

大模型高效微调框架 LLaMA Factory 配合 vLLM 的最佳实践

郑耀威

LLaMA Factory 团队负责人 | 北京航空航天大学

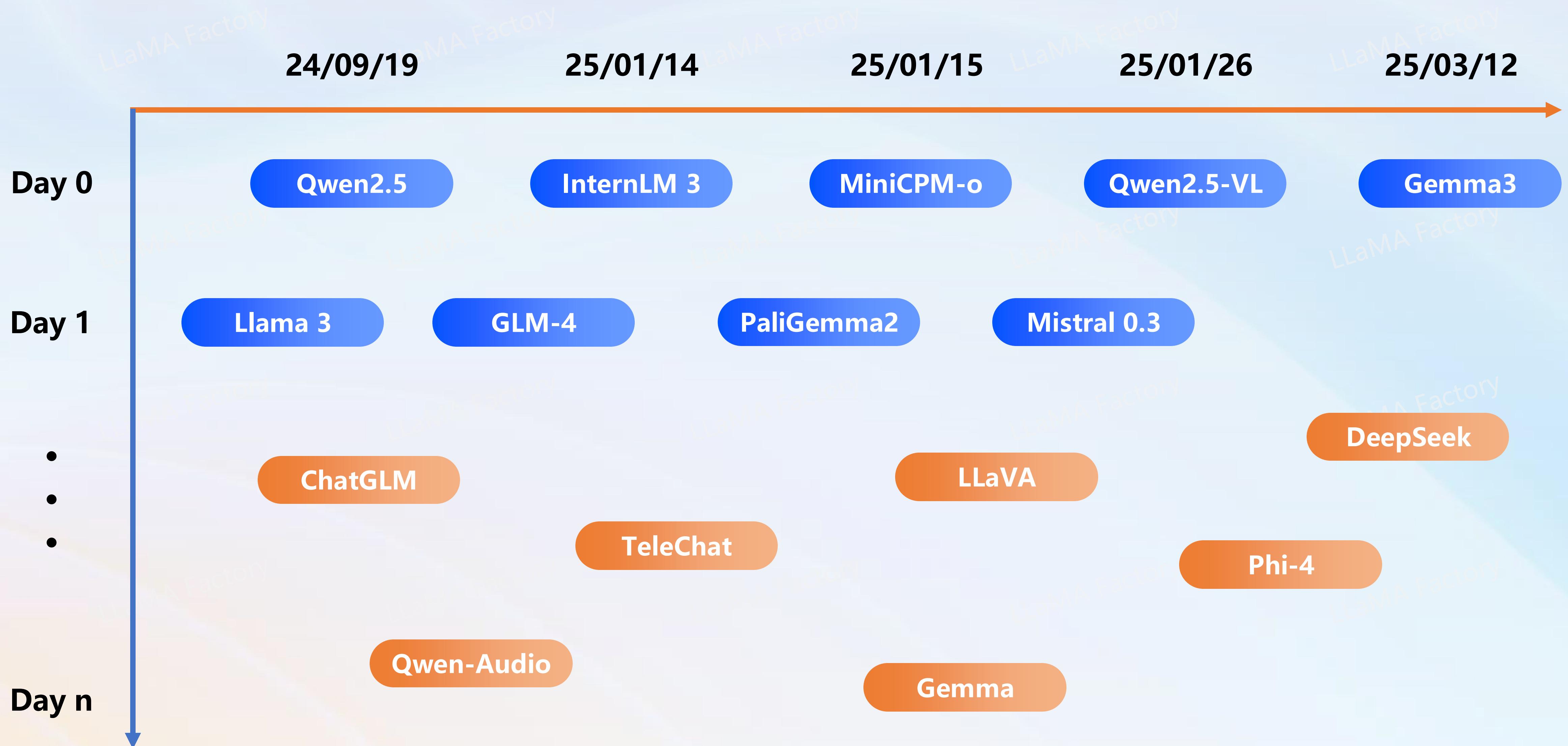
2025/03/16

使用 LLaMA-Factory 零代码微调百种大模型

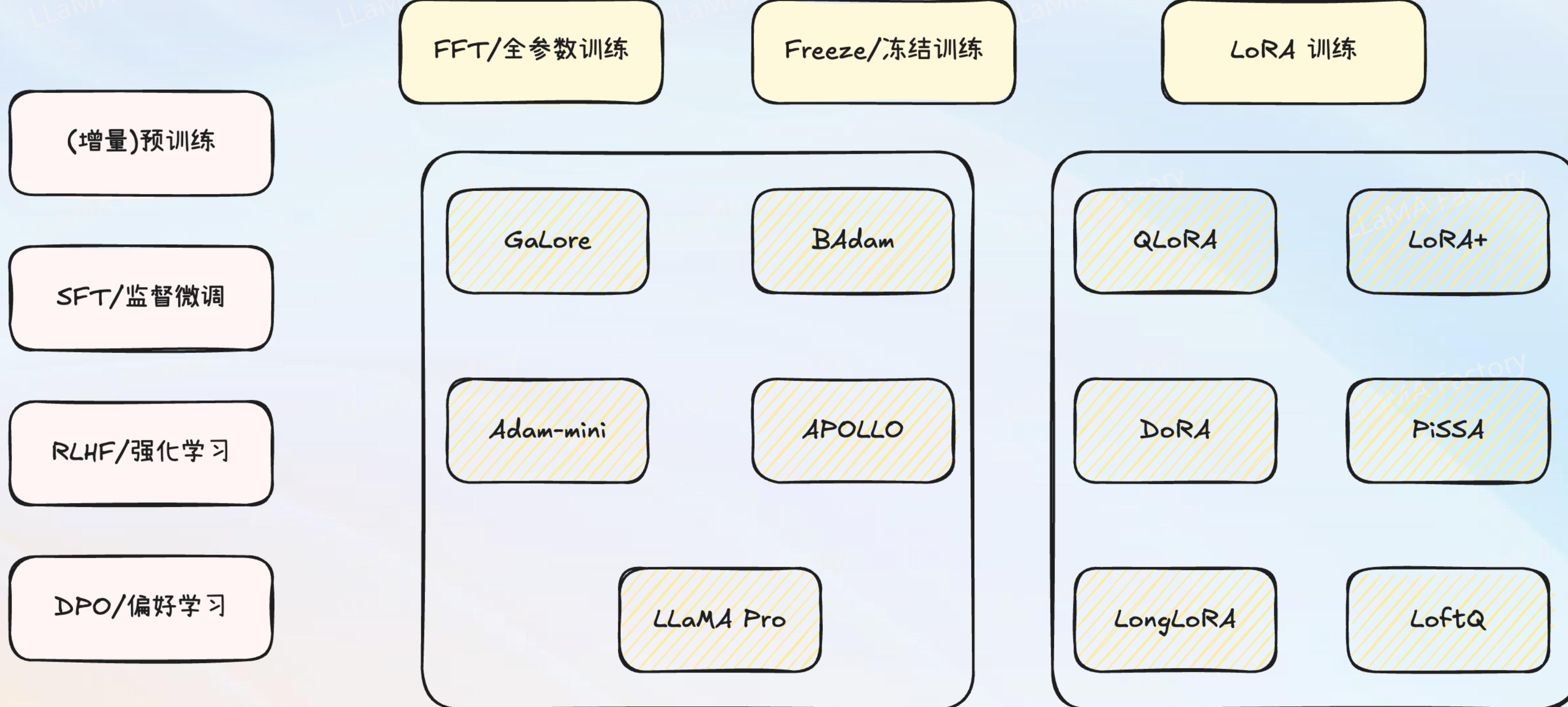
The screenshot displays the LLaMA-Factory web application interface for training large language models. The interface is organized into several sections:

- Language:** A dropdown menu set to "en".
- Model name:** An input field for searching a model, currently empty.
- Model path:** A field for specifying a pretrained model or identifier from Hugging Face, also empty.
- Finetuning method:** A dropdown menu set to "lora".
- Checkpoint path:** An input field for the checkpoint path, currently empty.
- Quantization bit:** A dropdown menu set to "none".
- Quantization method:** A dropdown menu set to "bitsandbytes".
- Chat template:** A dropdown menu set to "default".
- RoPE scaling:** A dropdown menu set to "none".
- Booster:** A dropdown menu set to "auto".
- Train:** The active tab, highlighted in orange.
- Evaluate & Predict**
- Chat**
- Export**
- Stage:** A dropdown menu set to "Supervised Fine-Tuning".
- Data dir:** An input field set to "data".
- Dataset:** A dropdown menu.
- Preview dataset:** A button to preview the dataset.
- Learning rate:** An input field set to "5e-5".
- Epochs:** An input field set to "3.0".
- Maximum gradient norm:** An input field set to "1.0".
- Max samples:** An input field set to "100000".
- Compute type:** A dropdown menu set to "bf16".
- Cutoff length:** A slider set to 2048.
- Batch size:** A slider set to 2.
- Gradient accumulation:** A slider set to 8.
- Val size:** A slider set to 0.
- LR scheduler:** A dropdown menu set to "cosine".
- Extra configurations:** A section with a dropdown arrow.
- Freeze tuning configurations:** A section with a dropdown arrow.
- LoRA configurations:** A section with a dropdown arrow.
- RLHF configurations:** A section with a dropdown arrow.

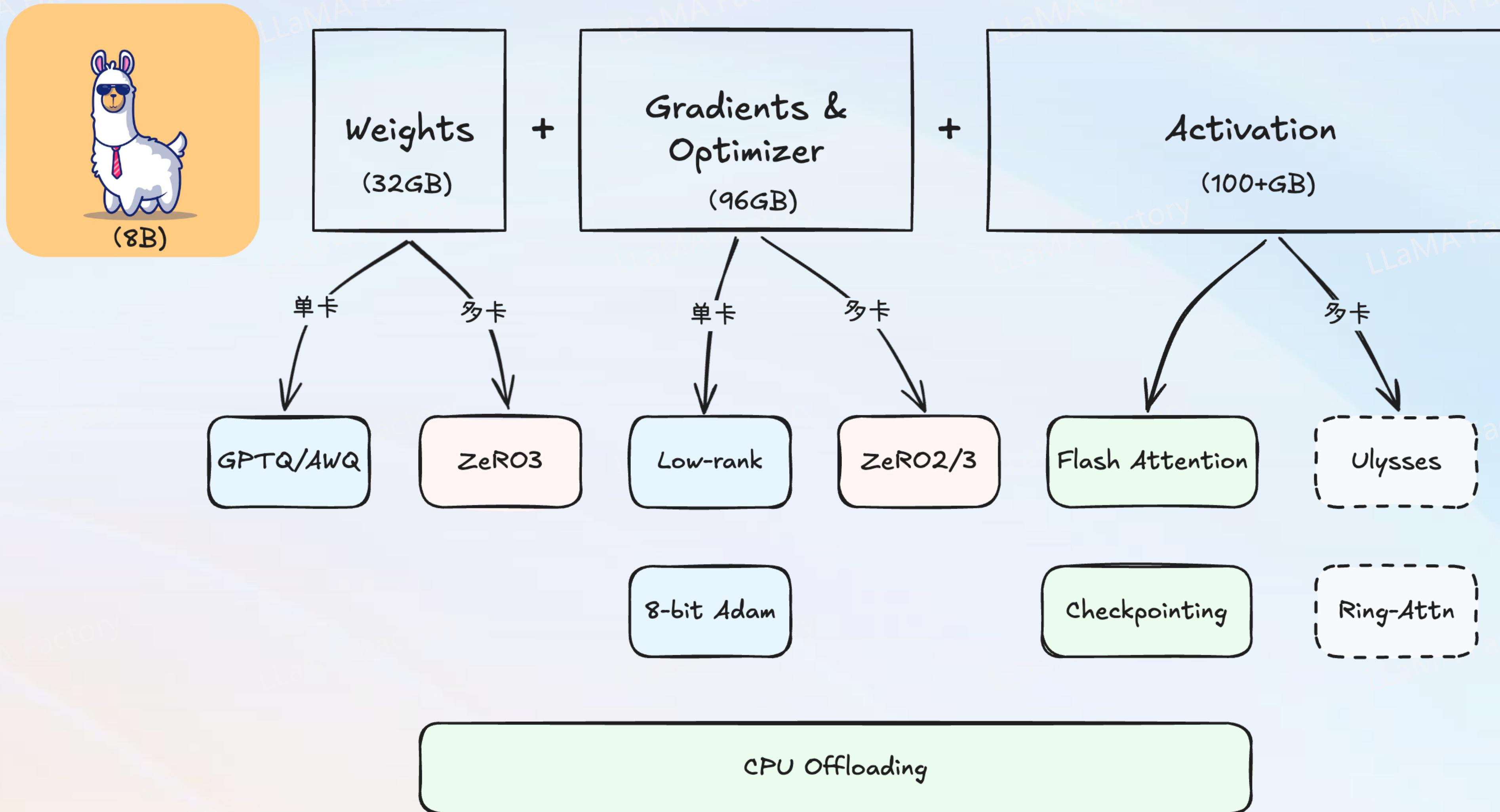
新模型架构 Day0 适配



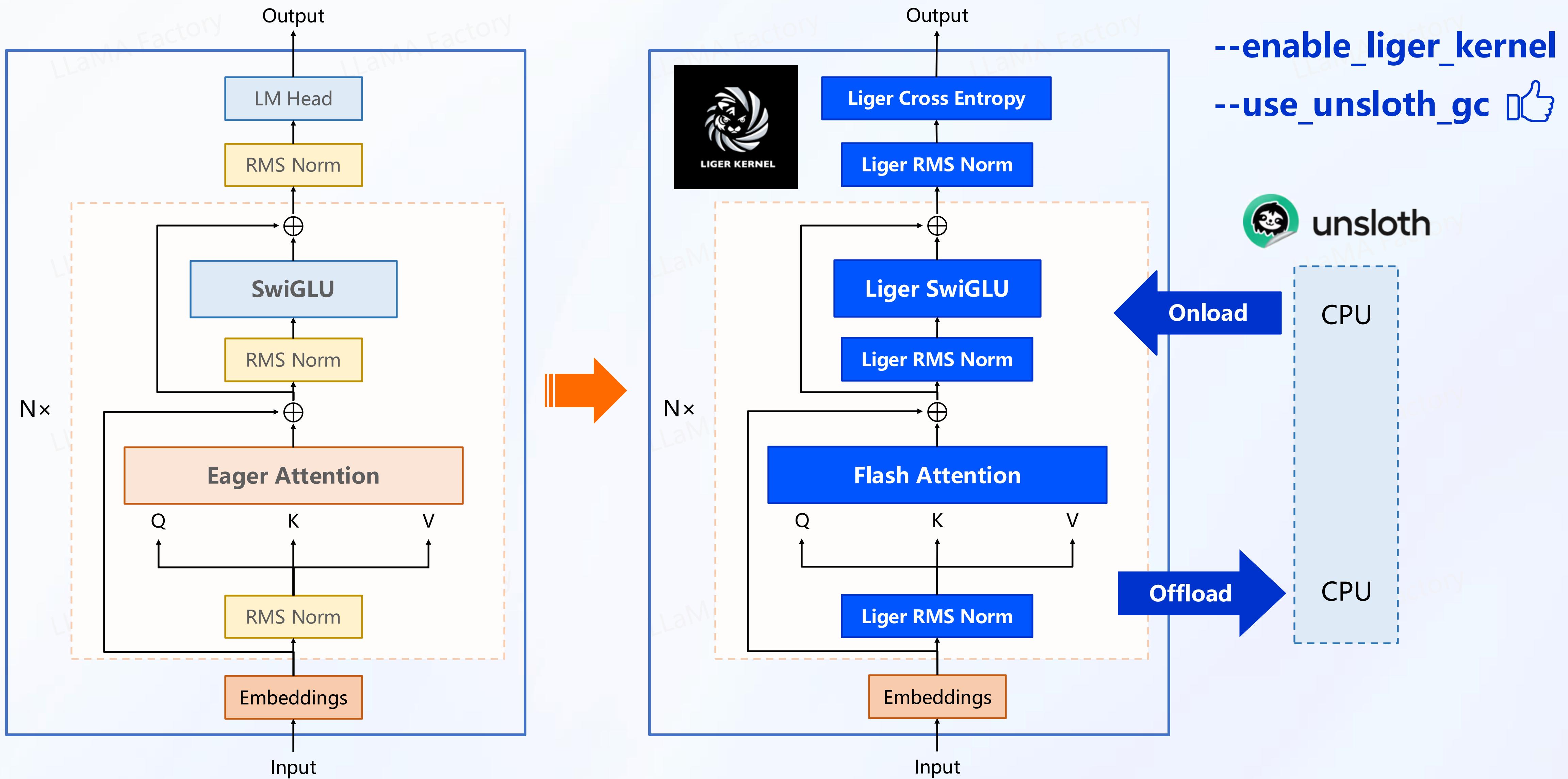
微调算法全覆盖



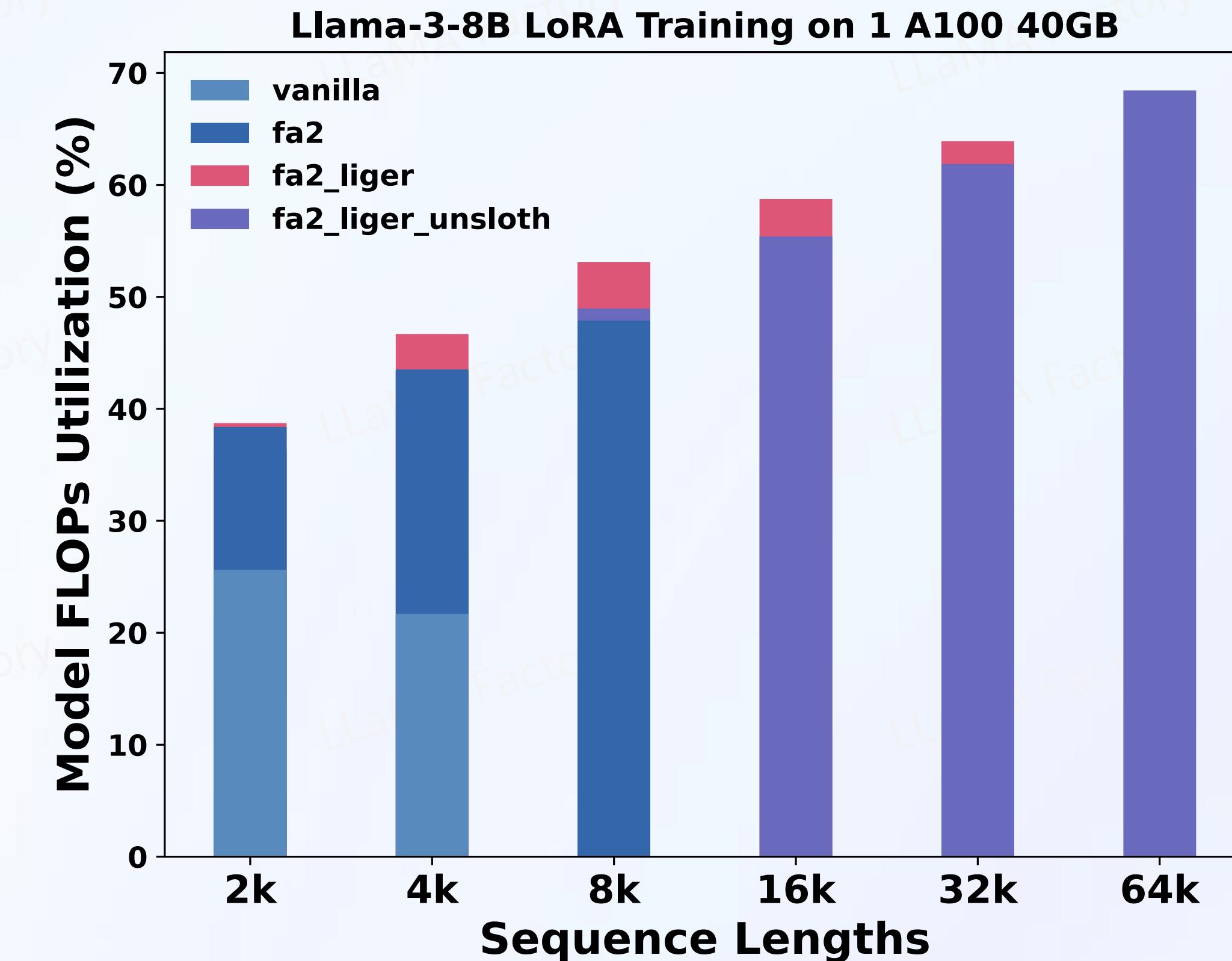
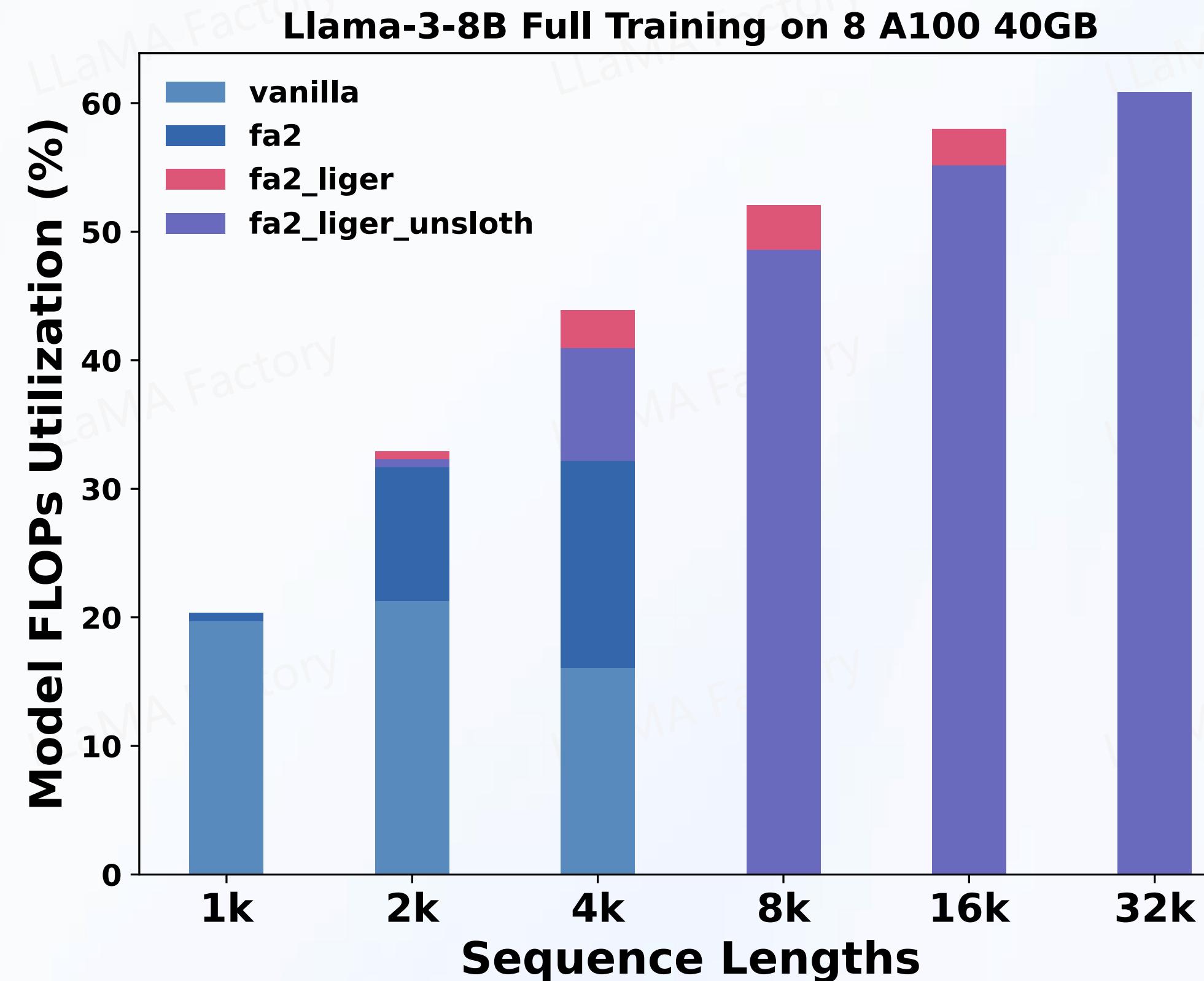
多场景显存管理



Transformer 算子加速

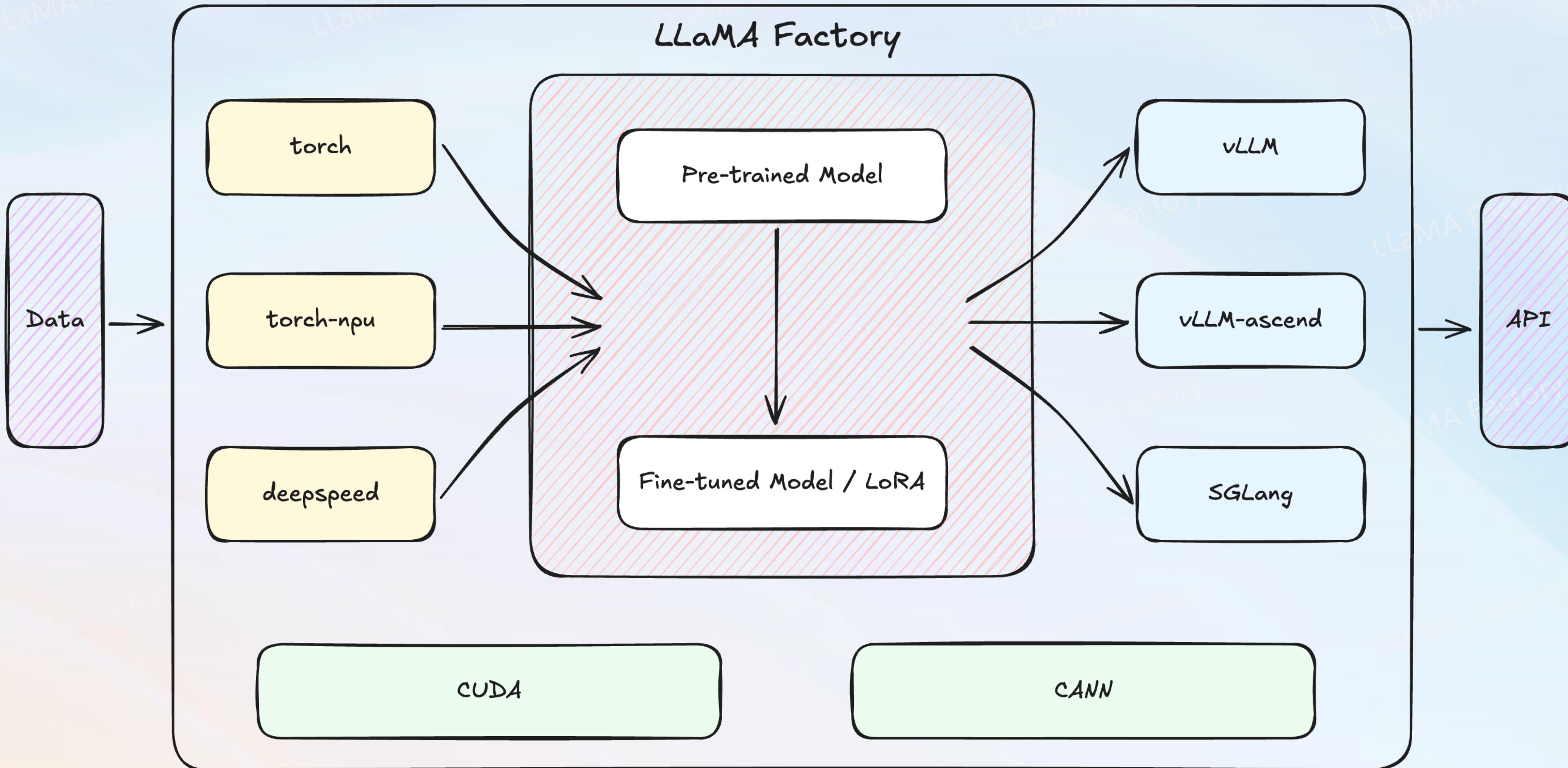


长文本训练性能



- Llama 3 8B 全参微调样本最大长度： **4k -> 32k**
- Llama 3 8B LoRA 微调样本最大长度： **8k -> 64k**

训练推理一站式服务



使用 vLLM 加速 DeepSeek R1 推理

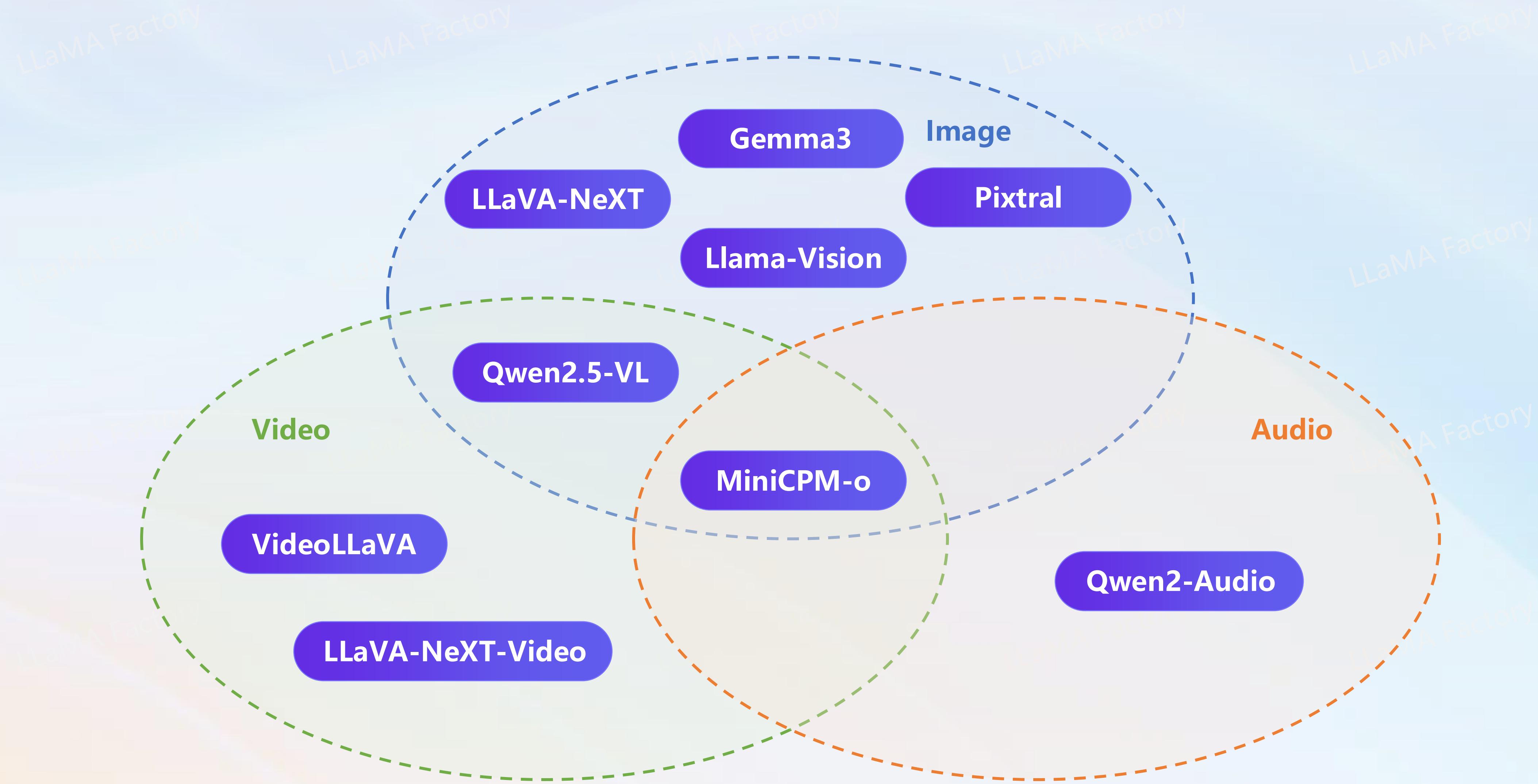
The screenshot shows the vLLM web interface with the following configuration settings:

- 语言:** zh
- 模型名称:** (Input field)
- 模型路径:** (Input field)
- 微调方法:** lora
- 检查点路径:** (Input field)
- 量化等级:** none
- 量化方法:** bitsandbytes
- 对话模板:** default
- RoPE 插值方法:** none
- 加速方式:** auto
- 推理引擎:** huggingface
- 推理数据类型:** auto

At the bottom, there are two large buttons: "加载模型" (Load Model) and "卸载模型" (Unload Model). A message box indicates: "模型未加载, 请先加载模型。" (Model not loaded, please load the model first.)

https://v.youku.com/v_show/id_XNjQ2OTYyMzl4MA==.html

多模态混合理解



多模态混合推理

The screenshot displays a user interface for managing and selecting large language models (LLMs). The interface includes the following sections:

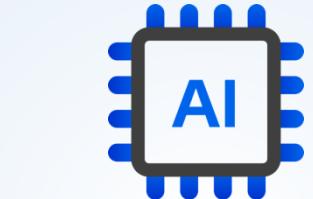
- 语言 (Language):** A dropdown menu set to "zh".
- 模型名称 (Model Name):** A search bar with placeholder text "输入首单词以检索模型。" (Input the first word to search for the model.)
- 微调方法 (Fine-tuning Method):** A dropdown menu set to "lora".
- 量化等级 (Quantization Level):** A dropdown menu set to "none".
- 推理引擎 (Inference Engine):** A dropdown menu set to "huggingface".
- 模型路径 (Model Path):** A text input field for specifying the local file path or Hugging Face model identifier.
- RoPE 插值方法 (RoPE Interpolation Method):** A dropdown menu set to "none".
- 加速方式 (Acceleration Method):** A dropdown menu set to "auto".

Below these settings, there is a large list of available LLM models:

- Aya-23-8B-Chat
- Aya-23-35B-Chat
- Baichuan-7B-Base
- Baichuan-13B-Base
- Baichuan-13B-Chat
- Baichuan2-7B-Base
- Baichuan2-13B-Base
- Baichuan2-7B-Chat
- Baichuan2-13B-Chat
- BLOOM-560M
- BLOOM-3B
- BLOOM-7B1
- BLOOMZ-560M
- BLOOMZ-3B
- BLOOMZ-7B1-mt
- BlueLM-7B-Base
- BlueLM-7B-Chat
- Breeze-7B
- Breeze-7B-Instruct
- ChatGLM2-6B-Chat

A prominent button labeled "卸载模型" (Unload Model) is located at the bottom right of the configuration area.

开源社区 & 外部应用



昇腾平台
兼容性认证



RTX AI Toolkit
工具集成



Online Merging
Optimizers



GitHub 星标
44000+



阿里云 PAI
云上训练



Tilearn Angel
训练加速案例



代码写入
Transformers

350+

论文引用

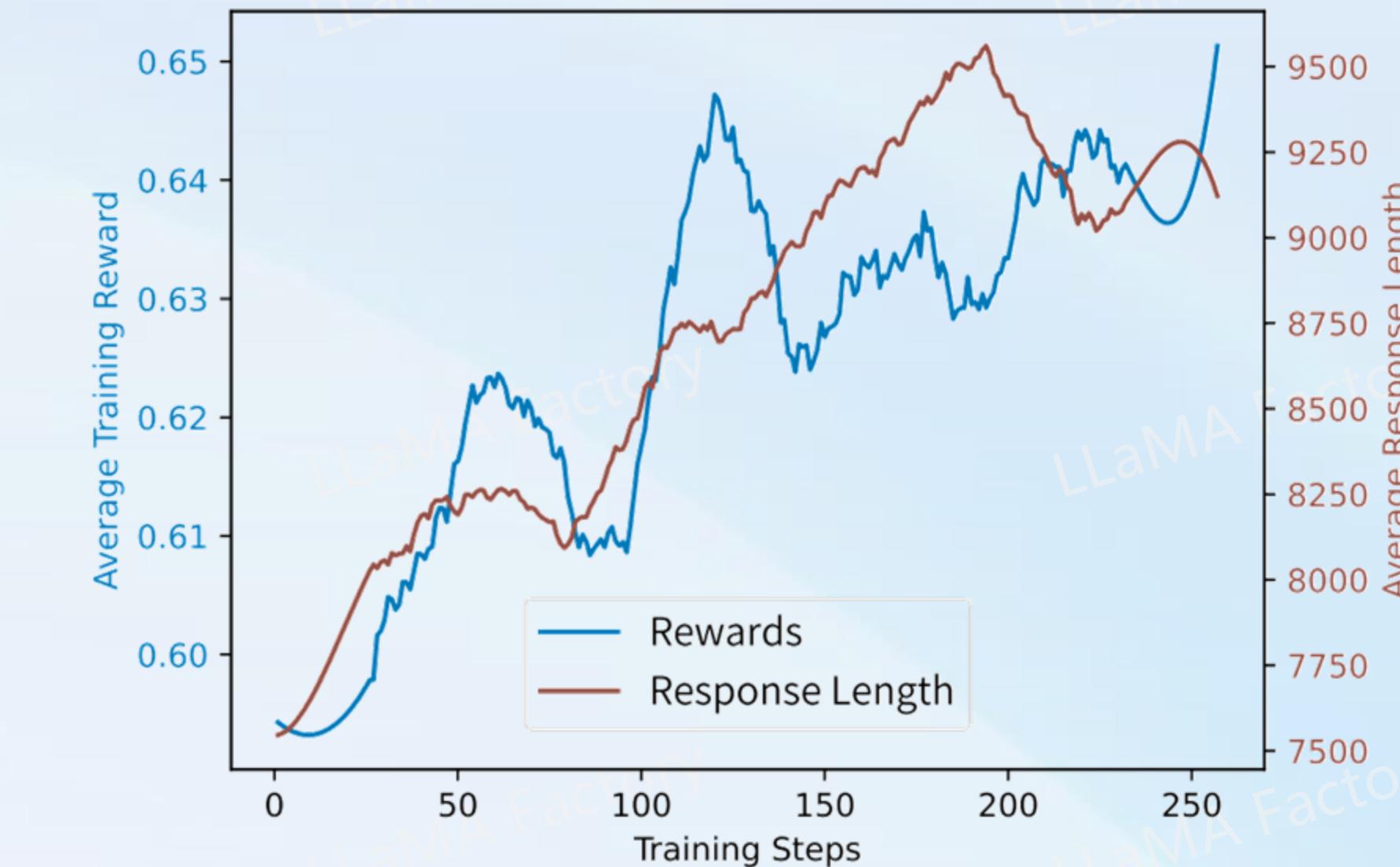
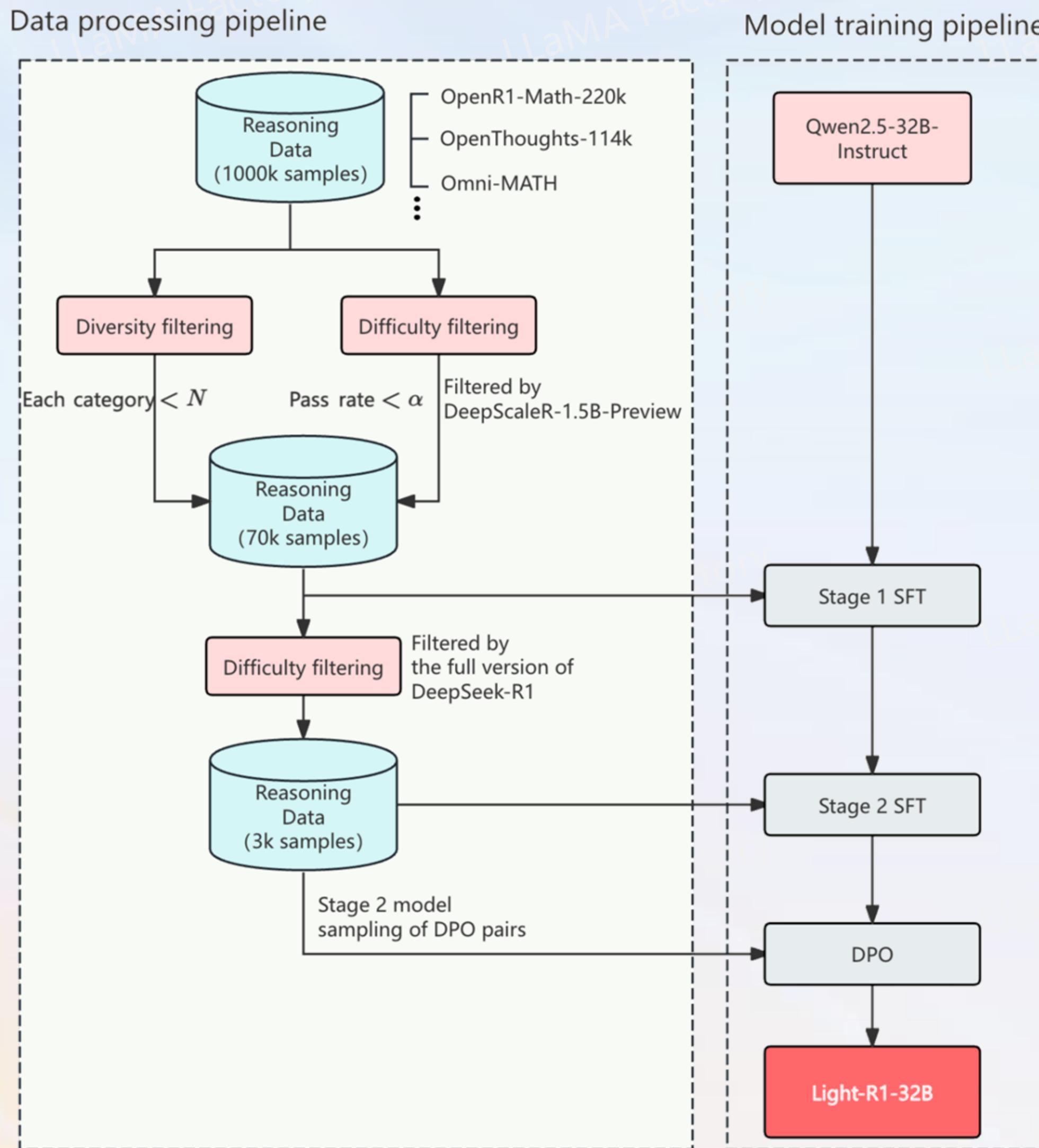


AWS Samples
案例使用

5000+
社区模型

150+
贡献者

Light-R1: Curriculum SFT, DPO and RL



Stage	AIME24	AIME25	GPQA Diamond
Qwen2.5-32B-Instruct (base model)	16.6	13.6	48.8
Light-R1-32B-SFT-stage1	69.0	57.4	64.3
Light-R1-32B-SFT-stage2	73.0	64.3	60.6
Light-R1-32B-DPO	75.8	63.4	61.8
Light-R1-32B (merged model)	76.6	64.6	61.8

DeepSeek-R1-Distill-Qwen-32B 72.6 54.9 62.1

GitHub: <https://github.com/Qihoo360/Light-R1>

Code: <https://github.com/Qihoo360/360-LLaMA-Factory>



What's NeXT?

基于 vLLM/闭源 API 的数据蒸馏工具

vllm serve Qwen/Qwen2.5-72B-Instruct -tp 8

```
2025-03-16 04:46:16,706 - logger - INFO - -----Questions-----  
2025-03-16 04:46:16,706 - logger - INFO - ['哪一天进行女子-78公斤级柔道比赛? ','女子-78公斤级柔道比赛的世界排名是多少的中国选手马振昭将出战? ','马振昭在女子-78公斤级柔道比赛中将尽力达成什么目标? ']  
2025-03-16 04:46:16,706 - logger - INFO - -----Answers-----  
2025-03-16 04:46:16,706 - logger - INFO - ['8月1日','世界排名第4','力争闯入决赛,创造佳绩']
```

Generate Dataset

Dataset Preview

question	answer
哪一天将产生巴黎奥运会的首金?	7月27日
哪几个项目将争夺巴黎奥运会的首金?	10米气步枪混团项目
哪两对选手代表中国队争夺巴黎奥运会的首金?	黄雨婷/盛李豪、韩佳予/杜林澍
昌雅妮和陈艺文计划在哪一天参加巴黎奥运会女子双人3米板比赛?	7月27日
昌雅妮和陈艺文参加的是哪个奥运项目的比赛?	巴黎奥运会女子双人3米板
昌雅妮和陈艺文的比赛将在哪里进行?	水上运动中心
李冰洁在巴黎奥运会女子400米自由泳比赛的具体日期是什么?	7月27日
李冰洁参加的是巴黎奥运会的哪一项目?	女子400米自由泳

Distill-Factory

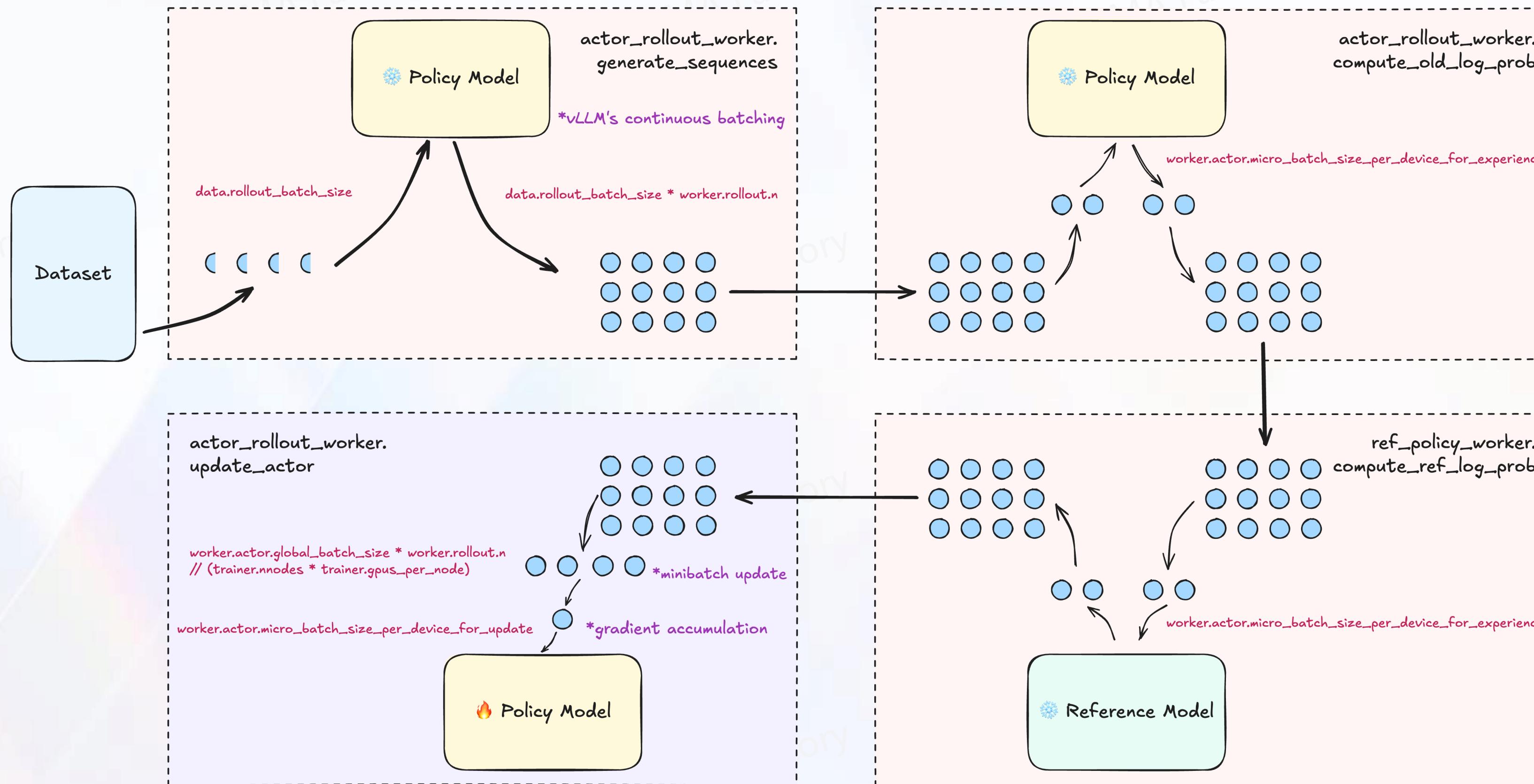
基于 vLLM 的模型评测工具

Command	Measured Acc	Reported Acc
mathruler gen meta-llama/Llama-3.1-8B-Instruct	50.8%	51.9%
mathruler gen Qwen/Qwen2.5-Math-7B-Instruct --temperature 0.5 --sample_num 4	82.6%	83.6%
--json_path data/gsm8k_splits/test.jsonl	90.4%	-
	95.6%	95.2%

```
from mathruler.grader import extract_boxed_content, grade_answer  
  
grade_answer(given_answer: str, ground_truth: str)  
grade_answer(extract_boxed_content(generated_result: str), answer: str)
```

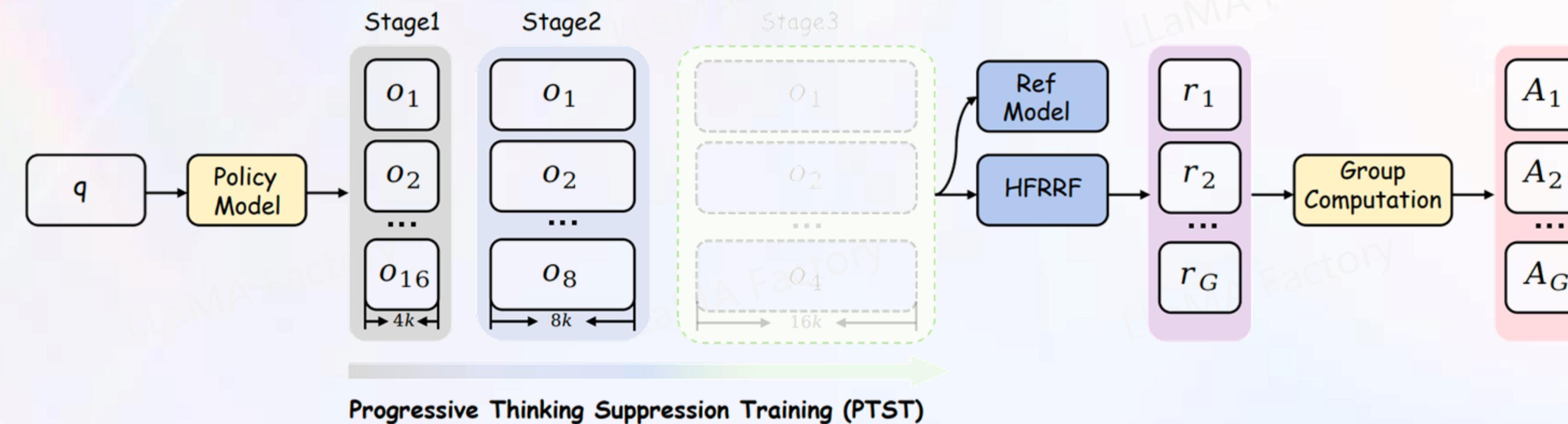
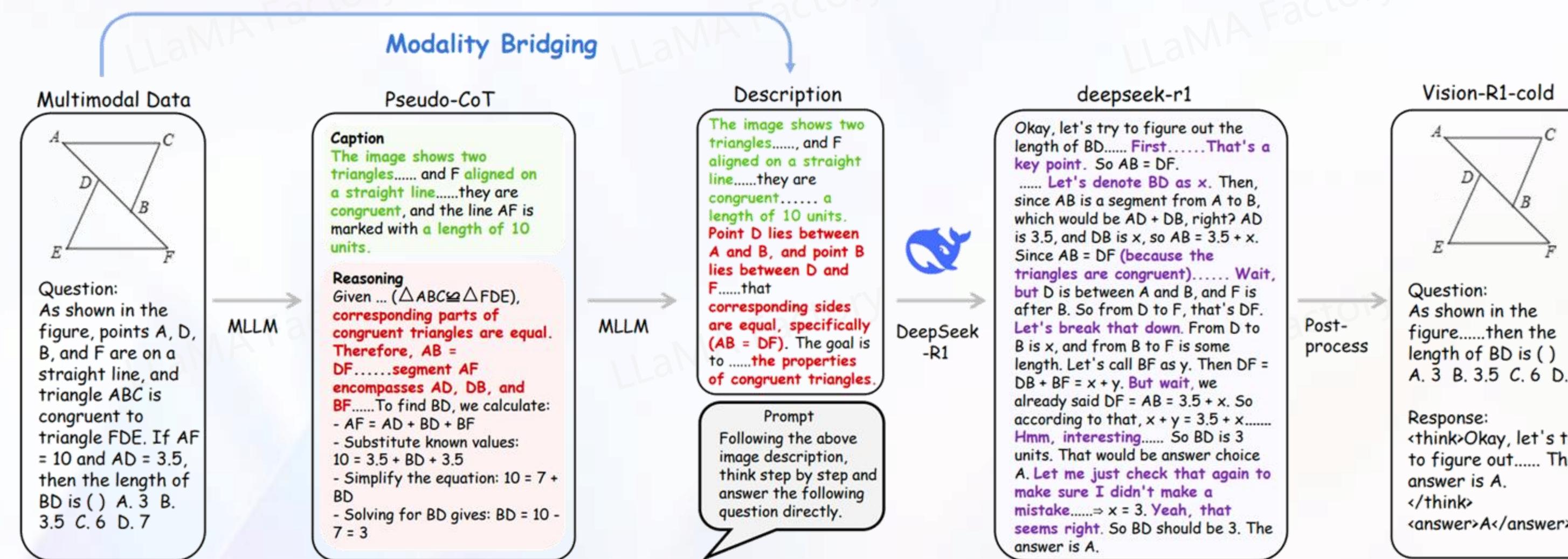
MathRuler

基于 vLLM 和 veRL 的多模态强化学习框架



EasyR1

Vision-R1: Incentivizing Reasoning Capability in MLLMs



GitHub: <https://github.com/Osilly/Vision-R1>

谢谢

Thank You

44k Stars

