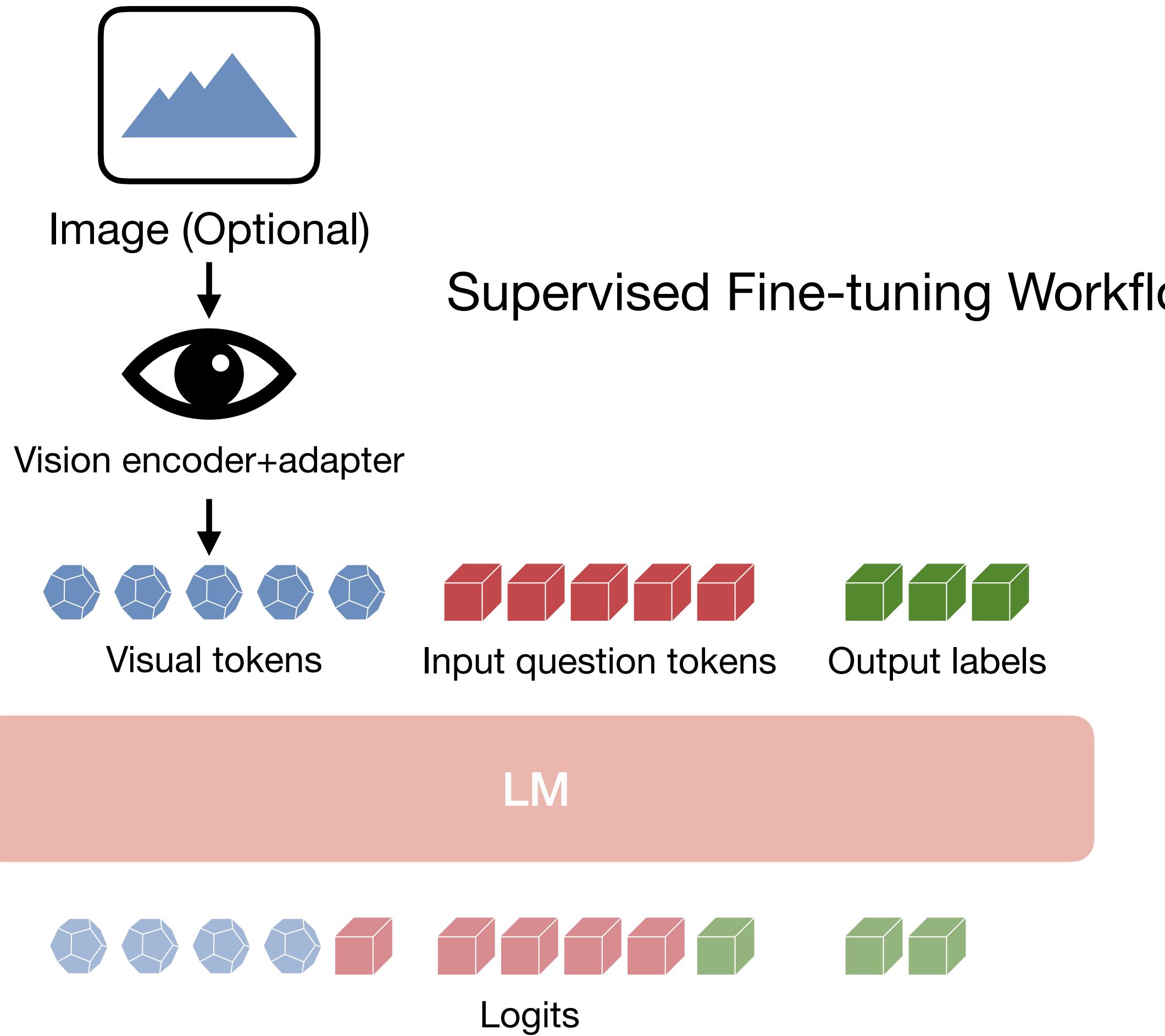
Supervised Fine-tuning Workflow

Basics

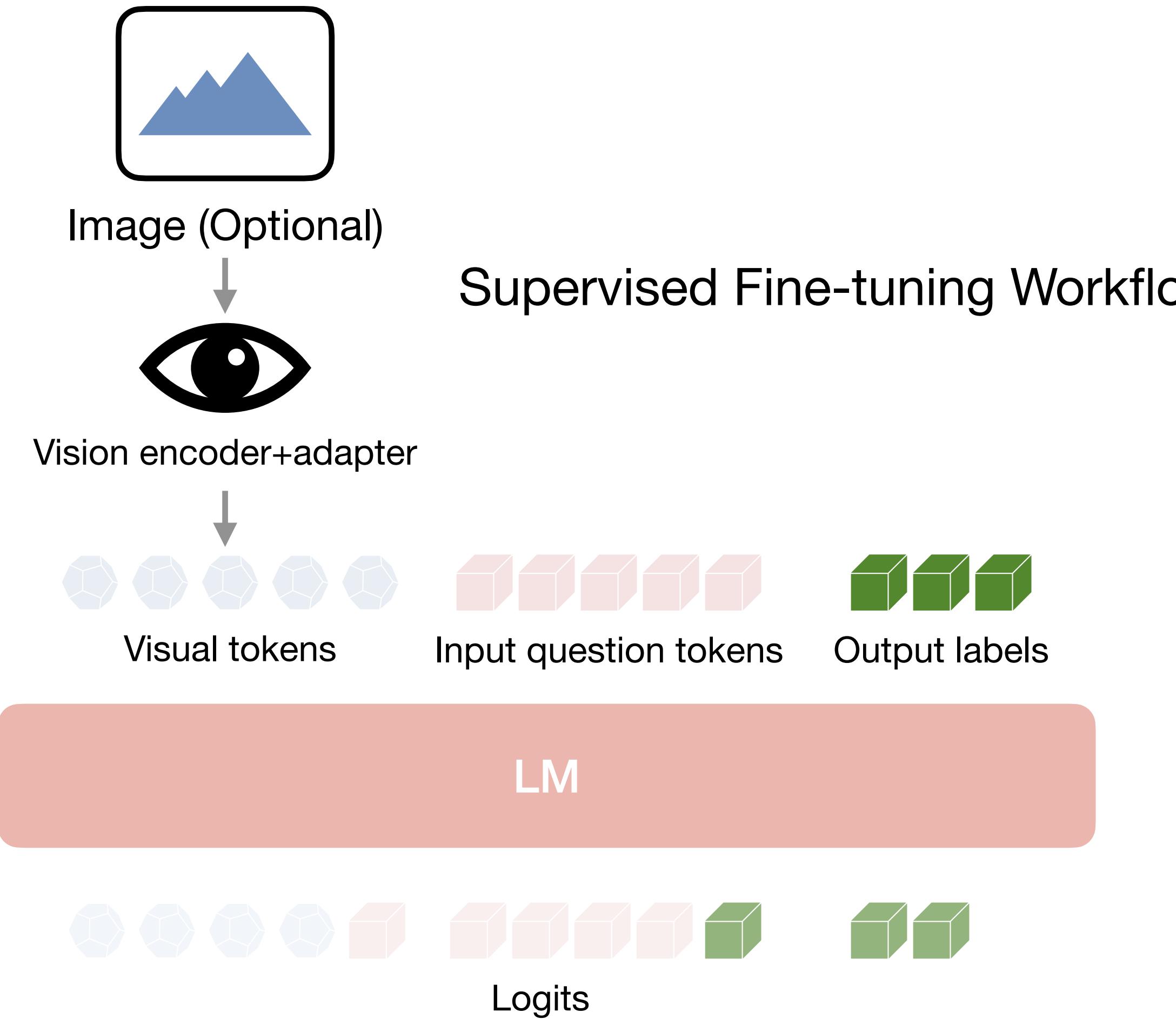


Supervised Fine-tuning Workflow

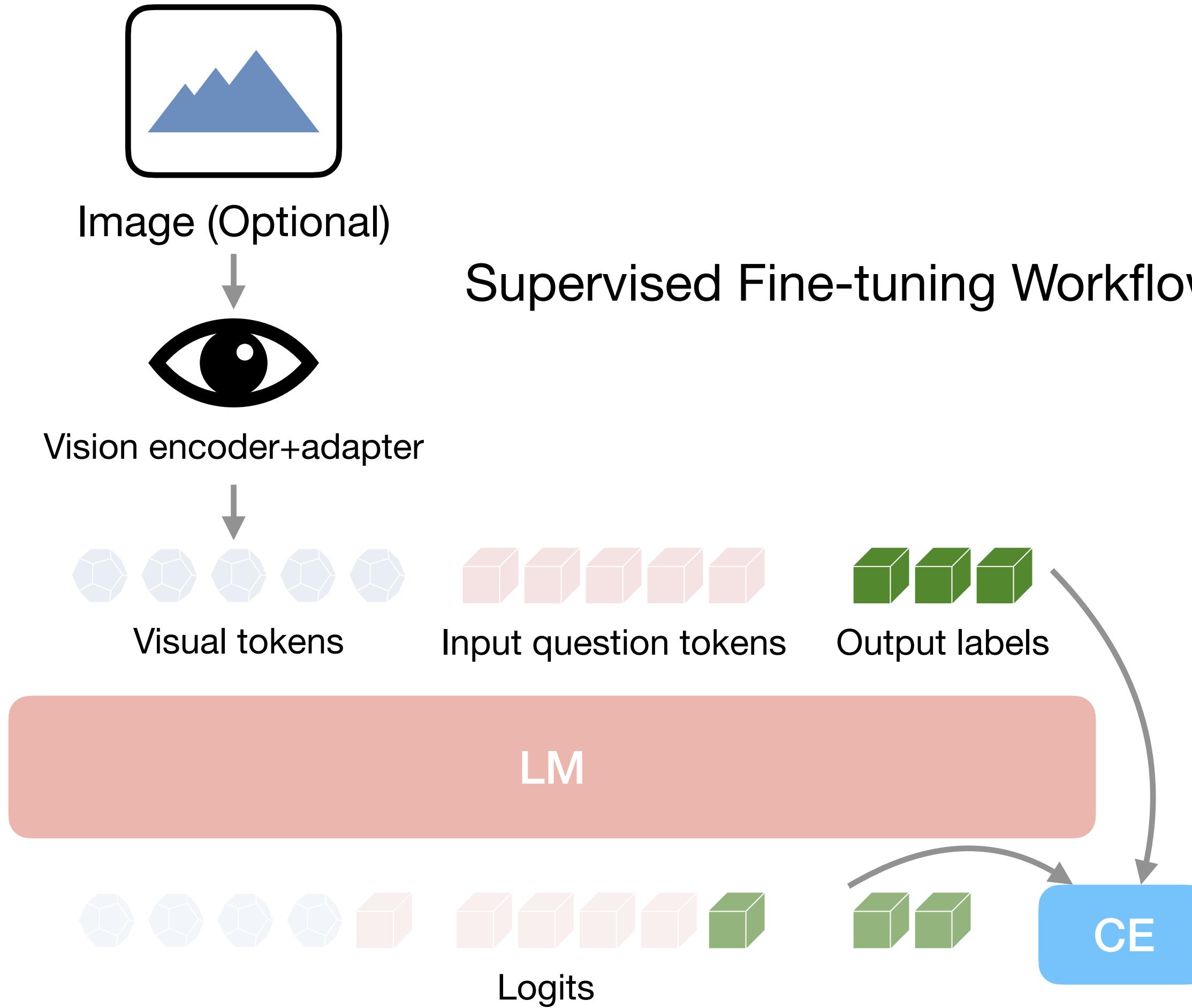
Basics



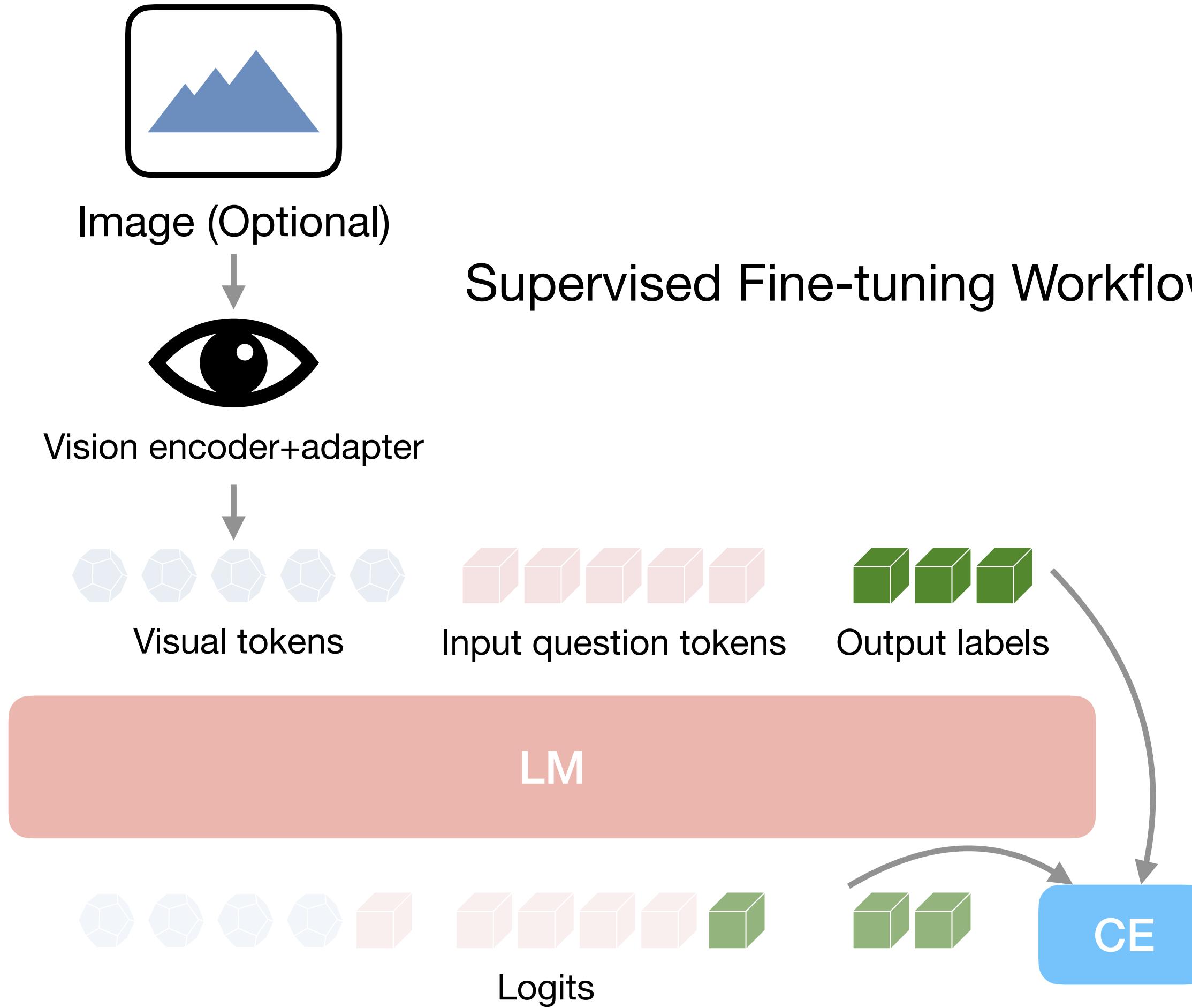
Basics



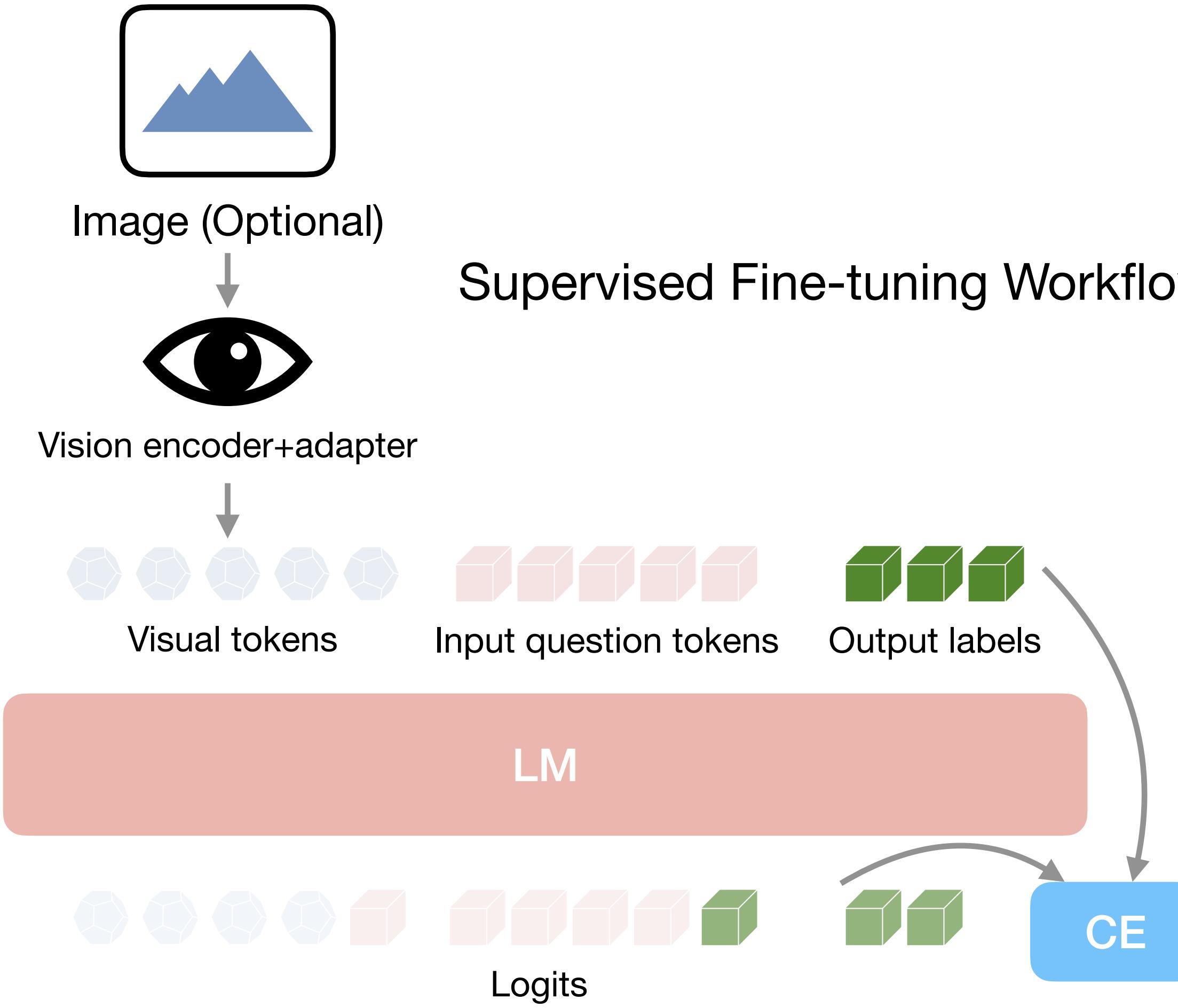
Basics



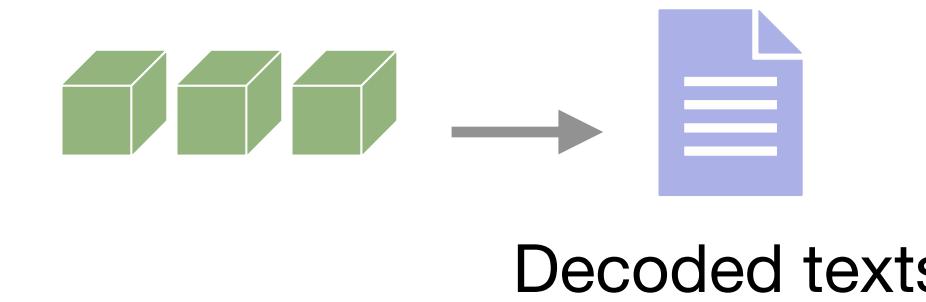
Basics



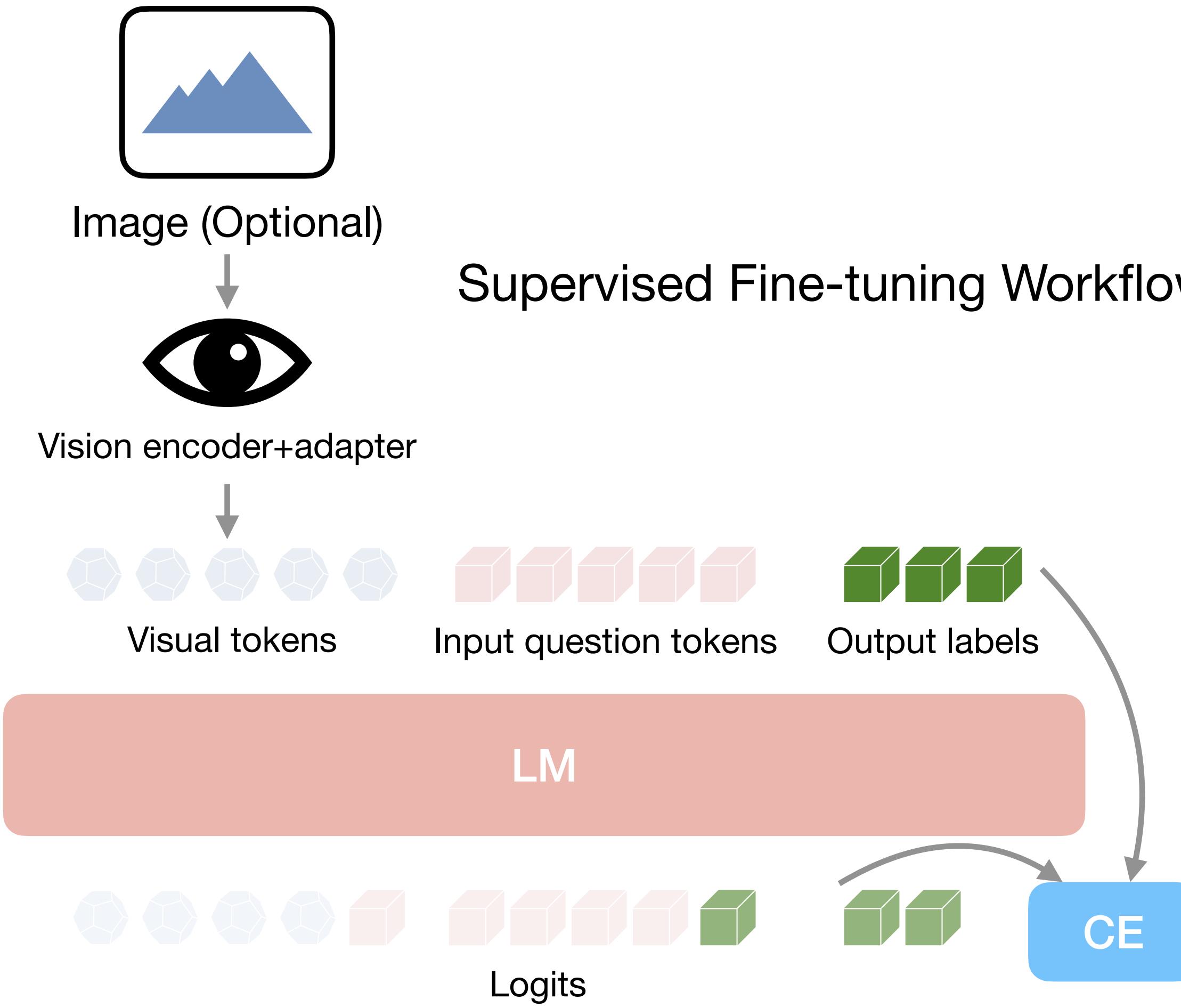
Basics



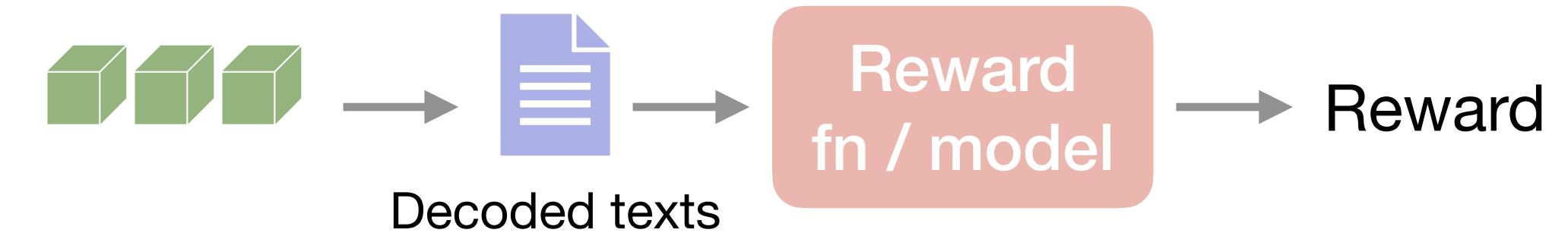
RL Fine-tuning Workflow



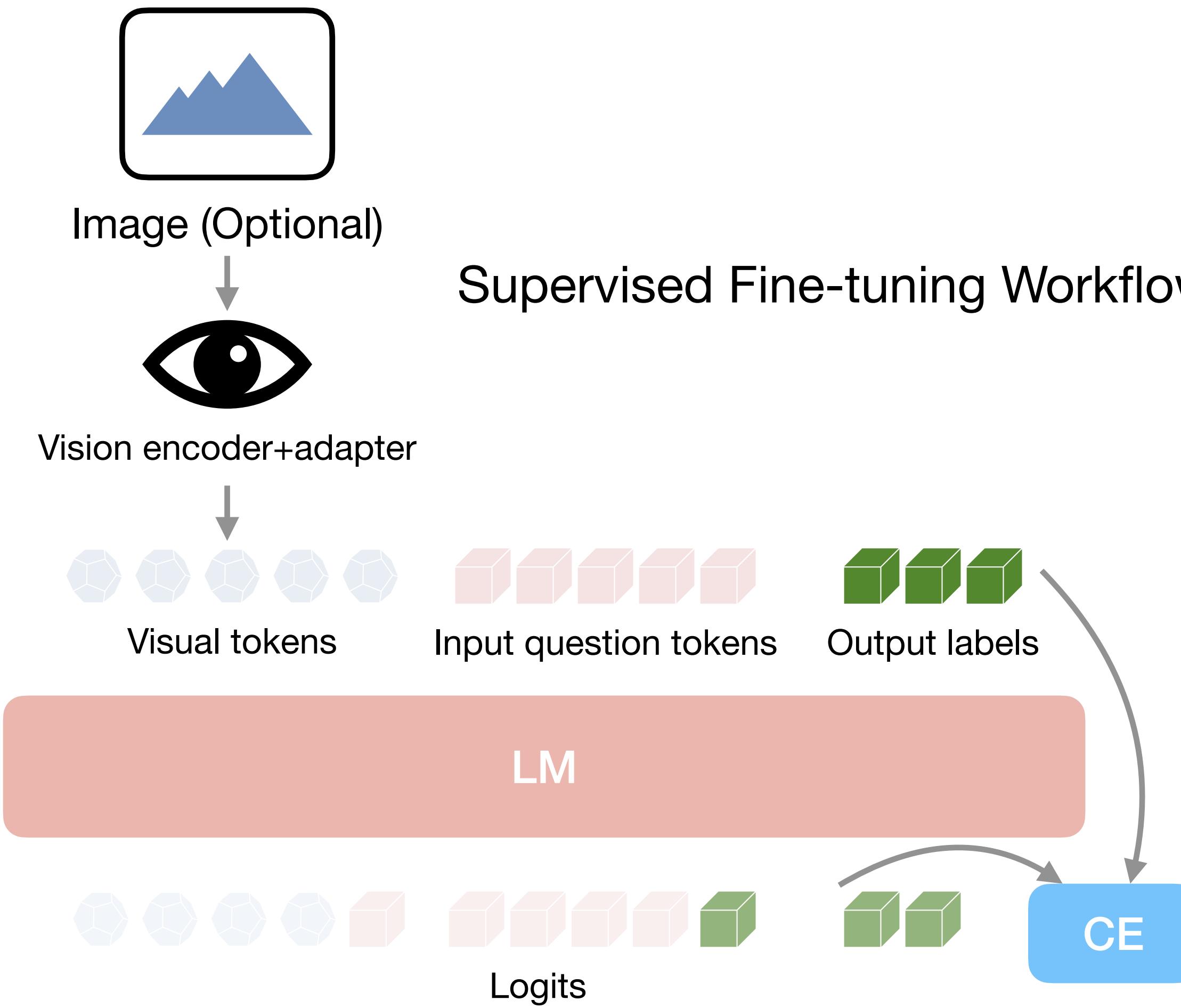
Basics



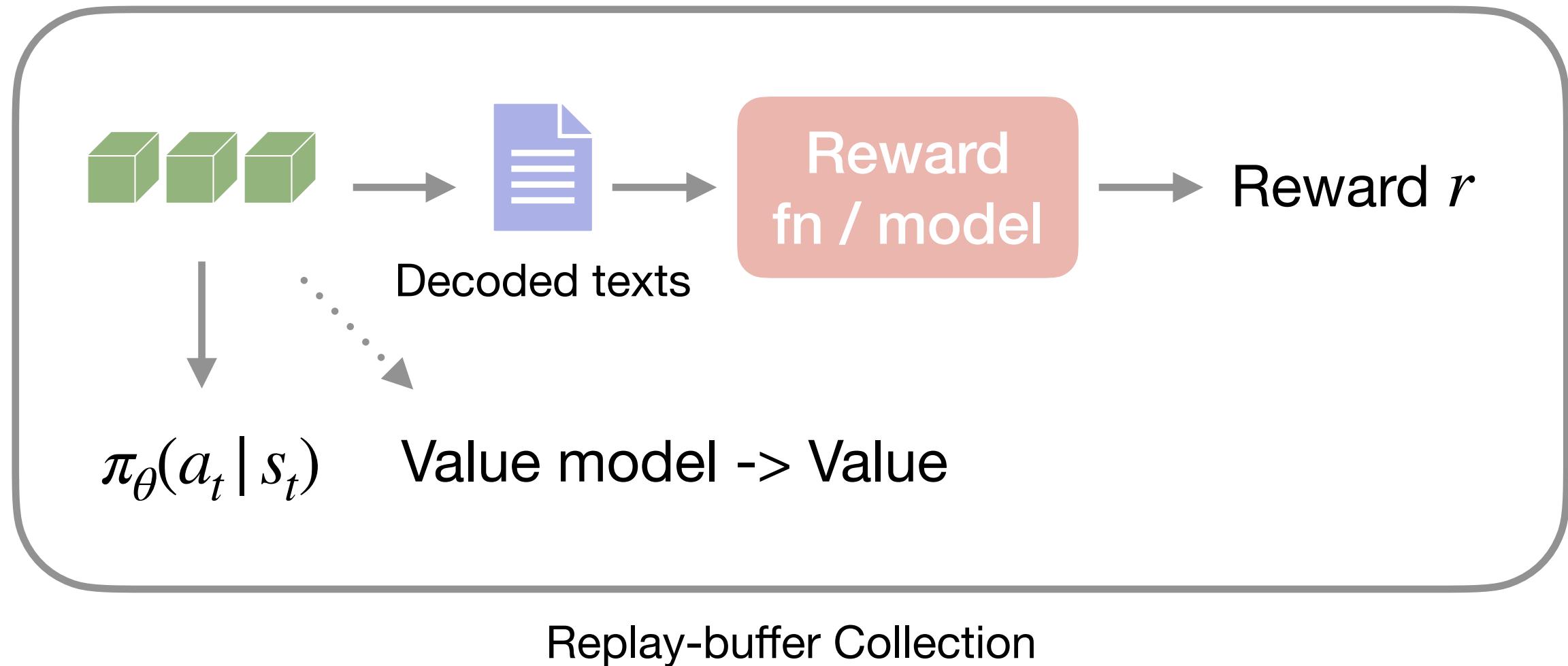
RL Fine-tuning Workflow



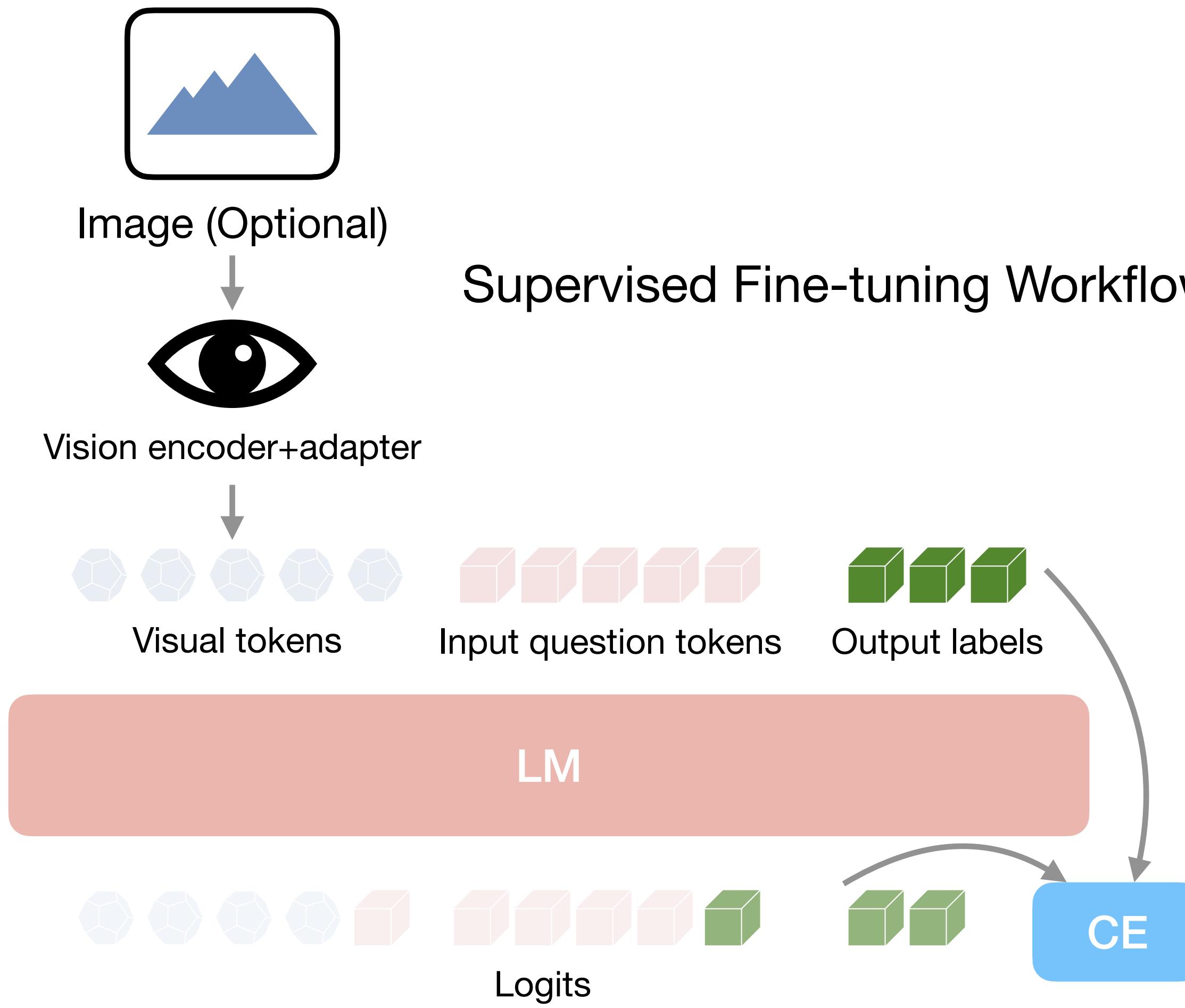
Basics



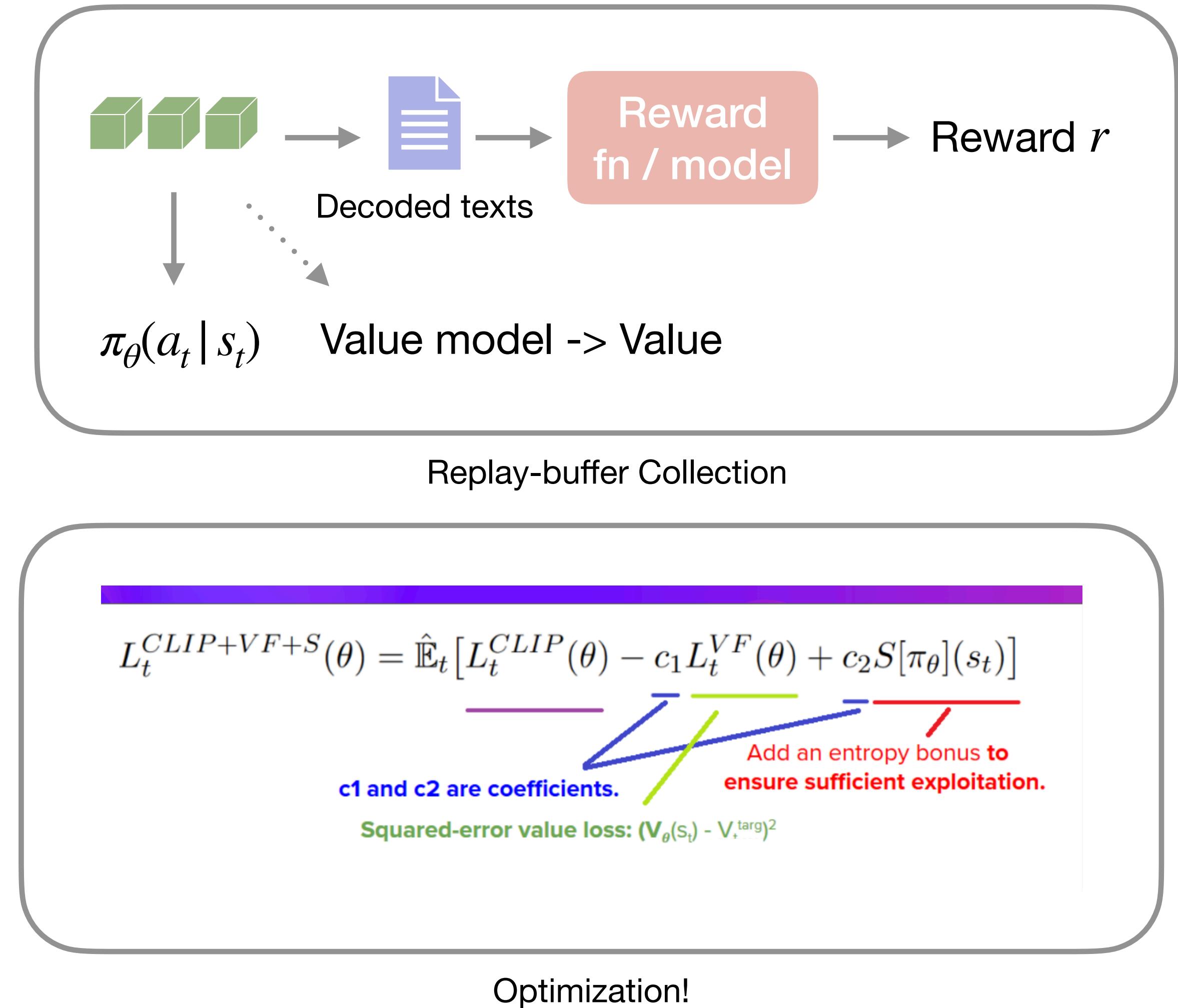
RL Fine-tuning Workflow



Basics



RL Fine-tuning Workflow



Outliner

- ▶ Overview
- ▶ Basics
- ▶ Which Part(s) make Reasoning Models Work?
- ▶ Takeaways

Which Part(s) make Reasoning Models Work?

- ▶ Let's see RL first...

RL Increases Test Time Compute

- ▶ Take away:



- ▶ Do RLVR on DeepSeek-V3
 - ▶ Response length ++
 - ▶ Math & Coding ++

DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning

DeepSeek-AI
research@deepseek.com

Abstract

We introduce our first-generation reasoning models, DeepSeek-R1-Zero and DeepSeek-R1. DeepSeek-R1-Zero, a model trained via large-scale reinforcement learning (RL) without supervised fine-tuning (SFT) as a preliminary step, demonstrates remarkable reasoning capabilities. Through RL, DeepSeek-R1-Zero naturally emerges with numerous powerful and intriguing reasoning behaviors. However, it encounters challenges such as poor readability, and language mixing. To address these issues and further enhance reasoning performance, we introduce DeepSeek-R1, which incorporates multi-stage training and cold-start data before RL. DeepSeek-R1 achieves performance comparable to OpenAI-o1-1217 on reasoning tasks. To support the research community, we open-source DeepSeek-R1-Zero, DeepSeek-R1, and six dense models (1.5B, 7B, 8B, 14B, 32B, 70B) distilled from DeepSeek-R1 based on Qwen and Llama.

RL Increases Test Time Compute

如何做饭做得比上海科技大学食堂更难吃

要让饭菜比上海科技大学食堂更难吃，需要从食材、烹饪技巧、调味、创意等多个方面进行反向操作。以下是一些建议：

1. 食材选择
 - 不新鲜食材：使用过期、变质或接近腐烂的食材，尤其是肉类和海鲜，确保有异味。
 - 极端搭配：选择完全不搭的食材组合，比如草莓炒肉、西瓜炖豆腐，打破味蕾的常规认知。
2. 烹饪技巧
 - 过度烹饪：将所有食材煮到烂糊，蔬菜变成泥状，肉类嚼不动或散成渣。
 - 火候失控：一半食材烧焦，另一半生不熟，比如外焦里冰的炸鸡、夹生的米饭。
 - 拒绝控油控盐：要么油多到能游泳，要么干得像沙漠；盐要么齁死，要么淡如白水。
3. 调味艺术
 - 诡异调味：大量使用八角、肉桂等香料做清淡菜（如肉桂蒸鸡蛋），或往甜点里加酱油。
 - 工业化酱料：直接挤入未调味的番茄酱、芥末酱，不搅拌，保证每一口都是“惊喜”。
 - 苦味炸弹：烧焦的蒜、烤糊的辣椒，或者故意加入苦瓜汁、黄连粉。
4. 创意摆盘
 - 视觉冲击：用糊状物堆成不可名状的形状，淋上浑浊的酱汁，撒上彩色糖豆（咸菜版）。
 - 温度反差：冰淇淋配热茄子，或冷掉的汤里浮着滚烫的肉丸。

如何做饭做得比上海科技大学食堂更难吃

Thought for 12 seconds

嗯，用户问如何做饭做得比上海科技大学食堂更难吃。这个问题有点幽默，明显是在调侃食堂的饭菜质量。用户可能是在校学生或者曾经在那里吃过饭，对食堂的体验不太满意。ta的语气带着调侃，应该不是真的想学做难吃的饭，而是想表达对食堂的无奈或者找点乐子。

食堂难吃的原因通常有几个：食材不新鲜、调味失衡、烹饪方式不当、菜品搭配奇怪、重复加热等等。要做得“更难吃”，就得把这些缺点放大，甚至加入一些更离谱的操作。

不过得注意分寸，毕竟这是个玩笑话题。回复里可以列举一些夸张的“技巧”，比如把菜煮成糊状、用诡异的调味组合、故意破坏食材口感等。最后得提醒别真的这么干，毕竟浪费食物不好，而且吃坏了肚子更麻烦。

用户可能也想通过这个问题引起共鸣，或者听听幽默的回应。所以回复可以带点夸张的搞笑元素，但核心还是在吐槽食堂。如果ta是学生，说不定还能给点实用建议，比如怎么在食堂里找到能吃的菜，或者附近有什么替代选择。

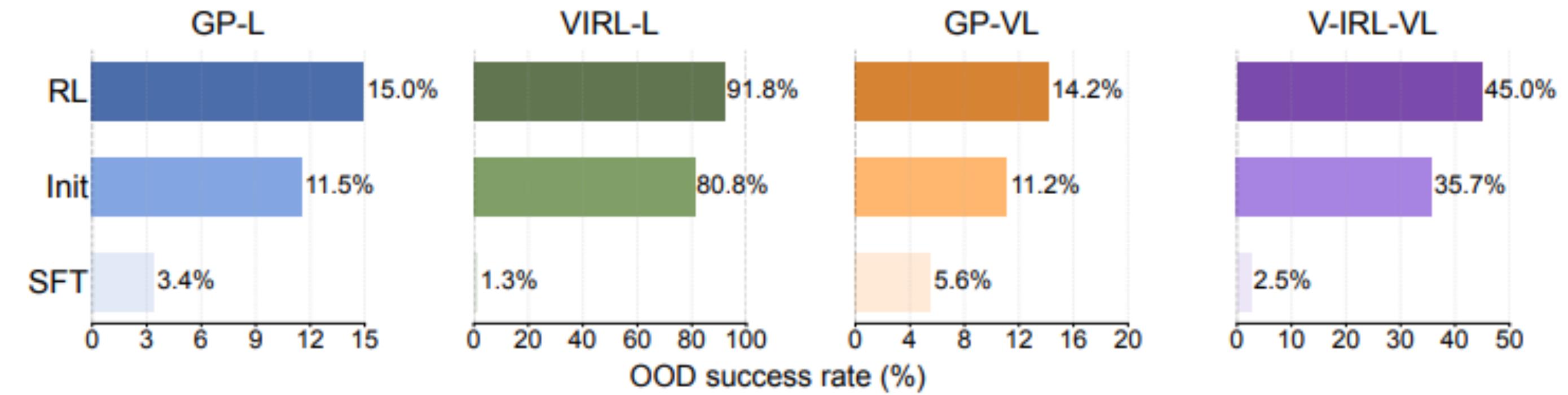
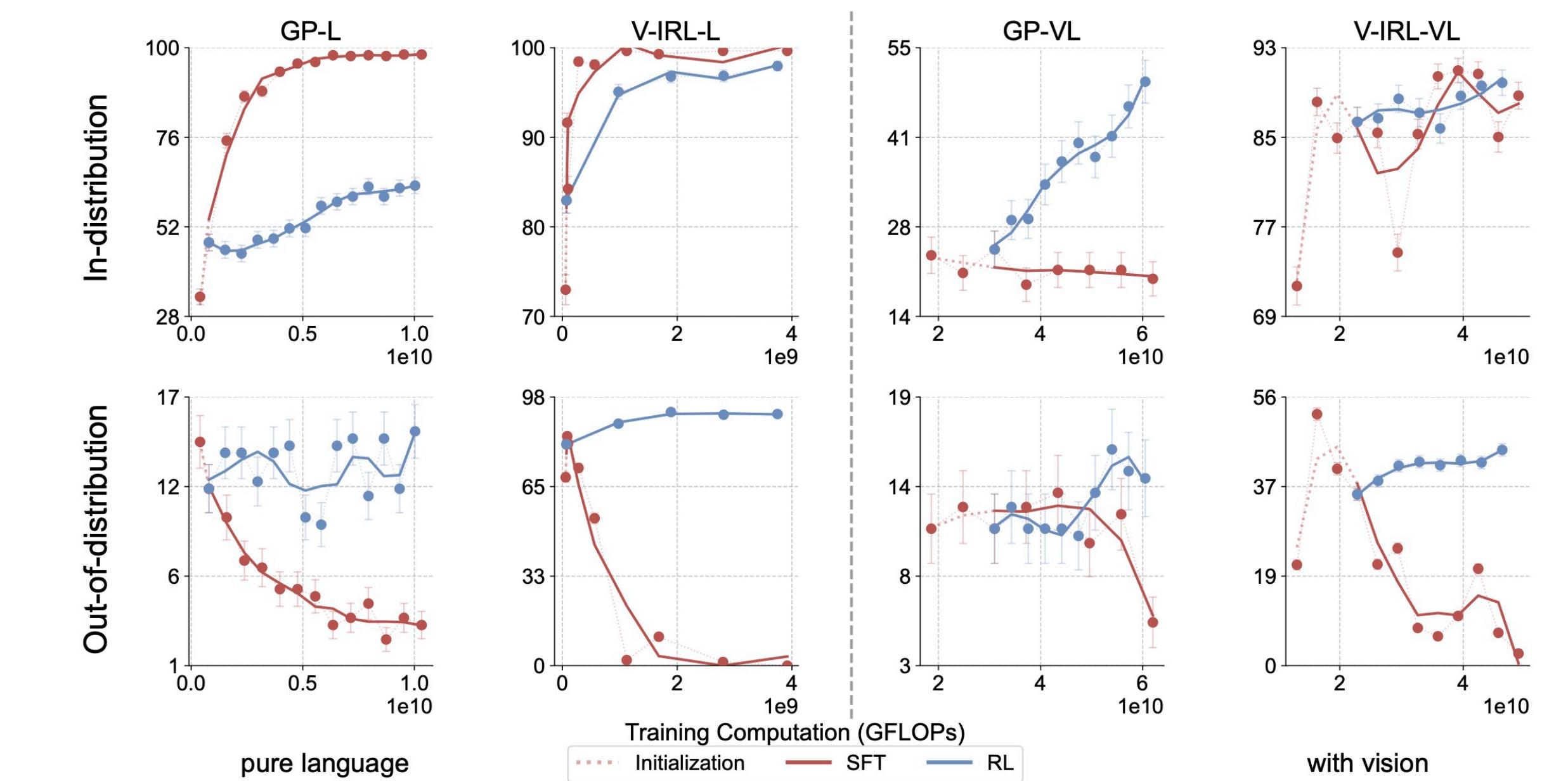
要做出比上海科技大学食堂更难吃的饭菜，需要将烹饪技巧反向运用到极致——核心原则是破坏食材本质、颠覆味觉逻辑、挑战人类食欲极限。以下为科学难吃配方（仅供娱乐，请勿浪费食物）：

难吃核心技术指南

1. 食材毁灭预处理
 - 蔬菜：绿叶菜水煮45分钟至发灰，根茎类微波炉干烤至脱水皱缩
 - 肉类：冷冻肉直接下锅（无需解冻），中火慢煎30分钟锁死柴韧口感
 - 主食：米饭加水1:3熬成粥糊状，放凉后冷藏24小时产生「老化回生」颗粒感

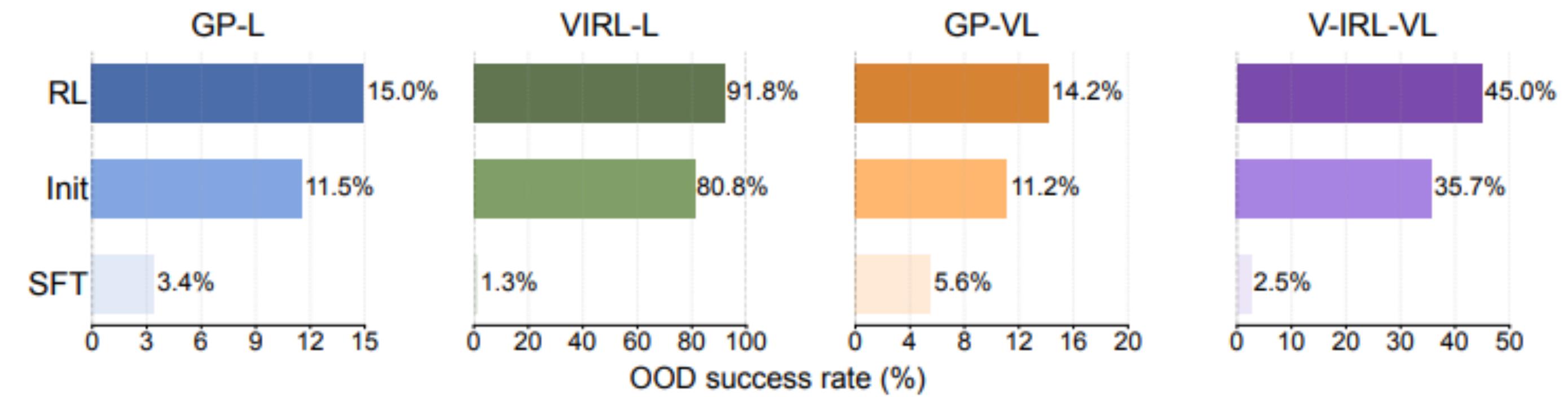
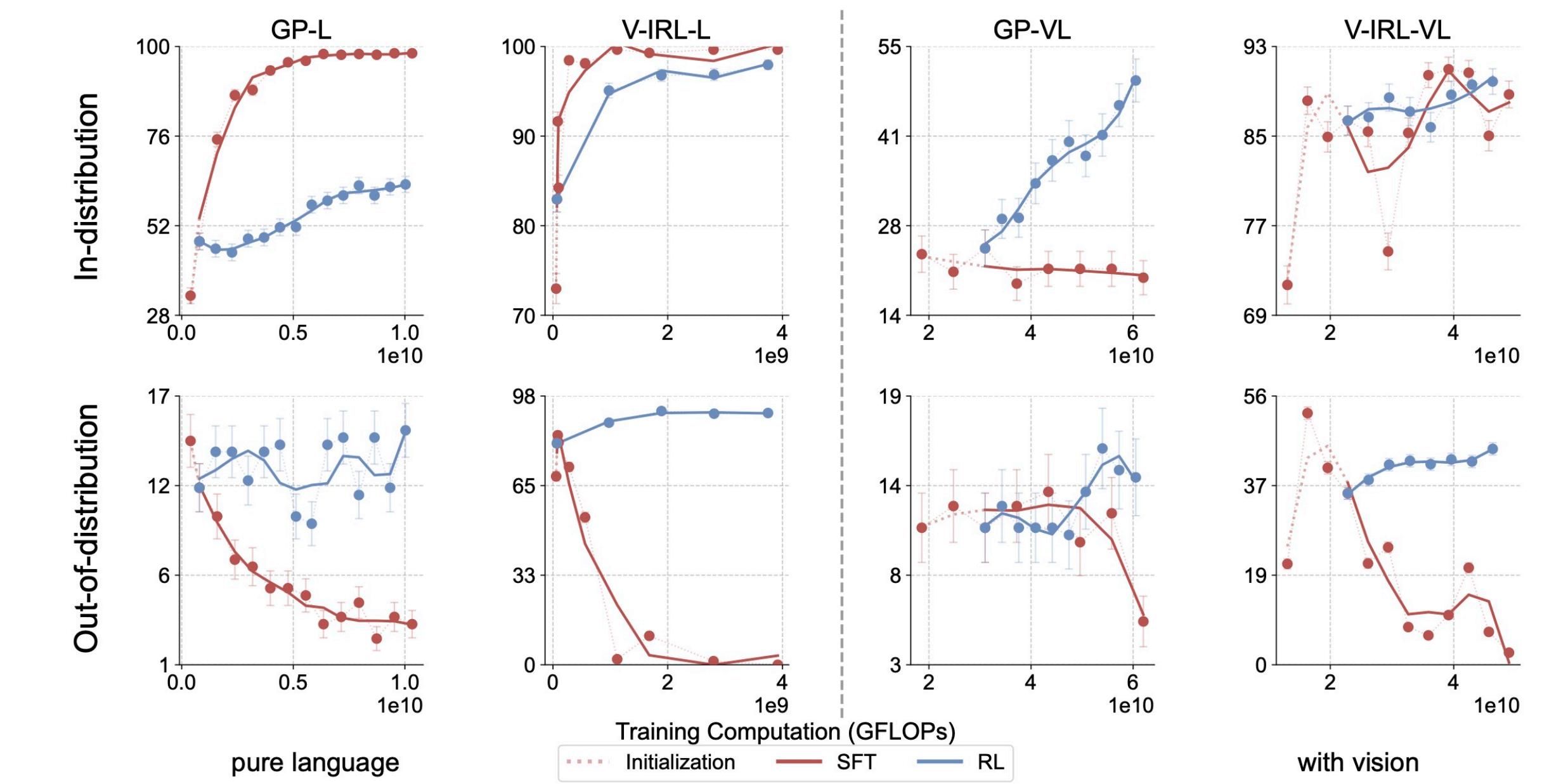
RL Generalizes

- Take away:
 - RL generalizes better than SFT



RL Generalizes

- Take away:
 - RL generalizes better than SFT
 - Equal test time compute
 - Equal training compute



Delving Deeper...

- ▶ RL tunes subnetwork [Mukherjee et al., 2025]
- ▶ Training data difficulty matters in RL [Zeng et al., 2025]
- ▶ ...

On the other hand

- ▶ RL on one example [Wang et al., 2025]
- ▶ RL without external rewards [Zhao et al., 2025]
- ▶ RL on random rewards [Shao et al., 2025]

On the other hand

- ▶ RL on one example [Wang et al., 2025]
- ▶ RL without external rewards [Zhao et al., 2025]
- ▶ RL on random rewards [Shao et al., 2025]
- ▶ ...

Qwen too strong?

AIME. too easy?

Format matters?

Test time scaling matters?

Which Part(s) make Reasoning Models Work?

⚠ Practical Warning to Future Works

Qwen models, with open weights and high performance on reasoning tasks, have become the de facto choice for RLVR research in the open-source community—a range of recent research on RLVR drew conclusions on Qwen-centric experiments (check our paper for a list).

However, we find that two recent works that show RLVR with weak supervision work well on Qwen models can not be generalized to other model families.

1. Test-Time Reinforcement Learning ([TTRL](#), Zuo et al., 2025): The paper proposes to conduct RLVR on test samples using on-policy majority-voted answers to compute rewards.
2. 1-shot Reinforcement Learning ([1-shot RL](#), Wang et al., 2025): This paper shows that RLVR with one example can achieve comparable performance as RLVR on a standard training set.

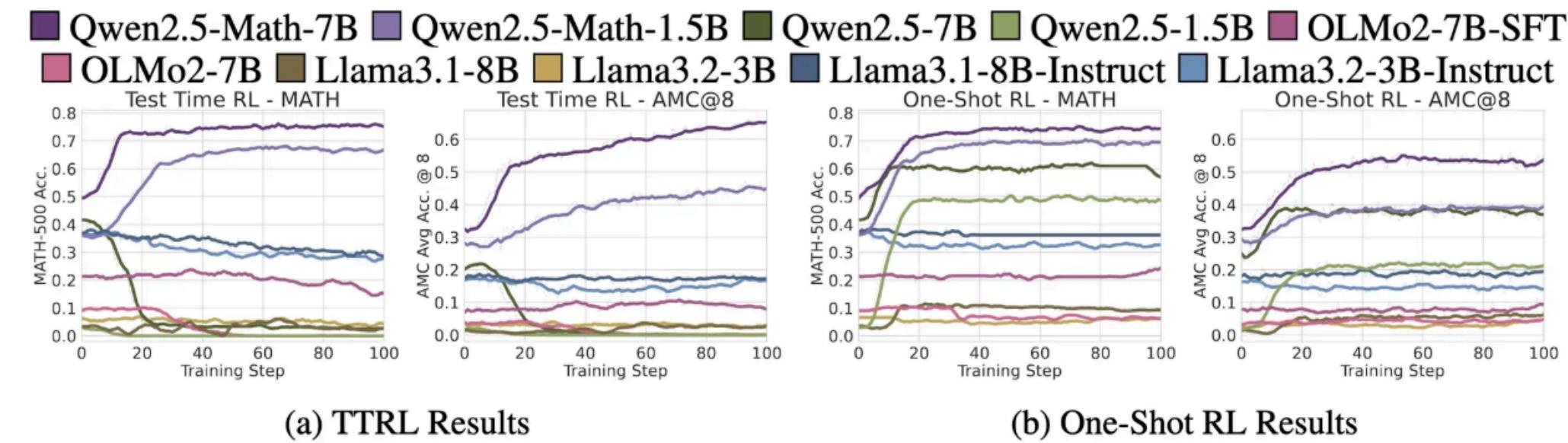


Figure 4: We evaluate two recent weak supervision RL methods—TTRL (Zuo et al., 2025) and One-Shot RL (Wang et al., 2025b)—on diverse base models. We find that the proposed training rewards can consistently work on Qwen models. Yet with few exceptions, those same proposed signals often yield no gains on other model families, mirroring the limited generalization observed when training with our own spurious rewards. See Appendix A.6 for setup details.

Therefore, we suggest that future RLVR research should be validated on other models.

Which Part(s) make Reasoning Models Work?

- ▶ Preparing base model for RL

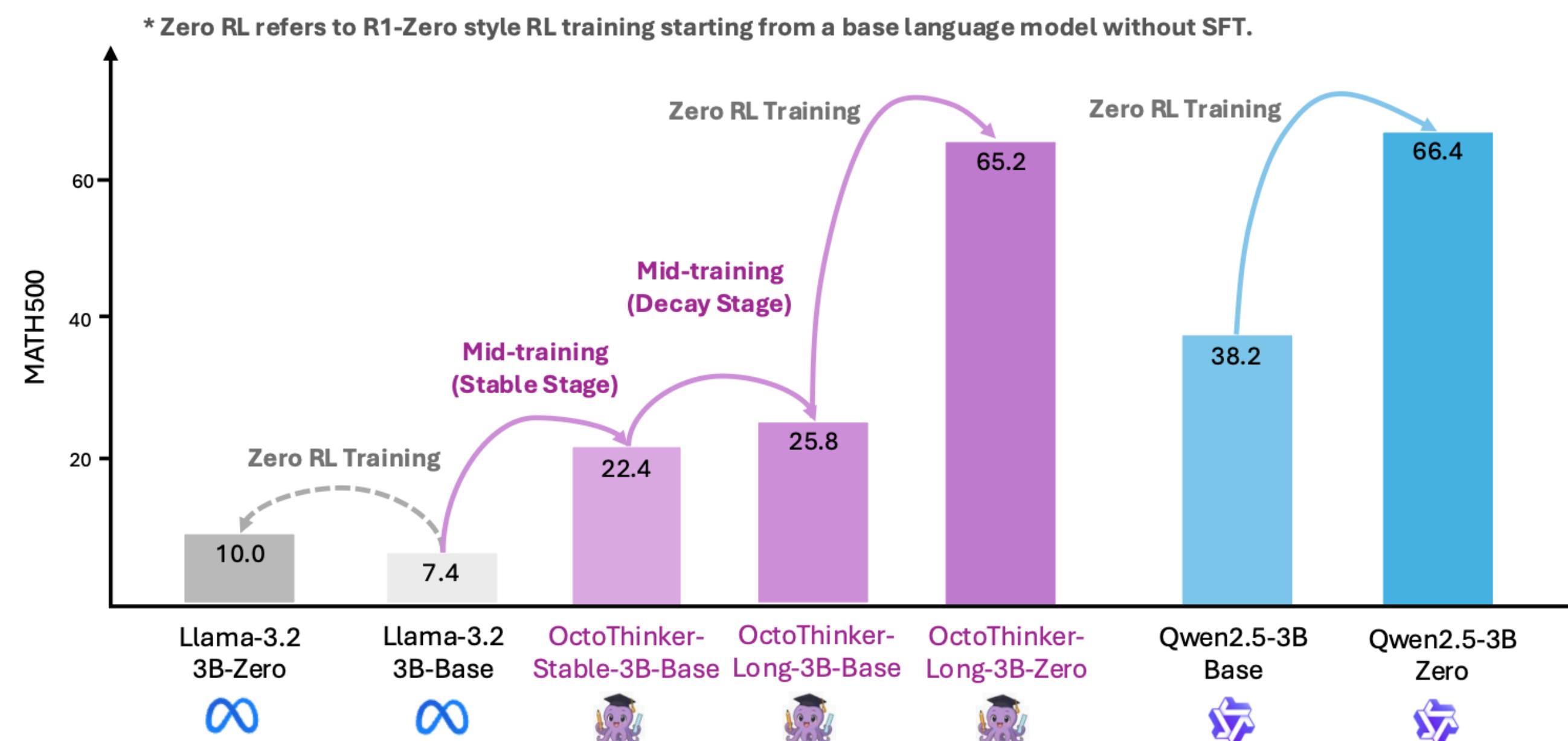


Figure 1 | Our strategic mid-training incentivizes Llama's RL scaling, matching Qwen2.5 performance.

Which Part(s) make Reasoning Models Work?

- ▶ Preparing base model for RL

What is Mid-training?

Mid-training is a mid-stage whose computational and data (token) requirements are intermediate between pre-training and post-training. It aims to achieve specific objectives—such as domain and language expansion (Dou et al., 2025, *inter alia*), long-context extension (Abdin et al., 2024a,b, *inter alia*), improving data quality (Hu et al., 2024a; OLMo et al., 2025, *inter alia*), leveraging large-scale synthetic data (Yang et al., 2024a, 2025, 2024b, *inter alia*), and preparing for post-training, among others—by significantly altering data quality and distribution (Dubey et al., 2024; Wake et al., 2024, *inter alia*) (and/or modifying model architecture to improve inference efficiency (Bercovich et al., 2024, 2025, *inter alia*)).^a

^aIn the absence of a precise or widely agreed-upon definition, here, we aim to introduce a concise and rigorous definition of *mid-training* within this context. The term was reportedly first mentioned in an OpenAI job description in mid-2024. A detailed blog for this term is available at <https://vintagedata.org/blog/posts/what-is-mid-training>

Outliner

- ▶ Overview
- ▶ Basics
- ▶ Which Part(s) make Reasoning Models Work?
- ▶ Takeaways

Takeaways

- ▶ RL fever cools down after 6 months

Takeaways

- ▶ RL fever cools down after 6 months
- ▶ (Principled) Understanding is still lacking

Takeaways

- ▶ RL fever cools down after 6 months
- ▶ (Principled) Understanding is still lacking
- ▶ Should you work on it?

Zeyuan Allen-Zhu, Sc.D. @ZeyuanAllenZhu · Jul 29

∅ ...

❓ Can I collect some feedbacks: Is fully open-source research necessary? Earlier, I released a family of 1-8B models (open data, code, weights): beating Llama3-8B with <10% pretrain time, beating most (all?) open-data models of this scale.

⚠ No shortcuts: 10+ legal debates for using open data, 10+ more for weights, months of blockers, endless nights scavenging GPUs.

↗ All to provide a strong, reproducible baseline — one I believed critical for understanding the physics of LLMs. Next planned was GLA + Canon, outperforming all modern linear models in tests.

Yet attention was low.

❓ Should I close-source to save time and focus on pure research? Honest feedback appreciated.

Zeyuan Allen-Zhu, Sc.D. @ZeyuanAllenZhu · Jul 27

Phase 1 of Physics of Language Models code release

- ✓ our Part 3.1 + 4.1 = all you need to pretrain strong 8B base model in 42k GPU-hours
- ✓ Canon layers = strong, scalable gains
- ✓ Real open-source (data/train/weights)...

Note: Huggingface access is restricted from CN/RU/HK/etc. per company policy. Please apply from an eligible location.

"Physics of Language Models" PhysicsLM4.2 Code release – Phase 1

Efficient pretrain for <8B base models 42k H200 GPU-hours for 8B model

1.3B models, our LlamaCanon
1.1B models, our Llama