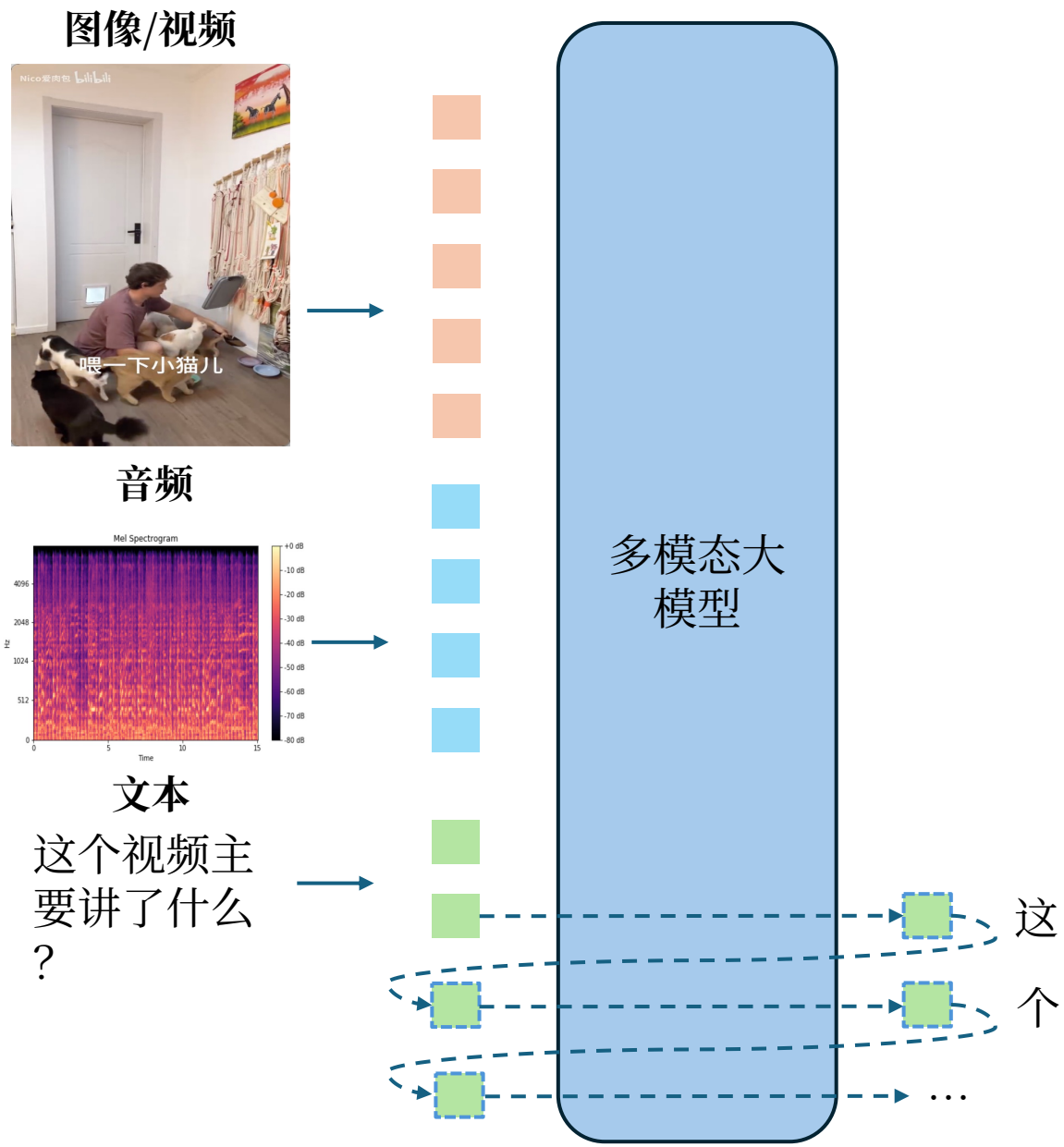
Post-Training of Multimodal Large Language Models

Data, Optimization and Evaluation

Qi Xu

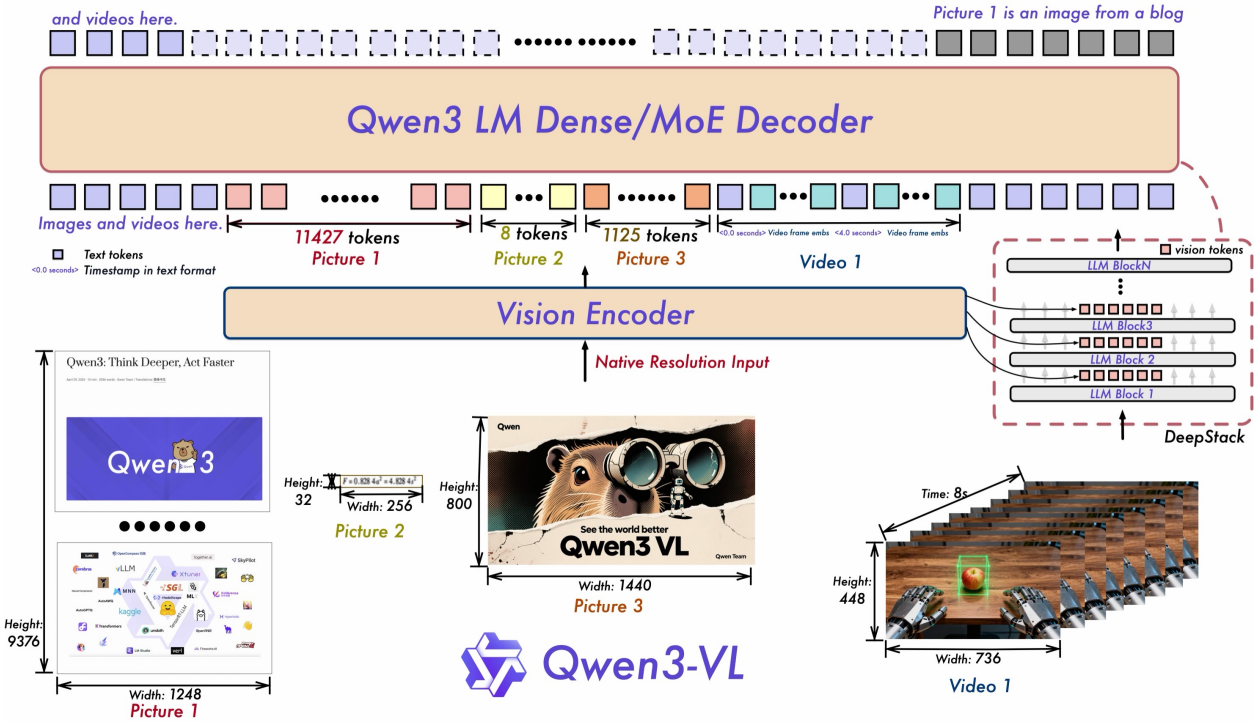
2026.08.06

什么是多模态大模型



它能够接收并理解文本以外的信息，例如
图像、视频和音频
它能够输出文本内容
预测下一个词 next-token-prediction

感知内容 → 融合模态 → 理解关系 → 回答问题



https://qwen.ai/blog?id=99f0335c4ad9ff6153e517418d48535ab6d8afef

我们希望模型如何理解这段视频

“看懂视频”不是一个足够明确的训练目标。后训练的第一步，是定义我们希望模型产生什么行为。

这个视频主要讲了什么？

这个视频记录了一个外国男生（Nico）在中国充满波折但又温馨的一天。主要内容包括：

1. **早晨的突发状况**：他起床后准备迎接一天的考试（概率统计和Python），但在喂猫时发现爱猫“大白”不吃东西。
2. **带猫就医**：他赶紧带大白去宠物医院，医生诊断大白是尿闭，需要导尿和手术，这让他一早就花掉了3000元，让他感叹六月对他好一点。
3. **参加考试**：安顿好猫咪后，他匆忙复习并去参加Python期末考试。考前他非常紧张，甚至开玩笑说“28岁还怕挂科”。
4. **成绩惊喜**：考完试后，他觉得题目很简单。随后去办公室查前一天概率统计的成绩，意外发现自己考了82分，这对于数学不好的他来说是个巨大的惊喜，连老师都感到惊讶。
5. **探望猫咪与手账解压**：得知好消息后，他去看了做完手术、戴着伊丽莎白圈的大白，确认它平安无事。晚上回家后，他通过做手账日记来记录这“起起伏伏”的一天，将小猫生病、考试、查成绩等经历画了下来，以此作为放松和总结。

总的来说，这是一个关于留学生活、宠物关爱以及面对压力时自我调节的日常Vlog。

请按时间顺序列出主要事件，并给出事件对应的时间戳。

基于视频内容，以下是按时间顺序排列的主要事件及其对应的时间戳：

- **00:00 - 00:02**: 男子起床，伸懒腰，视频配有“早上好”和“起床先发射激光”的字幕及特效。
- **00:03 - 00:08**: 男子在衣帽间涂抹止汗露，字幕预告他“一会儿将会痛失3000元”。
- **00:09 - 00:15**: 男子一边刮胡子一边提到昨天考了概率统计，下午还要考Python，并祈祷好运。
- **00:16 - 00:22**: 男子给家里的几只猫喂食，发现他的白猫“大白”没有吃饭，显得很担心。
- **00:23 - 00:27**: 男子带着白猫“大白”骑电动车去宠物医院，并在医院里抱着它，希望它没事。
- **00:28 - 00:31**: 兽医为“大白”做B超检查，诊断结果是尿闭，需要先导尿再做手术。
- **00:32 - 00:35**: 男子得知早上就为猫咪花了3000元，表情无奈，感叹六月对他好一点。
- **00:36 - 00:38**: 男子回到家中，在升降桌前站着学习，为下午的考试做准备。
- **00:39 - 00:43**: 男子抱着康复中的白猫，开心地跳舞庆祝。
- **00:44 - 00:47**: 男子骑车出发去期末考试，脸上用了搞怪的特效，表达不想考试和没准备好的心情。
- **00:48 - 00:51**: 男子在户外长廊里表现出对挂科的恐惧，然后跑开。
- **00:52 - 00:58**: 男子在教室里，抱怨班上很吵，并与画外音互动。
- **00:59 - 01:04**: 男子考完试后走出教学楼，高兴地说Python考试很简单，并开玩笑说老师可能觉得他们笨所以出题简单。
- **01:05 - 01:12**: 男子收到通知去办公室看昨天概率统计的成绩，他惊喜地发现自己考了82分，并兴奋地旋转拍摄。
- **01:13 - 01:16**: 男子去宠物医院看望戴着伊丽莎白圈的“大白”，看到它还活着很开心，并为它加油。
- **01:17 - 01:43**: 男子回到家，通过做手帐日记来解压，记录下今天起伏的心情和发生的各种事情（猫生病、考试、查成绩），最后享受放松时光。
- **01:44 - 01:44**: 男子展示完成的手帐，并向观众挥手告别。

主人公参加考试之前做了什么？

根据视频内容，主人公在参加下午的Python考试之前，做了以下几件事：

1. **起床并洗漱**：视频开始是他起床，然后涂止汗露、刮胡子。
2. **喂猫**：他给家里的几只猫喂食。
3. **带猫看病**：他发现其中一只叫“大白”的白猫不吃饭，于是带它去看兽医。医生诊断大白是尿闭，需要导尿和手术，这花了他3000块钱。
4. **回家学习**：从医院回来后，因为下午还要考试，他先在家的书桌前学习了一会儿。

完成这些事情后，他才出发去参加期末考试。

视频主人公喂猫的时候画面中一共出现了多少只猫？

要确定视频主人公喂猫时画面中出现的猫的数量，我们可以查看视频在00:16到00:20之间的帧。

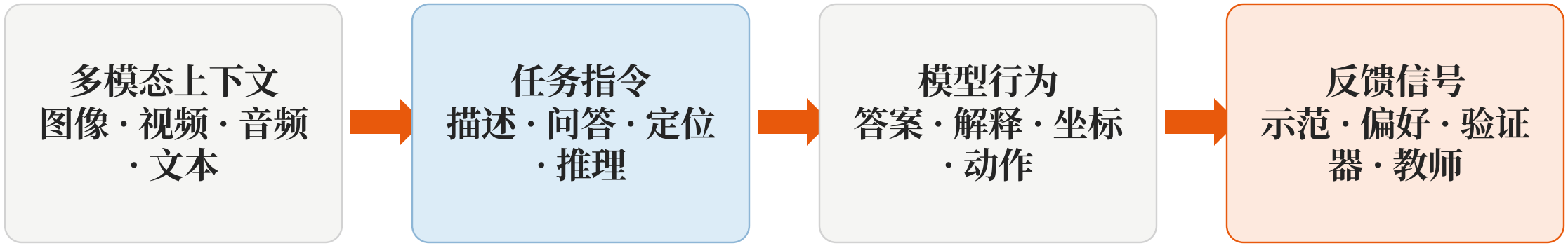
- 在00:17，我们可以看到一只橘猫，一只黑白相间的猫，和一只白色的猫。
- 在00:18，除了上述三只猫，还可以看到一只黑色的猫。
- 在00:19，所有的四只猫都在画面中清晰可见：一只橘猫在左下角，一只黑白相间的猫在右下角，一只黑色的猫在左上角，一只白色带橘色斑点的猫在右上角。

因此，视频主人公喂猫的时候画面中一共出现了4只猫。

同一个视频，可以对应完全不同的任务、答案与评价标准

后训练的基本单位：多模态交互与反馈

一条有效训练样本，需要同时说明：模型看到了什么、被要求做什么、实际做了什么，以及我们如何判断。



同一段视频，因为指令与反馈不同，可以被训练成完全不同的能力。

后训练数据，不只是 Input–Answer，而是 Input–Behavior–Feedback。

不同反馈，决定不同的训练方式

先问“我们能可靠提供什么反馈”，再问“用哪个算法”。

标准示范	**SFT**	教模型应该怎样回答
成对偏好	**DPO**	教模型哪个回答更好
可验证结果	**RLVR**	让模型通过 Rollout 探索
教师分布	**OPD**	在学生轨迹上获得稠密指导

算法不是起点；可信反馈的形态，才是方法选择的边界。

SFT：用示范建立行为起点


```
{
  "messages": [
    {
      "role": "user",
      "content": "<video>这个视频主要讲了什么？"
    },
    {
      "role": "assistant",
      "content": "这个视频记录了一个外国男生（Nico）在中国充满波折但又温馨的一天。主要内容包括：\n\n**早晨的突发状况**：他起床后准备迎接一天的考试（概率统计和 Python），但在喂猫时发现爱猫“大白”不吃东西。\n\n**带猫就医**：他赶紧带大白去宠物医院，医生诊断大白是尿闭，需要导尿和手术。这让他一早就花掉了 3000 元，也让他感叹“六月对他好一点”。\n\n**参加考试**：安顿好猫咪后，他匆忙复习并参加 Python 期末考试。考前他非常紧张，甚至开玩笑说“28 岁还怕挂科”。\n\n**成绩惊喜**：考完后，他觉得题目很简单。随后，他去办公室查询前一天概率统计考试的成绩，意外发现自己考了 82 分。对于不擅长数学的他来说，这是一个巨大的惊喜，连老师都感到惊讶。\n\n**探望猫咪与手账解压**：得知成绩后，他去看望做完手术、戴着伊丽莎白圈的大白，确认它平安无事。晚上回家后，他通过制作手账日记记录这“起起伏伏”的一天，将猫咪生病、参加考试和查询成绩等经历画下来，以此放松并总结一天。\n\n总的来说，这是一个关于留学生活、宠物关爱，以及面对压力时进行自我调节的日常 Vlog。"
    }
  ],
  "videos": [
    "30548099099-1-192.mp4"
  ]
}
```


SFT 最擅长

- 建立输出格式与基本任务行为
- 把领域知识与任务接口写进模型
- 为后续偏好或在线优化提供稳定起点

$$\mathcal{L}_{\text{SFT}} = -\mathbb{E}_{(x,y)\sim\mathcal{D}} \sum_{t} \log \pi_\theta(y_t | x, y_{<t})$$

SFT 学习“像示范那样回答”，并不自动学会探索与自我纠错。

SFT 之后，问题从“不会”变成“做不好”

当模型已经“会答”，新的问题往往是稳定性、分布与泛化。

Exposure Bias	分布收窄	教师上限	灾难性遗忘
训练看真实前缀 推理看自身前缀	输出趋于模板化 多样性下降	合成数据复制 教师偏差与盲点	专项能力上升 通用能力下降

因此：用偏好区分质量，用 On-policy 数据覆盖自身分布，用 Replay 保持已有能力

SFT 之后，优化目标从“模仿答案”转向“改善模型自己的行为分布”。

DPO：学习哪个回答更好

视频主人公喂猫的时候画面中一共出现了多少只猫?

已经完成思考 >

喂猫这一段镜头里总共出现了 **7 只猫**：

正在吃饭的 6 只：

1. 黑猫（画面左侧趴着吃）
2. 黑猫（男生身边低头吃）
3. 橘猫（画面下方吃蓝碗里的粮）
4. 奶牛猫（黑白，右下方吃绿碗里的粮）
5. 白底橘斑猫（右侧吃食）
6. 橘色小猫（右侧靠墙吃食）

另外还有一只纯白的"大白"，它没来吃饭，而是躺在旁边的沙发上，所以男生才发现它不对劲、后来带它去看医生。

合计：6 只吃饭的 + 1 只不吃的大白 = **7 只猫**。

chosen

视频主人公喂猫的时候画面中一共出现了多少只猫?

要确定视频主人公喂猫时画面中出现的猫的数量，我们可以查看视频在00:16到00:20之间的帧。

- 在00:17，我们可以看到一只橘猫，一只黑白相间的猫，和一只白色的猫。
- 在00:18，除了上述三只猫，还可以看到一只黑色的猫。
- 在00:19，所有的四只猫都在画面中清晰可见：一只橘猫在左下角，一只黑白相间的猫在右下角，一只黑色的猫在左上角，一只白色带橘色斑点的猫在右上角。

因此，视频主人公喂猫的时候画面中一共出现了4只猫。

reject

适合优化

- 事实性与视觉忠实度
- 指令遵循与输出风格
- 安全性与拒答边界

$$\mathcal{L}_{\text{DPO}}(\pi_\theta; \pi_{\text{ref}}) = -\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}}\left[\log\sigma\left(\beta\log\frac{\pi_\theta(y_w \mid x)}{\pi_{\text{ref}}(y_w \mid x)} - \beta\log\frac{\pi_\theta(y_l \mid x)}{\pi_{\text{ref}}(y_l \mid x)}\right)\right]$$

限制：只能在比较数据覆盖到的行为之间学习；偏好标签若含糊，优化方向也会含糊。

DPO 适合“能比较好坏，但难写出精确奖励”的任务。

RLVR：让模型从可验证结果中学习

GT: 7只

视频主人公喂猫的时候画面中一共出现了多少只猫?

要确定视频主人公喂猫时画面中出现的猫的数量，我们可以查看视频在00:16到00:20之间的帧。

- 在00:17，我们可以看到一只橘猫，一只黑白相间的猫，和一只白色的猫。
- 在00:18，除了上述三只猫，还可以看到一只黑色的猫。
- 在00:19，所有的四只猫都在画面中清晰可见：一只橘猫在左下角，一只黑白相间的猫在右下角，一只黑色的猫在左上角，一只白色带橘色斑点的猫在右上角。

因此，视频主人公喂猫的时候画面中一共出现了4只猫。

→ Answer1: 4只 → Reward1: 0.43

$$reward = 1 - \frac{|Answer - GT|}{GT}$$

$$\hat{A}_{i,t} = \tilde{r}_i = \frac{r_i - \text{mean}(\mathbf{r})}{\text{std}(\mathbf{r})}$$

视频主人公喂猫的时候画面中一共出现了多少只猫?

已经完成思考 >

喂猫这一段镜头里总共出现了 **7 只猫**：

正在吃饭的 6 只：

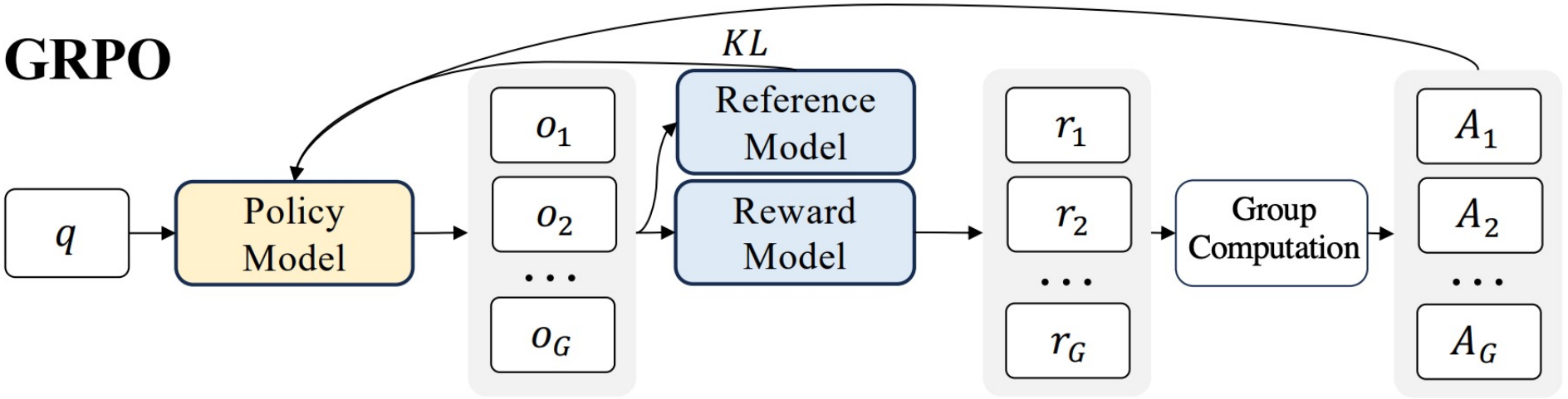
1. 黑猫（画面左侧趴着吃）
2. 黑猫（男生身边低头吃）
3. 橘猫（画面下方吃蓝碗里的粮）
4. 奶牛猫（黑白，右下方吃绿碗里的粮）
5. 白底橘斑猫（右侧吃食）
6. 橘色小猫（右侧靠墙吃食）

另外还有一只纯白的"大白"，它没来吃饭，而是躺在旁边的沙发上，所以男生才发现它不对劲、后来带它去看医生。

合计：6 只吃饭的 + 1 只不吃的大白 = **7 只猫**。

→ Answer2: 7只 → Reward1: 1

RLVR：让模型从可验证结果中学习



$$\mathcal{J}_{GRPO}(\theta) = \mathbb{E}[q \sim P(Q), \{o_i\}_{i=1}^{G} \sim \pi_{\theta_{old}}(O|q)]$$

$$\frac{1}{G}\sum_{i=1}^{G}\frac{1}{|o_i|}\sum_{t=1}^{|o_i|}\left\{\min\left[\frac{\pi_\theta(o_{i,t}|q,o_{i,<t})}{\pi_{\theta_{old}}(o_{i,t}|q,o_{i,<t})}\hat{A}_{i,t}, \text{clip}\left(\frac{\pi_\theta(o_{i,t}|q,o_{i,<t})}{\pi_{\theta_{old}}(o_{i,t}|q,o_{i,<t})}, 1-\varepsilon, 1+\varepsilon\right)\hat{A}_{i,t}\right] - \beta\mathbb{D}_{KL}\left[\pi_\theta||\pi_{ref}\right]\right\},$$

RLVR：让模型从可验证结果中学习

当结果可以被可靠验证，模型就能在自身 Rollout 中探索更好的策略。

更容易验证：结构化抽取 · OCR / 计数 · 时空定位 · 工具执行

风险：若 Verifier 只检查最终答案，模型可能找到“得分捷径”，而不是学会可靠地看与推理。

RLVR 的瓶颈通常不是优化器，而是验证器的覆盖率与可信度。

Reward 设计决定模型真正优化什么

R = Answer + Evidence + Format − Hallucination

- Answer：最终答案是否正确
- Evidence：是否引用了正确视觉 / 时间证据
- Format：输出是否可解析、可执行
- Hallucination：是否添加了输入中不存在的内容

Model	C-Acc	tF1@0.3	tF1@0.5	tF1@0.7	tIoU	EtF1
Base	0.31	37.07	26.75	17.93	30.42	0.21
SFT	44.06	69.57	61.23	45.63	56.94	34.81
RL	**55.63**	**73.46**	**65.40**	**48.96**	**61.24**	**43.65**

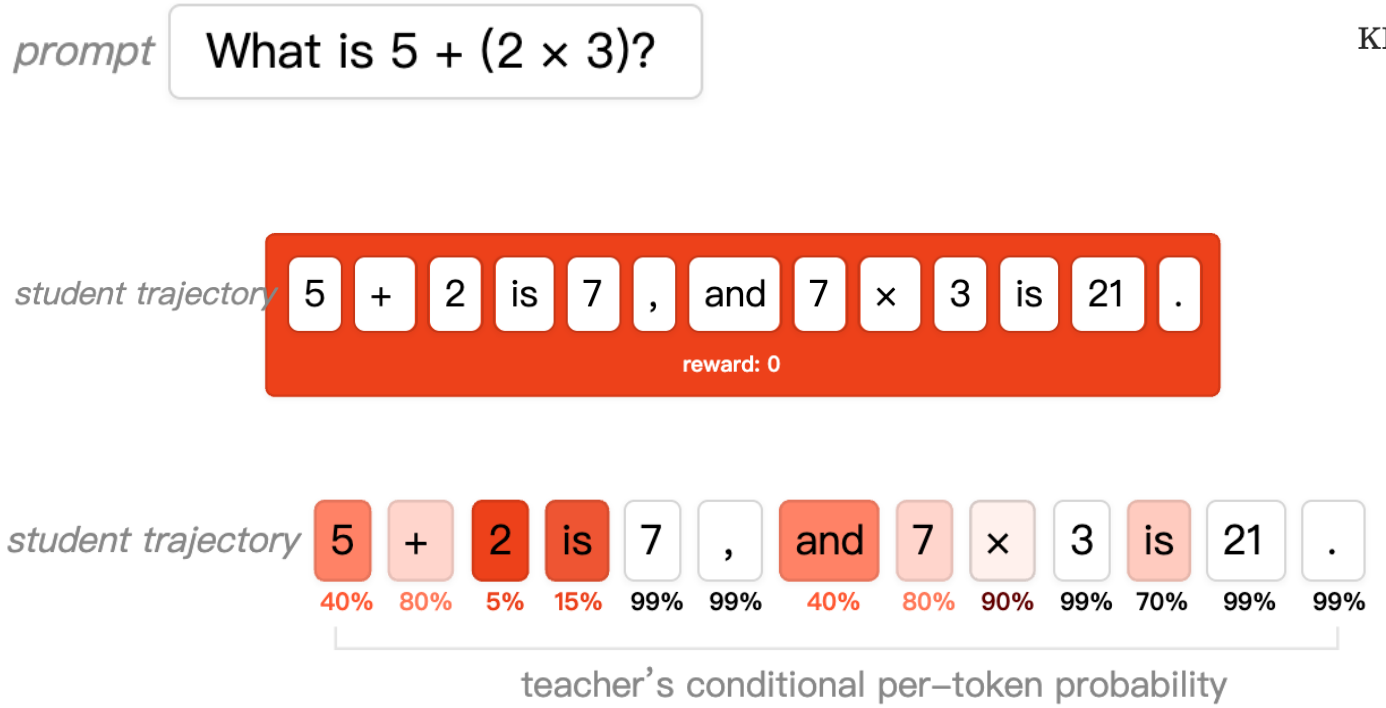
Reward Functions	C-Acc	tF1@0.3	tF1@0.5	tF1@0.7	tIoU	EtF1
R_{tIoU}	+0.31	+2.01	+1.82	+0.28	+2.61	+0.74
$R_{\text{tIoU}} + R_{\text{C-Acc}}$	+7.50	+4.69	+3.86	+3.27	+4.03	+5.87
$R_{\text{tIoU}} + R_{\text{C-Acc}} + R_{\text{Caption}}$	**+11.57**	**+3.89**	**+4.17**	**+3.33**	**+4.30**	**+8.84**

实践建议：分解奖励、记录每个分项，并检查高分样本的真实行为。

能被奖励的行为会被持续放大；遗漏的目标则不会自动出现。

OPD：在学生自己的轨迹上蒸馏

蒸馏的关键变化：让教师回答“学生真正会走到的状态”。



$$\mathrm{KL}\left(\pi_\theta || \pi_{\text{teacher}}\right) = \mathbb{E}_{x\sim\pi_\theta}\left[\log \pi_\theta(x_{t+1}|x_{1..t}) - \log \pi_{\text{teacher}}(x_{t+1}|x_{1..t})\right]$$

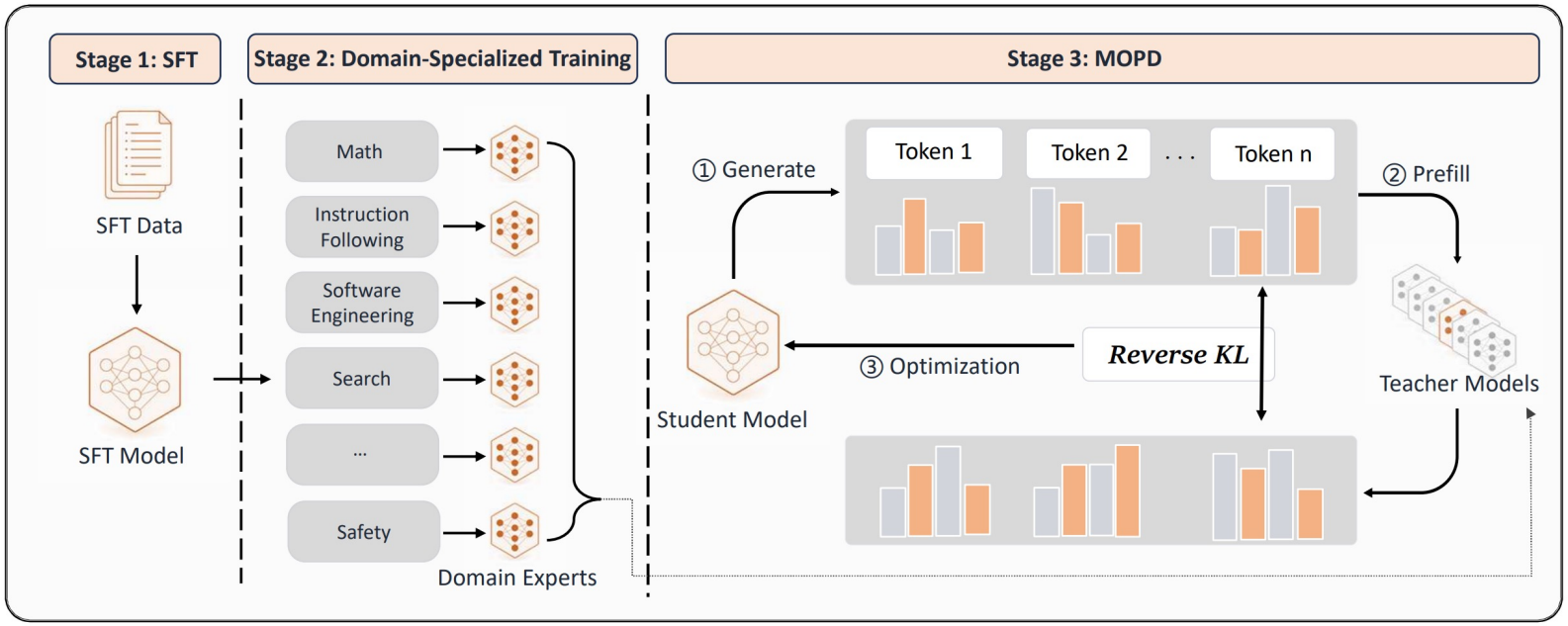
稠密 token-level 信号

减轻 Exposure Mismatch

代价是教师推理成本与误差传播

OPD 把教师从“答案生成器”变成“学生轨迹上的在线教练”。

OPD：融合多个教师能力



通过 OPD 来对多个专家模型进行融合

MODEL	PASS@1 ↑	NORM. LEVENSHTEIN ↓	ADDED CC ↓	LIVECODEBENCH V6 ↑
Teachers				
SFT	0.775	0.450	0.450	0.286
RL	0.792	0.063	0.206	0.320
Students (OPD)				
OPD SFT teacher	**0.800**	0.059	**0.206**	0.297
OPD RL teacher	0.787	**0.055**	0.228	**0.314**

Teachers vs. on-policy distillation students on the minimal editing task (out-of-domain corruptions) and LiveCodeBench v6.

OPD 后的模型不仅性能上能获得提升，也能更少遗忘

先看反馈条件，再选择算法

把方法选择改写成四个可回答的问题。

有稳定的标准示范吗?	**SFT**	建立行为起点
能稳定比较两个回答吗?	**DPO**	优化相对偏好
有可靠且便宜的 Verifier 吗?	**RLVR**	在线探索与自改进
能访问强教师的分布或反馈吗?	**OPD**	学生轨迹上的稠密指导

现实 Recipe 往往是组合：SFT 起步，偏好 / RL / 蒸馏改善，Replay 保持。

数据从哪里来：人工、合成与 Rollout

数据的价值不只取决于来源，而取决于它经过了怎样的筛选与验证。

人工数据
任务定义清楚、质量高
但昂贵且难扩展

合成数据
覆盖长尾、可控生成
但会继承教师偏差

Rollout 数据
来自当前模型分布
最接近真实失败模式

高质量数据集，更像一条持续运行的生产线，而不是一次性收集。

高质量数据，要同时满足四件事

规模只有在质量成立之后才有意义。

01

正确

事实与标注可信

02

相关

反馈对准目标行为

03

多样

覆盖任务、场景与难度

04

可验证

能自动或低成本复核

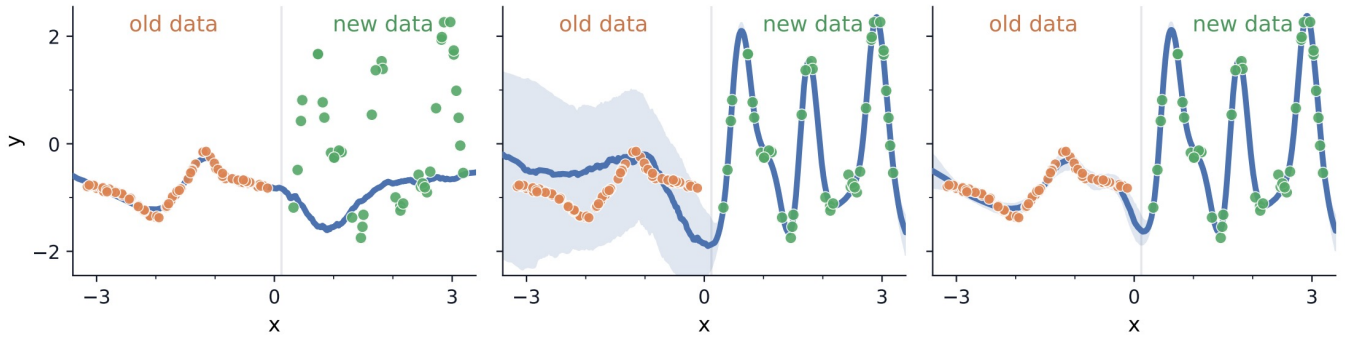
数据筛选的核心问题：这条反馈是否真的能把模型推向我们想要的行为?

“更多数据”不是策略；明确的数据质量标准才是。

数据配比决定能力增益，也决定遗忘

目标能力数据提高专项表现；Replay 决定旧能力能否保住。

- 目标任务数据：直接推动专项能力
- 通用多模态 Replay：保持感知与问答能力
- 通用文本 / 推理 Replay：保持语言与推理底座
- 安全与拒答数据：约束行为边界
- Hard / Failure Cases：覆盖当前模型的盲点



不要把 Replay 当作“顺手加一点旧数据”；它是显式的能力保持机制。

数据配比同时决定学习速度、遗忘程度与最终能力边界。

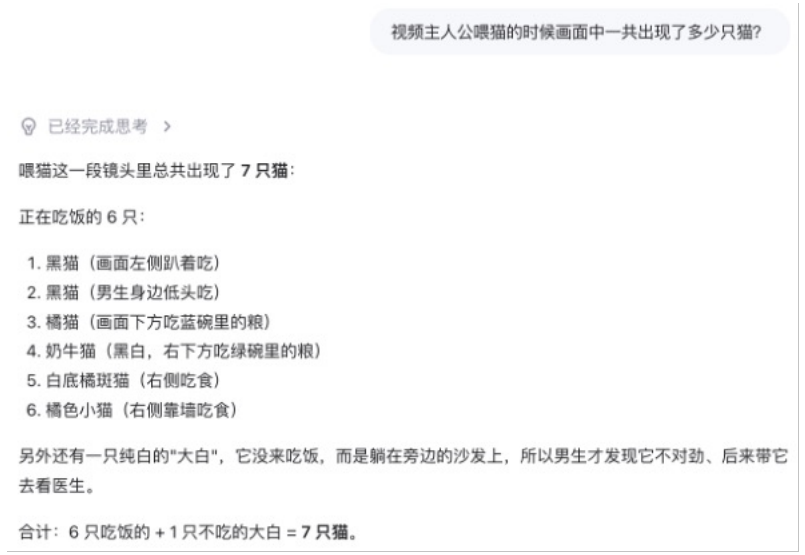
评估要同时检查答案、证据与代价

单一 Benchmark 分数无法回答：模型为什么变好，以及代价是什么。

答案
正确率 · 完整性
任务成功率

证据
视觉忠实度 · 定位
时序 / 空间一致性

代价
通用能力保持
鲁棒性 · 延迟 · 成本

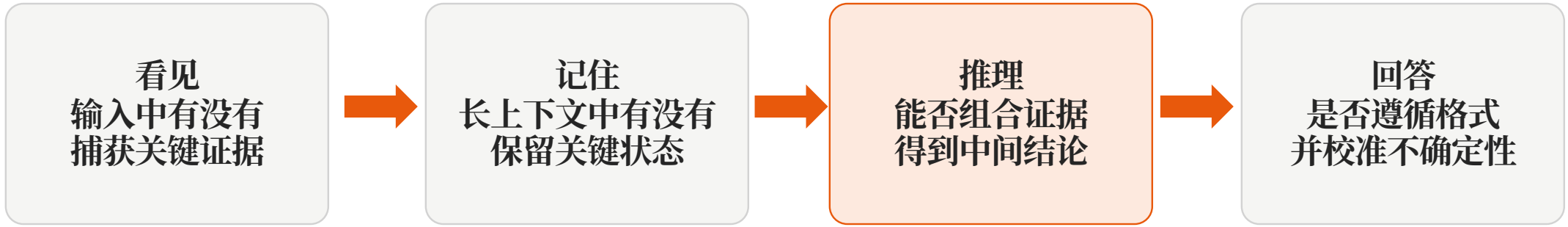


评估切片：任务 × 模态 × 难度 × 数据来源 × 能力阶段

答案正确是必要条件；证据可靠与能力保持决定模型是否真的可用。

诊断失败，要拆开看能力链

不要把所有错误都归因于“推理能力不足”。



先定位失败发生在哪一环，再决定改数据、改反馈、改模型还是加计算。

一套可落地的多模态后训练 Recipe

一个最小可行闭环：先把目标和评估钉住，再逐步增加训练复杂度。

每轮 Gate：目标能力 ↑　通用能力 ↔　训练 / 推理成本可控

先建立可复现的小闭环，再扩数据、算法与算力。

2026+：后训练走向持续学习闭环

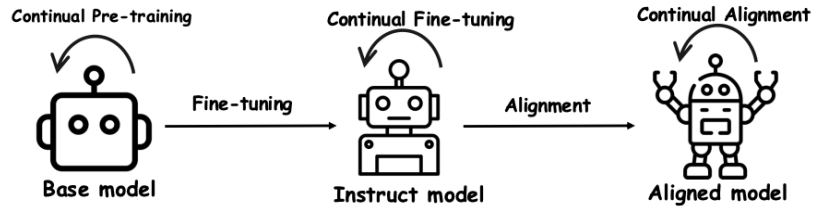
新的趋势不是更多方法名，而是反馈闭环变得更在线、更细粒度、更持续。

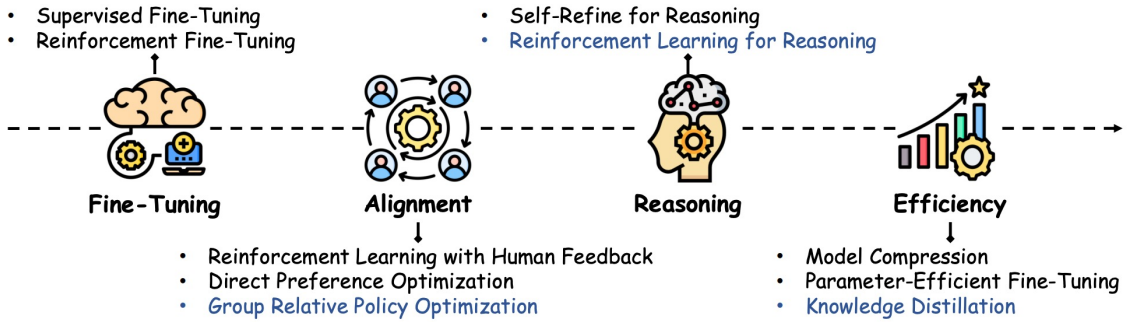
Offline Dataset → **On-policy Experience**

Terminal Reward → **Process / Evidence Feedback**

Single Skill → **Capability Integration**

One-off Training → **Continuous Learning Flywheel**





后训练的核心，不是某个方法
而是一套可持续产生可信反馈的闭环。