

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

Language

en

Model name

Input the initial name to search for the model.

Model path

Path to pretrained model or model identifier from Hugging Face.

Finetuning method

lora

Checkpoint path

Quantization bit

Enable quantization (QLoRA).

none

Quantization method

Quantization algorithm to use.

bitsandbytes

Chat template

The chat template used in constructing prompts.

default

RoPE scaling

RoPE scaling method to use.

none

Booster

Approach used to boost training speed.

auto

Train

Evaluate & Predict

Chat

Export

Stage

The stage to perform in training.

Supervised Fine-Tuning

Data dir

Path to the data directory.

data

Dataset

Preview dataset

Learning rate

Initial learning rate for AdamW.

5e-5

Epochs

Total number of training epochs to perform.

3.0

Maximum gradient norm

Norm for gradient clipping.

1.0

Max samples

Maximum samples per dataset.

100000

Compute type

Whether to use mixed precision training.

bf16

Cutoff length

Max tokens in input sequence.

2048

4

131072

Batch size

Number of samples processed on each GPU.

2

1

1024

Gradient accumulation

Number of steps for gradient accumulation.

8

1

1024

Val size

Percentage of validation set from the entire dataset.

0

0

1

LR scheduler

Name of the learning rate scheduler.

cosine

Extra configurations

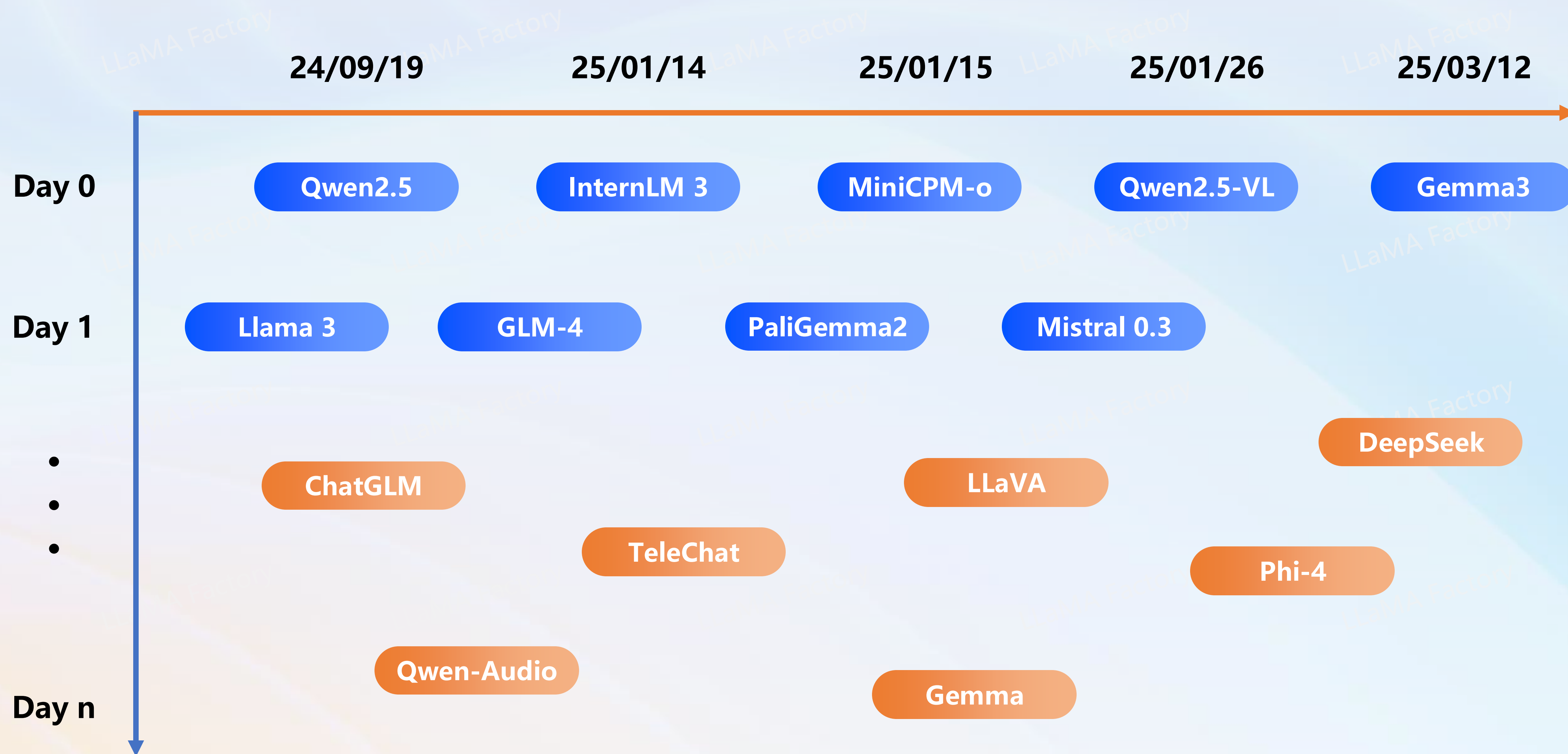
Freeze tuning configurations

LoRA configurations

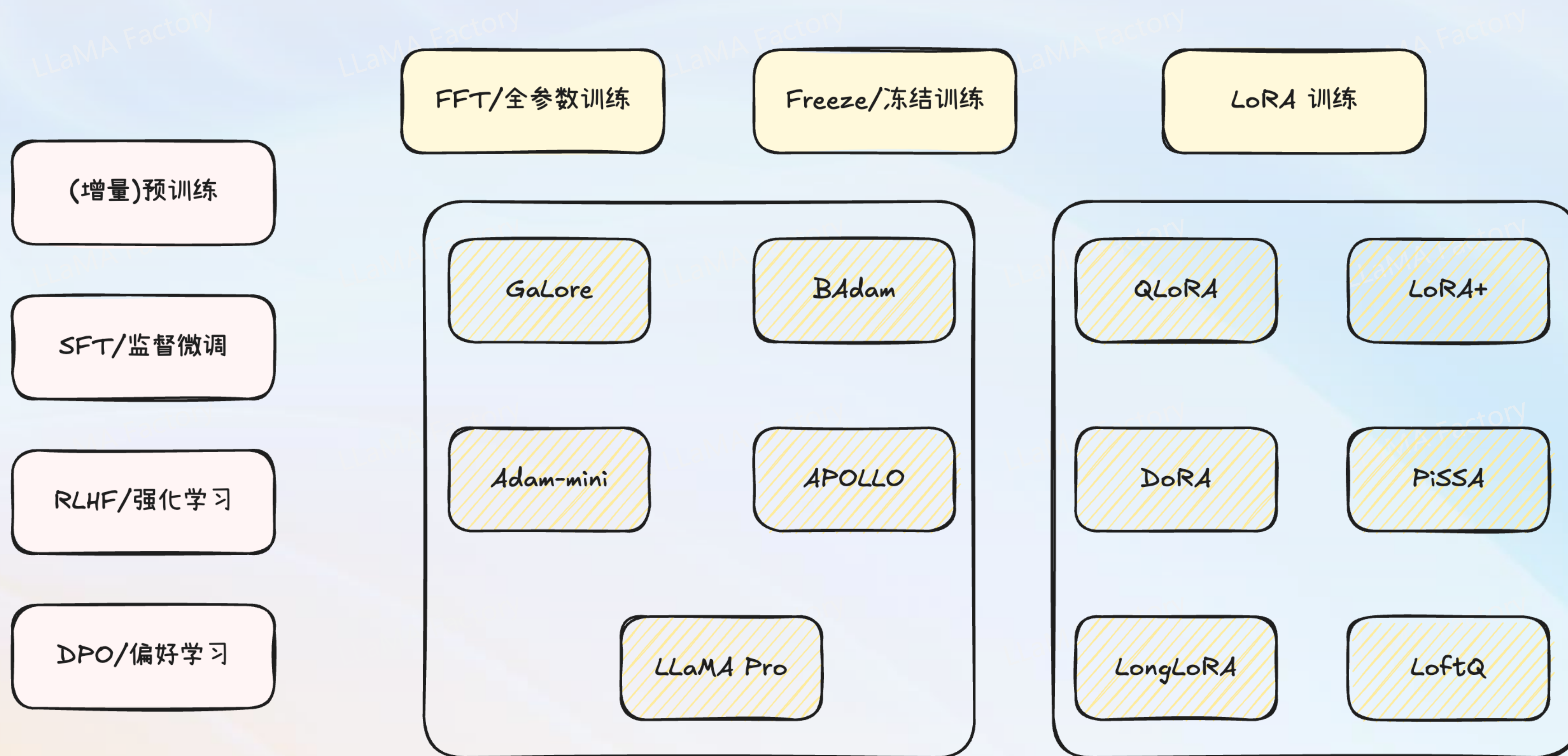
RLHF configurations

https://v.youku.com/v_show/id_XNjQ2OTYxNTg2NA==.html

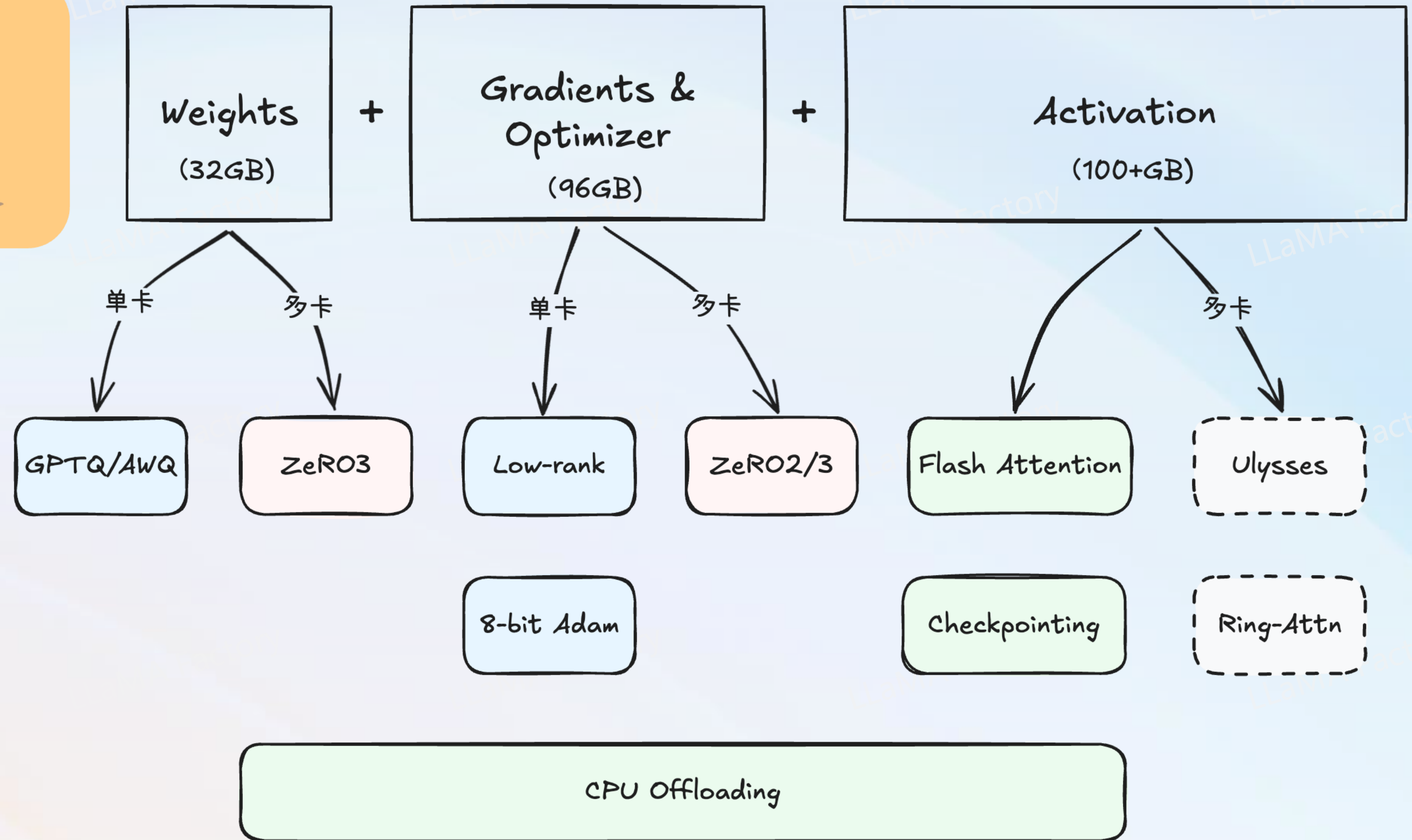
新模型架构 Day0 适配



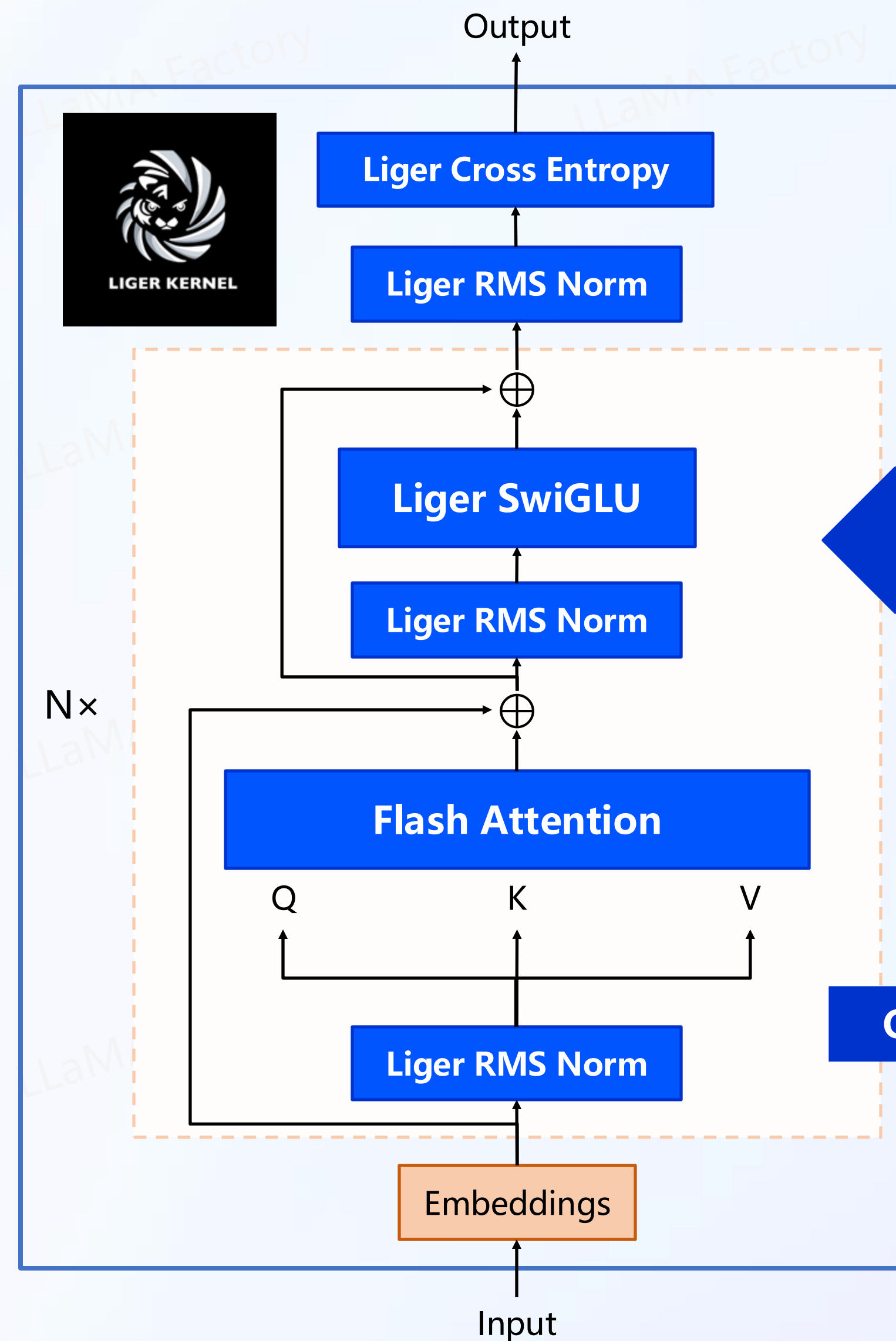
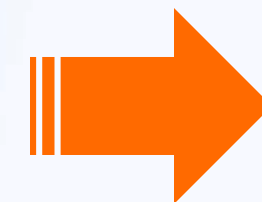
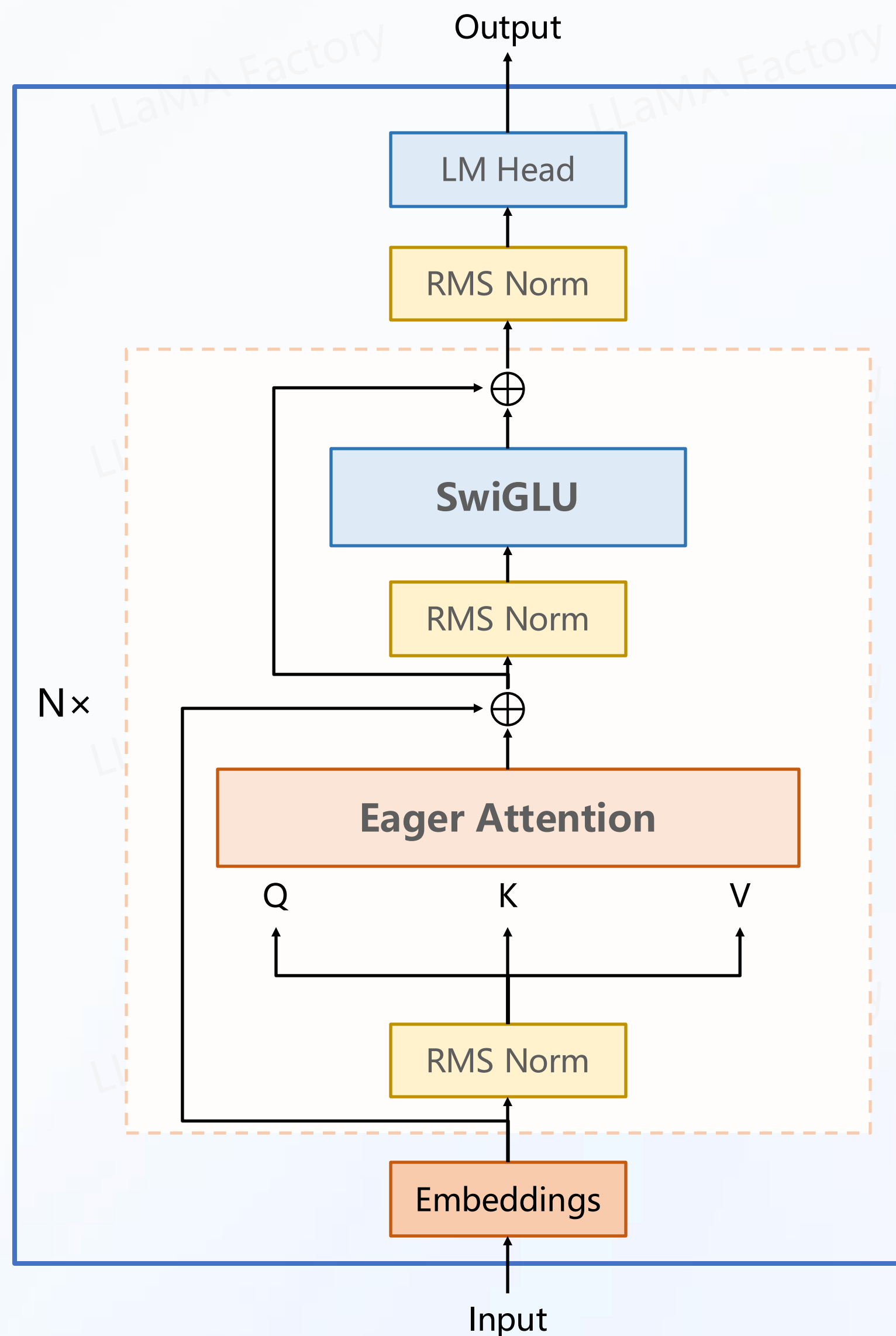
微调算法全覆盖



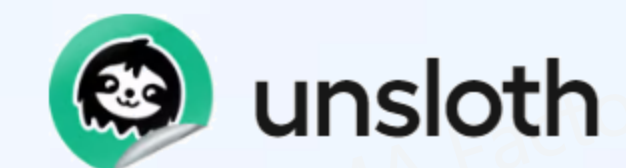
多场景显存管理



Transformer 算子加速



--enable_liger_kernel
--use_unsloth_gc 👍

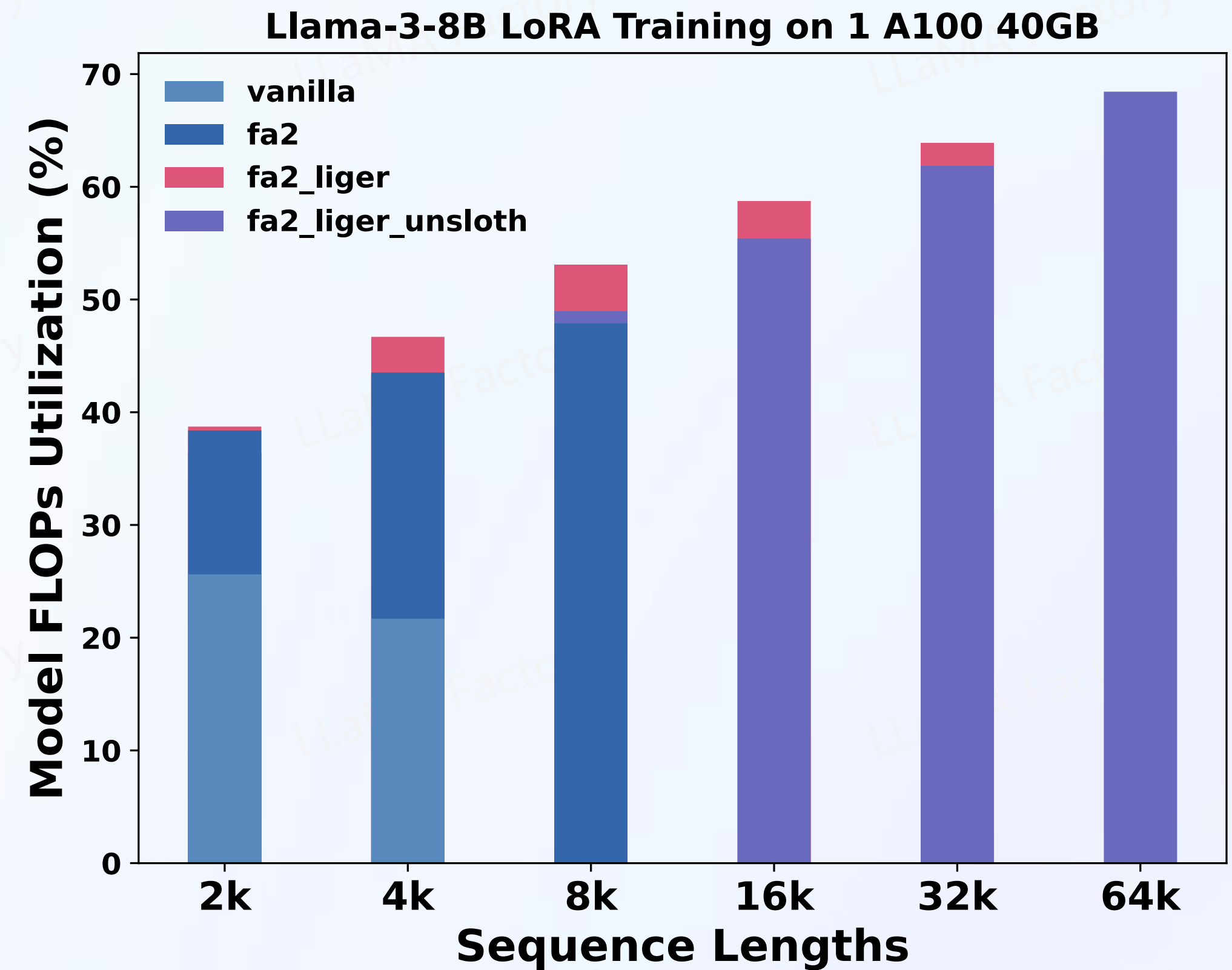
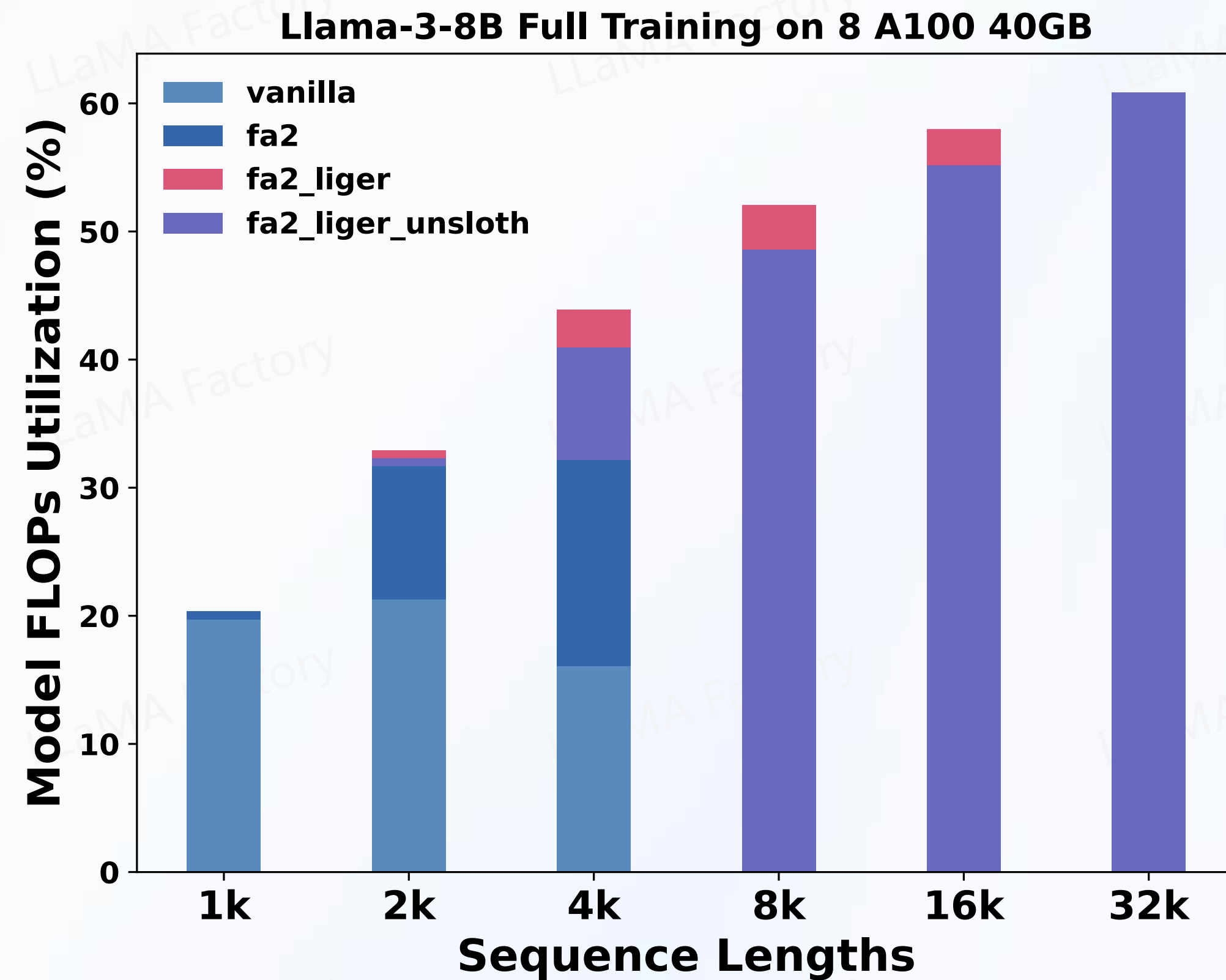


Onload

Offload

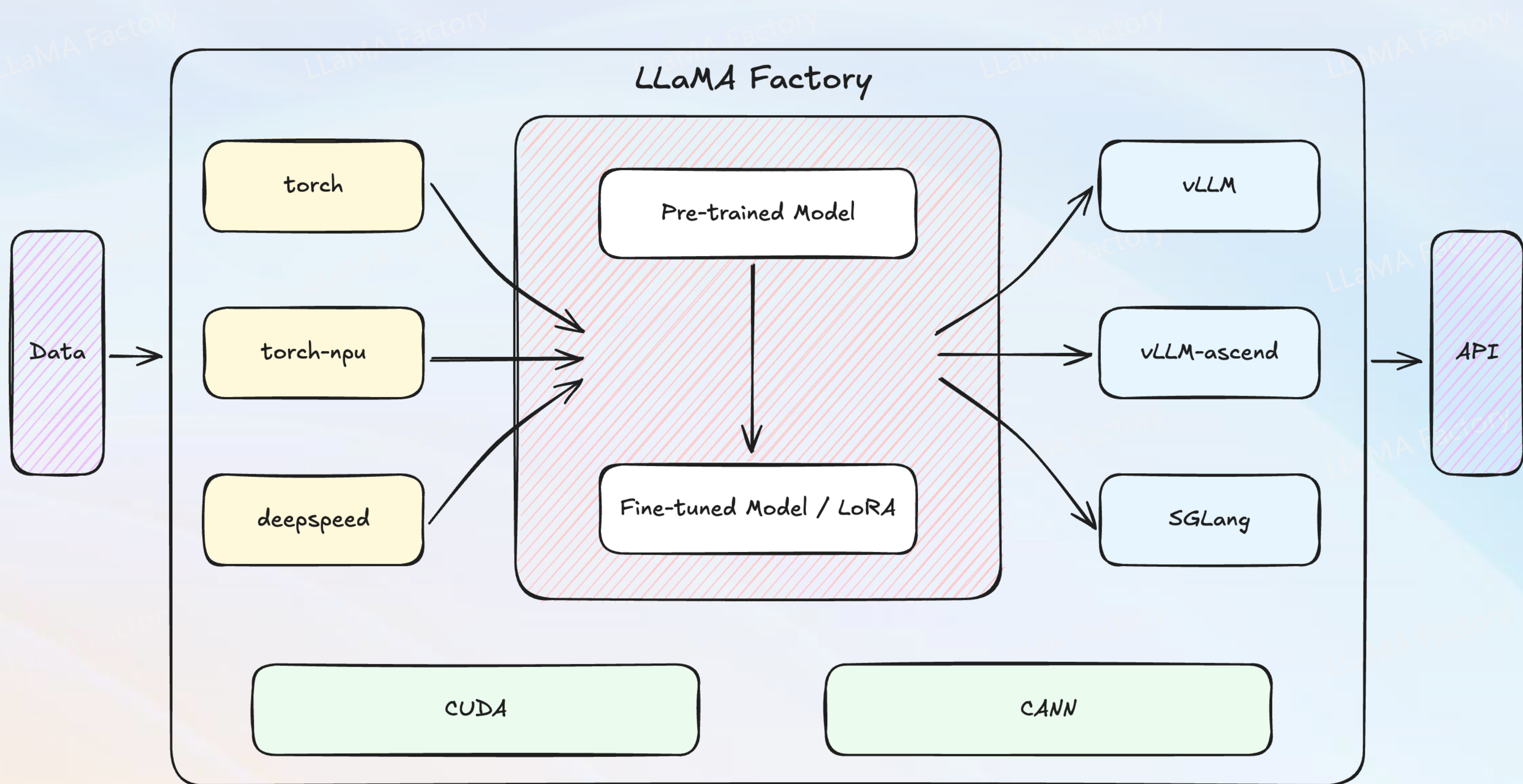


长文本训练性能



- 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

语言

zh

模型名称

输入首单词以检索模型。

模型路径

本地模型的文件路径或 Hugging Face 的模型标识符。

微调方法

lora

检查点路径

量化等级

启用量化 (QLoRA)。

none

量化方法

使用的量化算法。

bitsandbytes

对话模板

构建提示词时使用的模板。

default

RoPE 插值方法

RoPE 插值时使用的方法。

none

加速方式

使用的加速方法。

auto

Train

Evaluate & Predict

Chat

Export

推理引擎

huggingface

推理数据类型

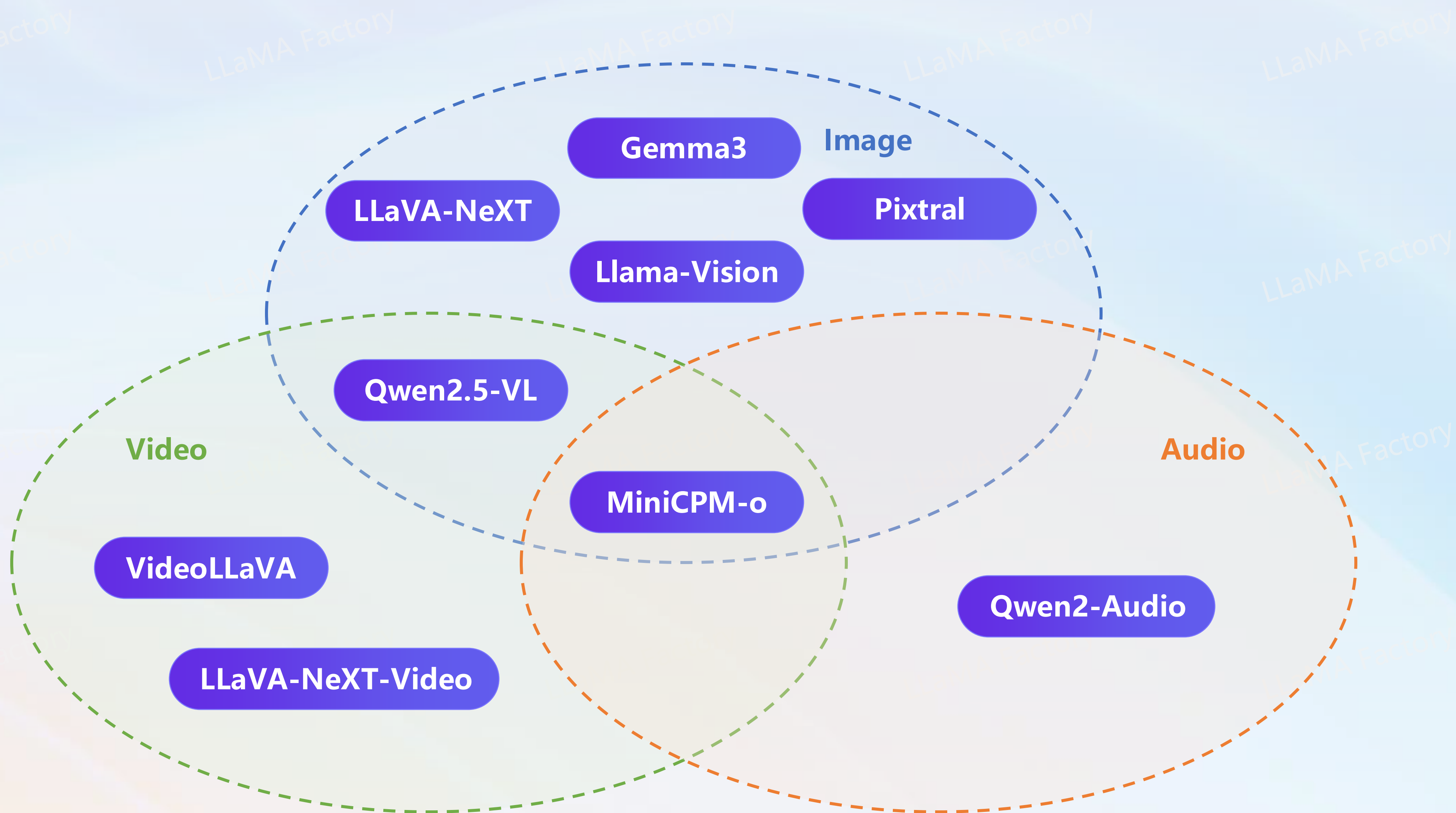
auto

加载模型

卸载模型

模型未加载，请先加载模型。

多模态混合理解



多模态混合推理

语言

zh

微调方法

lora

量化等级

启用量化 (QLoRA)。

none

Train

Evaluate & Predict

Chat

推理引擎

huggingface

模型名称

输入首单词以检索模型。

模型路径

本地模型的文件路径或 Hugging Face 的模型标识符。

RoPE 插值方法

RoPE 插值时使用的方法。

none

加速方式

使用的加速方法。

auto

模型未加载，请先加载模型。

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

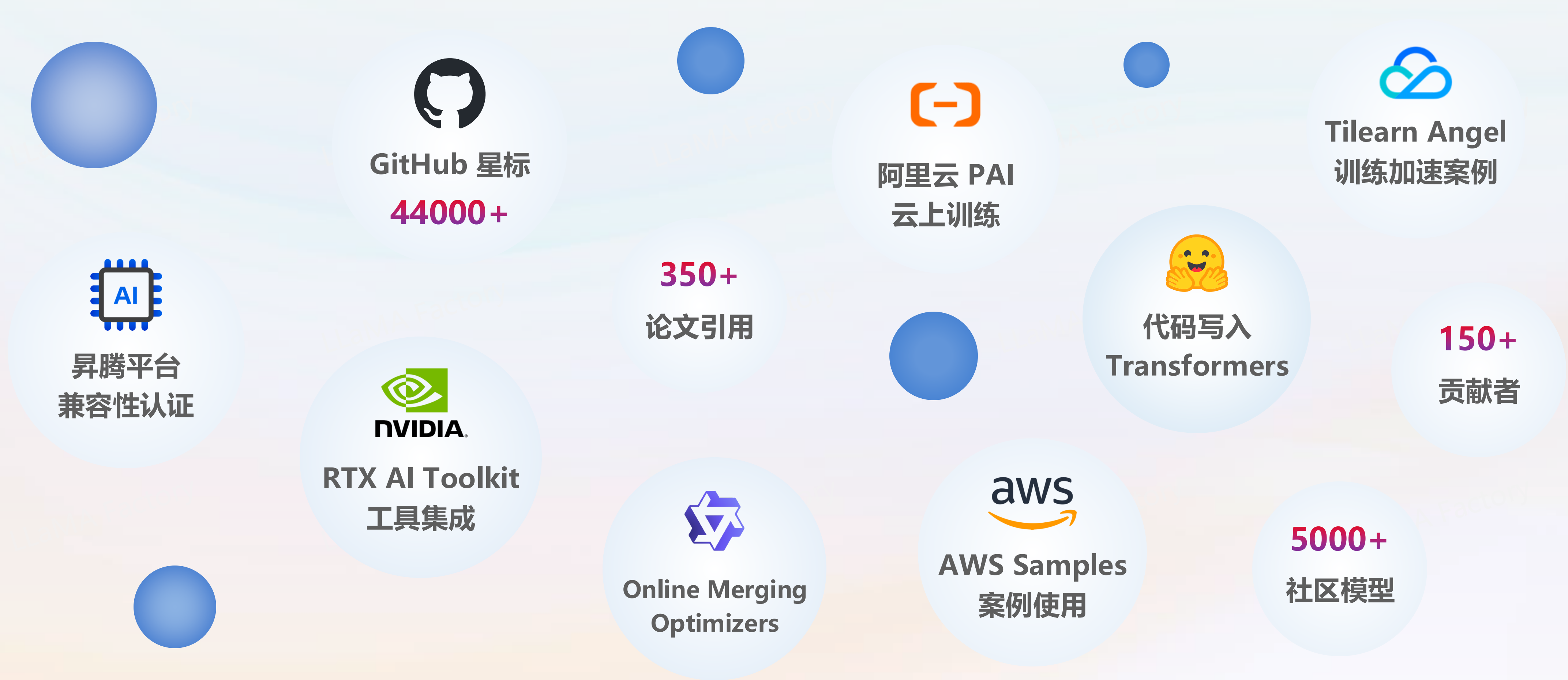
Breeze-7B

Breeze-7B-Instruct

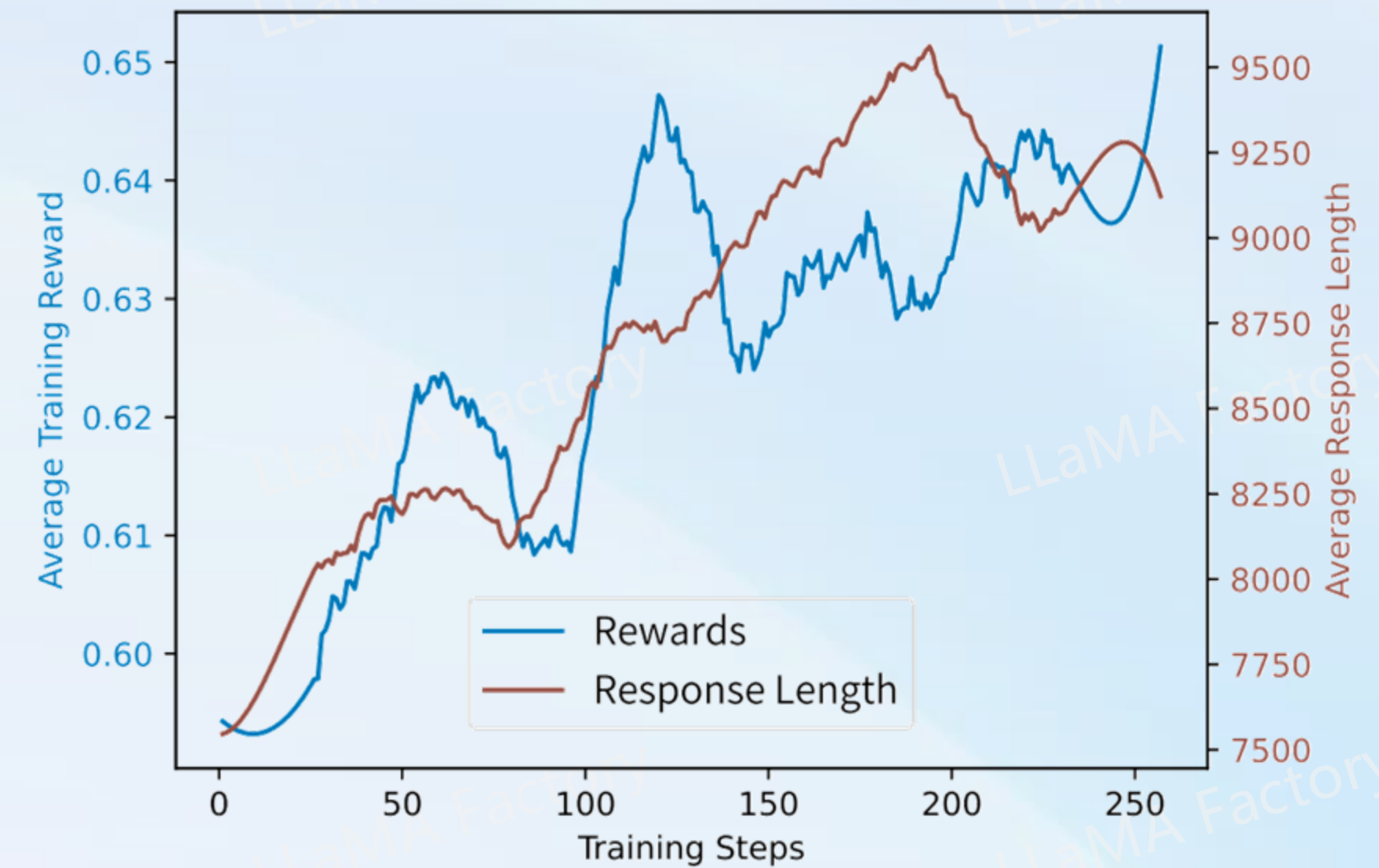
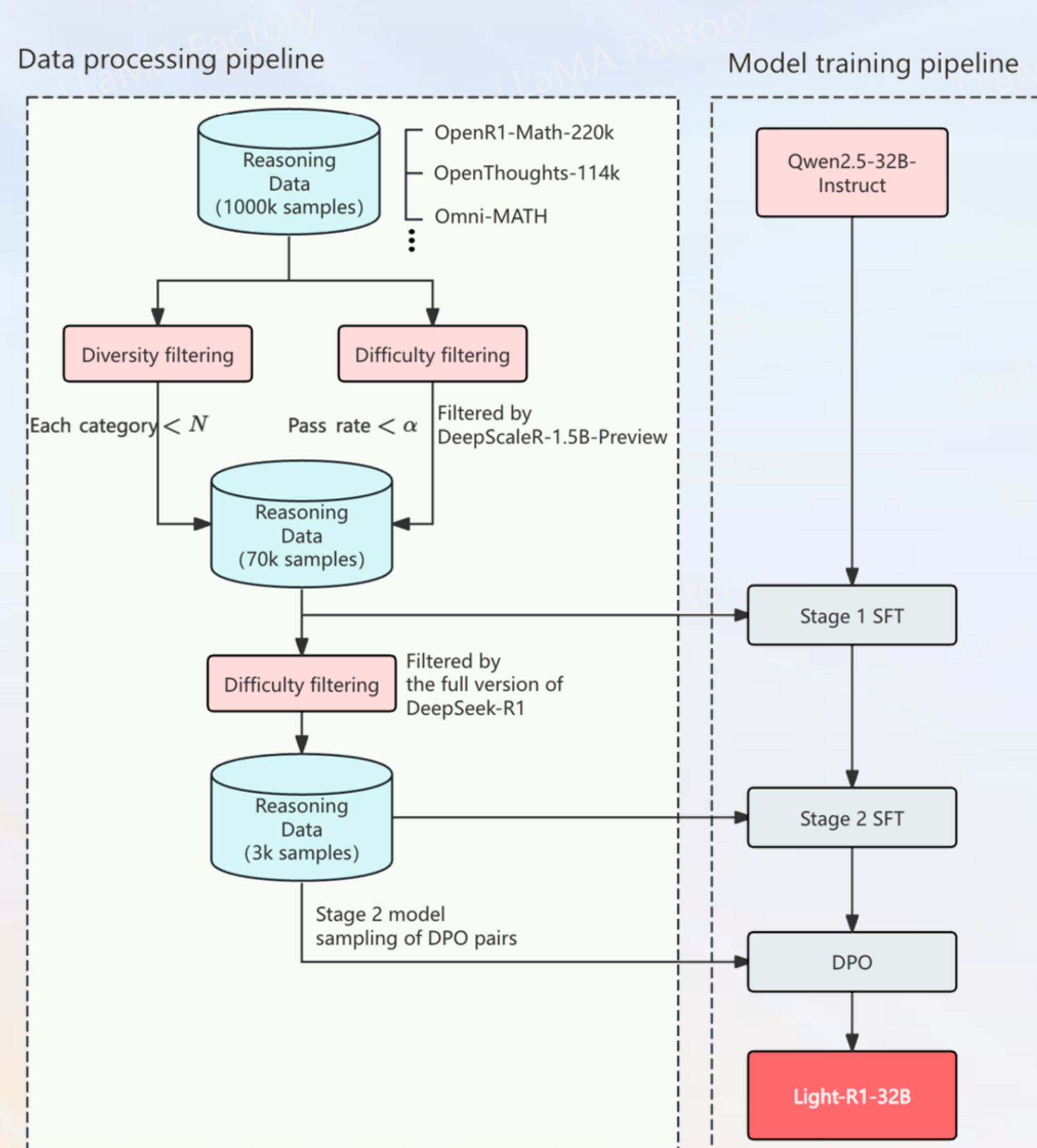
ChatGLM2-6B-Chat

https://v.youku.com/v_show/id_XNjQ2OTYyMDM0NA==.html

开源社区 & 外部应用



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

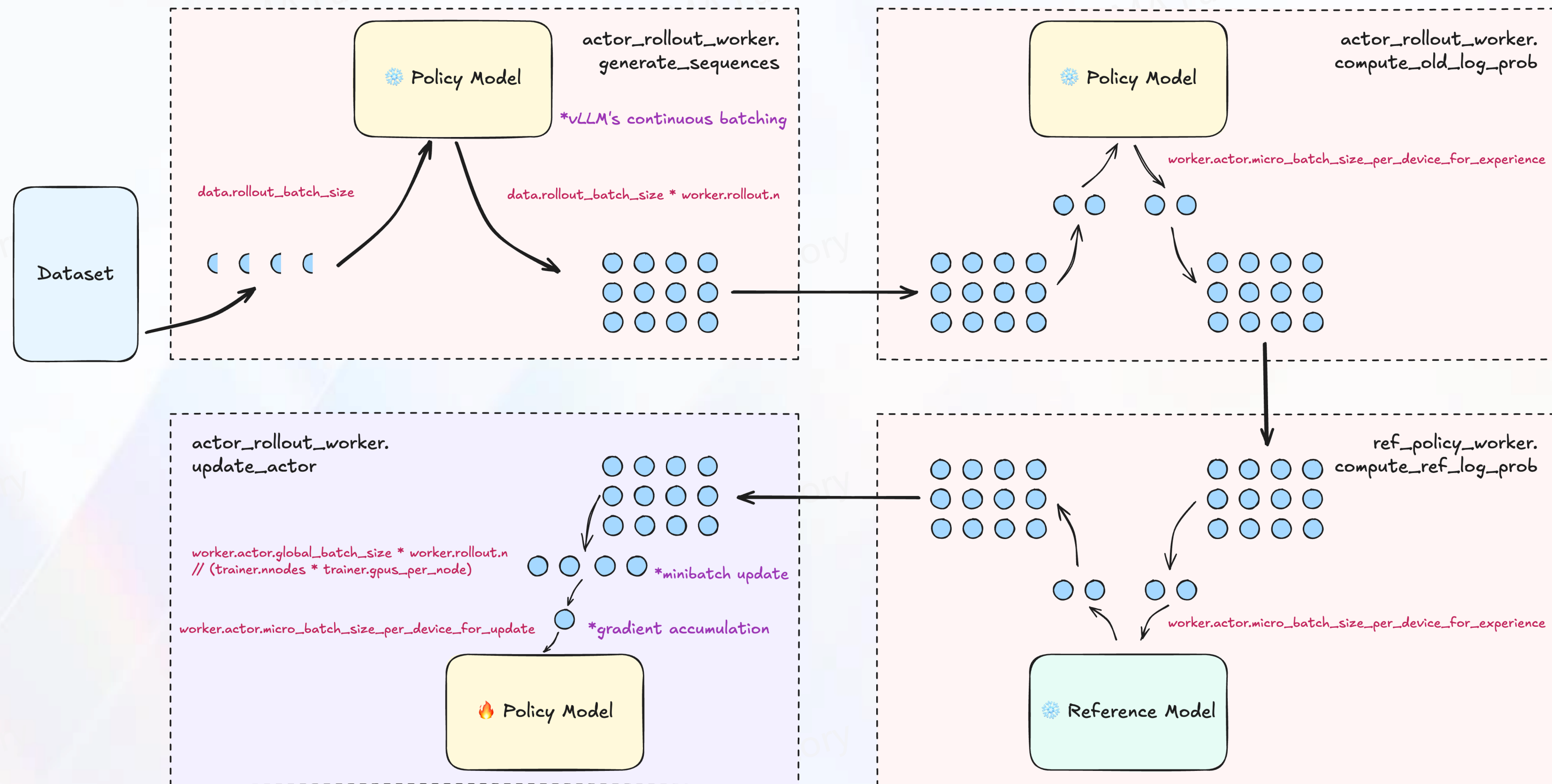
Command	Measured Acc	Reported Acc
mathruler gen meta-llama/Llama-3.1-8B-Instruct	50.8%	51.9%
mathruler gen Qwen/Qwen2.5-Math-7B-Instruct	82.6%	83.6%
--temperature 0.5 --sample_num 4	90.4%	-
--json_path data/gsm8k_splits/test.jsonl	95.6%	95.2%

```
from mathruler.grader import extract_boxed_content, grade_answer

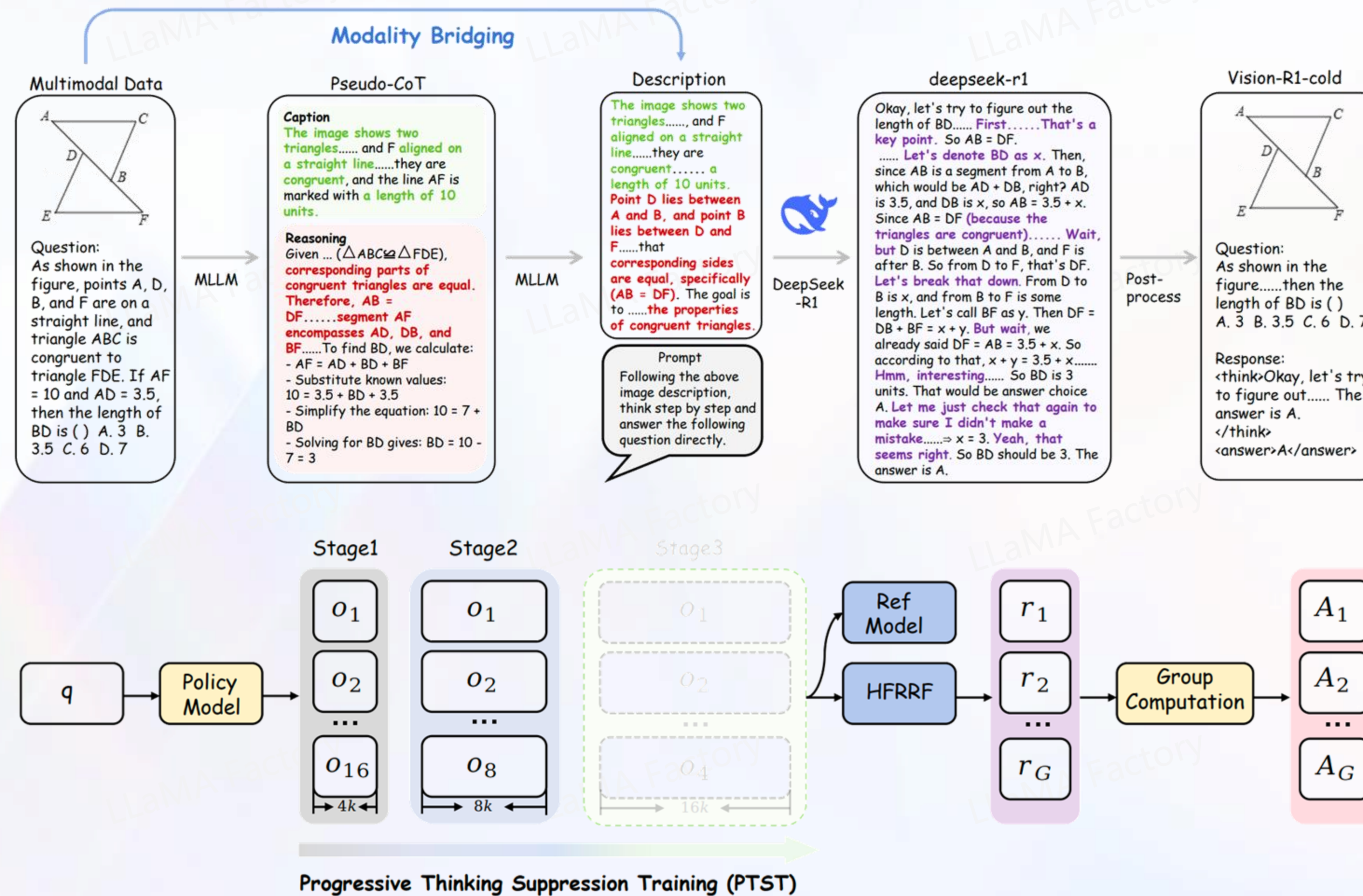
grade_answer(given_answer: str, ground_truth: str)
grade_answer(extract_boxed_content(generated_result: str), answer: str)
```

MathRuler

基于 vLLM 和 veRL 的多模态强化学习框架



Vision-R1: Incentivizing Reasoning Capability in MLLMs



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



44k Stars

谢谢
Thank You