第一章

（第1节）

大模型对齐算法8年演进四大家族一次讲透

唐国梁Tommy

唐国梁Tommy | TGLTommy 原创出品 © 版权所有

课程专享知识库使用说明

本课程除 PDF 资料外，还配套提供课程专享知识库文档，内容包括核心知识点梳理、补充资料、代码说明、常见问题和更新内容，建议结合课程视频同步学习。课程知识库将随课程持续更新，建议定期查看最新内容。

- 通过官网学习的同学：可直接在对应课程视频下方查看知识库文档。
- 通过 B 站学习的同学：可通过飞书知识库查阅相同的内容。如尚未开通知识库权限，请联系课程助理，并提供订单信息进行核验。（请注意：每个课程订单仅可绑定一个飞书账号，绑定后不支持更改。请确认使用本人常用飞书账号申请开通。）

官方网站：TGLTommy.com（国内访问需科学上网）

2025 年初 · DeepSeek-R1

DeepSeek-R1的冲击

把一个问题推到了台前

核心追问

数学题能自动判对错，代码能跑测试 ——

模型能不能 不依赖人类示范，直接靠奖励信号学会推理?

唐国梁Tommy | TGLTommy © 版权所有 原创出品

R1-Zero

✕ 不喂人工推理轨迹

✓ 仅规则奖励 · 大规模 RL

↓ 模型自发涌现 ↓

自我反思
Self-Reflection

验证
Verification

长链推理
Long-Chain Reasoning

唐国梁Tommy | TGLTommy 原创出品 © 版权所有

后训练到底在干什么？

这是 R1 真正改变的认知问题

观点 A · TEACHING

RL 在「教」

奖励信号塑造出 全新的能力

模型从 0 学会推理、反思、验证

?

观点 B · ELICITING

RL 在「引出」

预训练里 早已埋下能力

RL 只是把它们激活、显化

唐国梁Tommy | TGLTommy 原创出品 © 版权所有

要回答这个问题，先把整条演化史画出来 →



八年演化史 · 算法地图 · 从此展开

唐国梁Tommy | TGLTommy 原创出品 © 版权所有



对齐问题的本质

大语言模型，为什么需要"驯化"？

唐国梁Tommy | TGLTommy 原创出品 © 版权所有

预训练 LLM： 能力是 中性 的

既能写优美的诗，也能编造极其可信的谎言

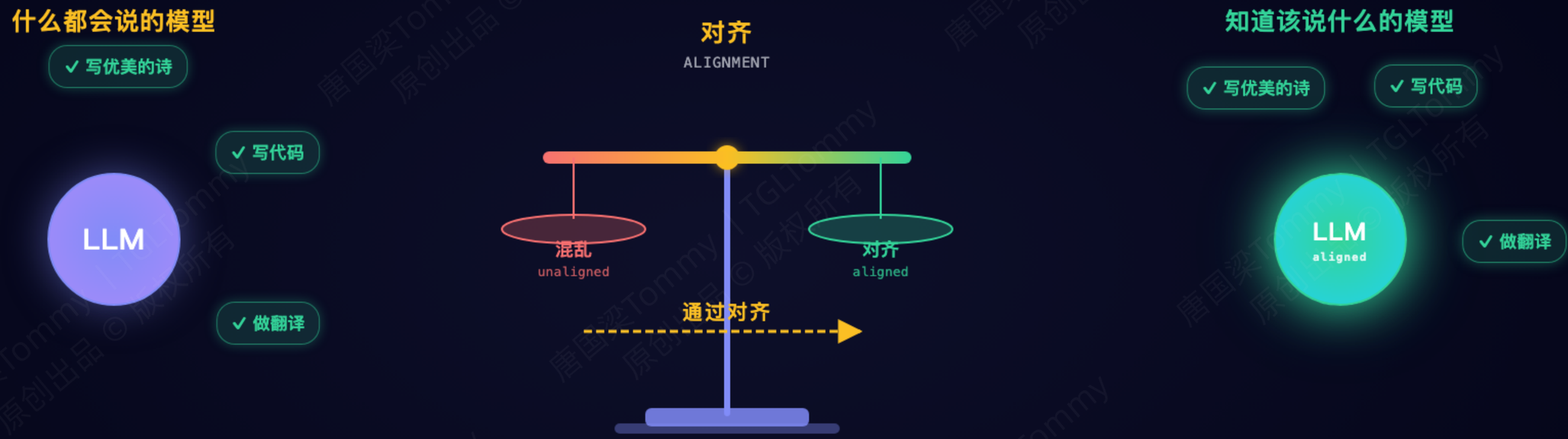


预训练目标 = 预测下一个 token，不是 "做一个有用、安全、可靠的助手"

唐国梁Tommy | TGLTommy 原创出品 © 版权所有

对齐问题的本质

把"什么都会说"的模型，变成"知道该说什么"的模型



知道该说什么、怎么说、什么时候不该说 —— 同时不能损伤核心能力

唐国梁Tommy | TGLTommy 原创出品 © 版权所有

贯穿全领域的 一组张力

一边要让模型听话，一边要让模型聪明



过去 8 年 的核心技术，基本都围绕这组张力展开

唐国梁Tommy | TGLTommy 原创出品 © 版权所有

2017

Deep RL from Human Preferences

Christiano et al., NeurIPS 2017 · RLHF 的思想原点

唐国梁Tommy丨TGLTommy 原创出品 © 版权所有

核心洞察

比较 ≠ 定义

人类 不擅长 描述「什么是好行为」
却 很擅长 告诉你「A 比 B 更好」

✕ 写规则

✕ 标尺打分

✓ 选 A 还是 B



唐国梁Tommy | TGLTommy 原创出品 © 版权所有

三段式流水线

偏好比较 → 奖励建模 → 策略优化



唐国梁Tommy | TGLTommy 原创出品 © 版权所有

◆ RLHF · CORE SKELETON ◆

RLHF 的基本骨架

1



人类比较

Preference Comparison

标注员选 A 或 B

训练奖励模型

2



奖励模型

Reward Modeling

拟合偏好为标量奖励 r(x)

RL 优化

3



策略优化

Policy Optimization

RL 最大化 r(x)

01 人类比较

02 奖励模型

03 策略优化

唐国梁Tommy | TGLTommy 原创出品 © 版权所有

2017 · 控制任务 人类比较 → 奖励模型 → 策略优化

★ 迁移到大语言模型

Apply the same recipe to LLM alignment

LLM

大语言模型对齐

从控制任务 → 从指令遵循到 ChatGPT

真正改变了产品形态

RLHF 时代开启

唐国梁Tommy | TGLTommy 原创出品 © 版权所有

◆ RLHF · SPOTLIGHT MOMENT ◆

2022

RLHF 真正走进聚光灯

The year RLHF stepped into the spotlight

唐国梁Tommy | TGLTommy 原创出品 © 版权所有

—— OpenAI · InstructGPT → ChatGPT ——

2017
Deep RL from
Human Preferences

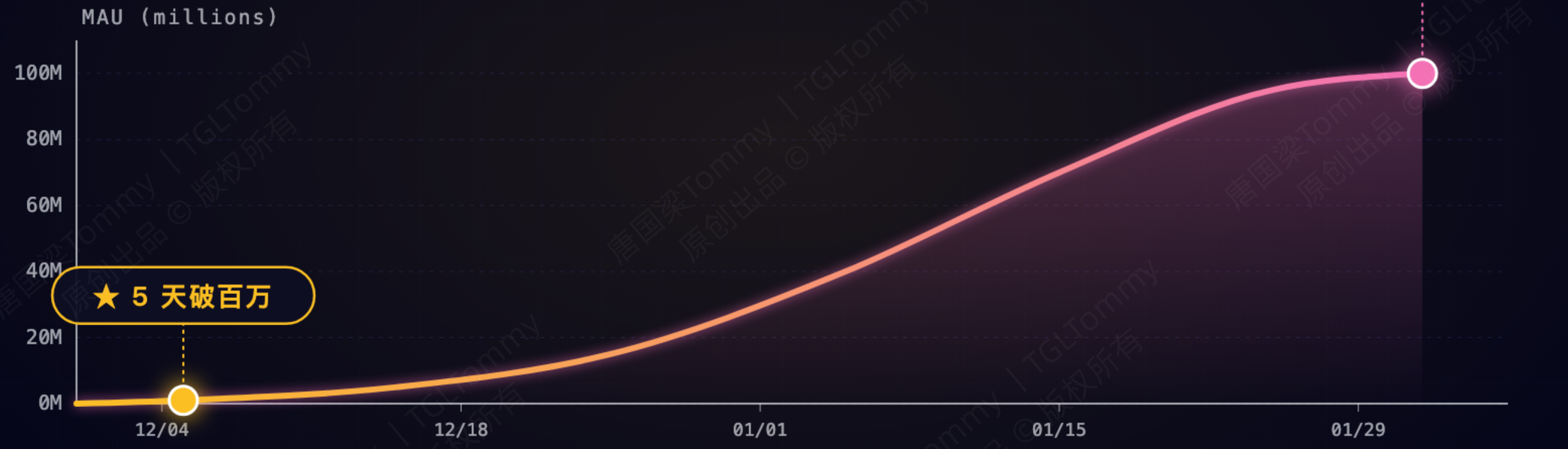
GPT-3
2020.05

InstructGPT
2022.03 · OpenAI

ChatGPT破亿用户
2022.11.30 · 引爆全球
2023.01

ChatGPT 月活用户（百万）

有史以来最快增速



唐国梁Tommy | TGLTommy 原创出品 © 版权所有

★ INDUSTRY STANDARD ★

三段式流水线 · 行业共识

01

偏好收集

Preference Collection

标注员对答案对打偏好标签

→

02

奖励建模

Reward Modeling

拟合偏好为标量奖励 r(x)

→

03

策略优化

Policy Optimization

RL 最大化奖励 + KL 约束

从 OpenAI 到 Anthropic、Google、Meta — **所有主流 LLM 团队都采用此三段式**

唐国梁Tommy | TGLTommy © 版权所有 原创出品

◆ DEFAULT ALGORITHM ◆

PPO

Proximal Policy Optimization

策略优化算法 · 最常见的选择

稳定	可调	易用
Clipped Surrogate	KL Penalty	Actor-Critic

唐国梁Tommy | TGLTommy 原创出品 © 版权所有

CAPABILITY LEAP

从「续写文本」到「遵循指令」

✗ Pre-RLHF · 续写文本

next-token prediction · GPT-3 default

用户： 请帮我写一封正式的辞职信。

模型： 请帮我写一封正式的辞职信。请帮我写一封正式的辞职信...请帮我写一封...或者关于离职的

（重复 / 离题 / 拒绝有缺陷 / 无组织）

× 重复　× 离题　× 无结构　× 不分语境

RLHF

✓ Post-RLHF · 遵循指令

instruction following · ChatGPT style

用户： 请帮我写一封正式的辞职信。

模型： 好的，下面是一封正式辞职信的范例：

1. 称呼 + 表达离职意愿
2. 简述离职原因（保持专业）
3. 工作交接安排
4. 致谢 + 落款

如需调整语气（更感性 / 更简短），请告诉我。

✓ 遵循指令　✓ 拒绝危险请求　✓ 结构化输出　✓ 按用户意图

大模型不再只是「会续写」 — 而是 **会遵循、会拒绝、会按意图组织**

唐国梁Tommy | TGLTommy 原创出品 © 版权所有

⚠ 关键代价 · CRITICAL COST

RLHF 有效，但很贵

训练时要同时维护 四个模型，加上 KL 约束、显存、工程、调参一齐压上来

唐国梁Tommy | TGLTommy 原创出品 © 版权所有

◆ FOUR-MODEL ARCHITECTURE ◆

四个模型同时在线

策略模型 `π_θ`

Policy / Actor

被训练的主模型

参考模型 `π_ref`

Reference

冻结的初始策略

GPU · HBM

共享显存

奖励模型 `r_φ`

Reward Model

拟合人类偏好

价值网络 `V_ψ`

Value / Critic

估计状态价值

唐国梁Tommy | TGLTommy 原创出品 © 版权所有

◆ KL CONSTRAINT + 3 COSTS ◆

KL 约束 · 三重代价

KL ≤ ε
防止策略跑偏

策略模型 π_θ
Policy / Actor
被训练的主模型

参考模型 π_ref
Reference
冻结的初始策略

GPU · HBM
共享显存

奖励模型 r_φ
Reward Model
拟合人类偏好

价值网络 V_ψ
Value / Critic
估计状态价值

显存占用
GPU HBM

HIGH ▲

4× 模型权重

工程复杂度
Pipeline Stages

HIGH ▲

多阶段串联

调参成本
GPU Hours

HIGH ▲

KL系数 / β / lr

唐国梁Tommy | TGITommy 原创出品 © 版权所有

本章只是地图视图 · 细节后续展开

CHAPTER 4 · DEEP DIVE

→ 第 4 章 · RLHF 全流程拆解

奖励建模 · PPO 训练 · KL 调度 · 工程实现

1 SFT 微调

2 奖励模型训练

3 PPO + KL 强化学习

唐国梁Tommy | TGLTommy 原创出品 © 版权所有

——— 一句话总结 · ONE-LINE VERDICT ———

偏好对齐有效
ALIGNMENT WORKS

- 证明可以从人类比较中学
- InstructGPT / ChatGPT 落地
- 范式拐点，真正可用

张力 TENSION

成本高昂
COSTLY TO RUN

- 四模型 + KL，显存吃紧
- PPO 调参经验门槛高
- 暴露偏好对齐的成本上限

下一站 → DPO 出场 · 砍掉两个模型

唐国梁Tommy | TGLTommy 原创出品 © 版权所有

2023

★ 第一个重要转折点 ★

DPO

Direct Preference Optimization

唐国梁Tommy | TGLTommy 原创出品 © 版权所有

◆ CORE INSIGHT ◆

核心洞察 · 数学等价

不一定要 显式训练 一个奖励模型 —— 奖励能藏进概率比

RLHF · 需要奖励模型

Reward Model 显式拟合

$$r_\phi(x, y) = \mathrm{RM}_\phi(x, y)$$

× 需要单独训练参数 ϕ

?

数学变换

≡

equivalent

DPO · 奖励藏进概率比

Implicit Reward via π_θ / π_ref

$$r_\theta(x, y) = \beta \log \frac{\pi_\theta(y \mid x)}{\pi_{\mathrm{ref}}(y \mid x)}$$

✓ 无需新参数 — 复用 π_θ 与 π_ref

唐国梁Tommy | TGLTommy 原创出品 © 版权所有

流程简化对比

复杂 RL 流水线 → 分类训练

RLHF 5 框流程 · 复杂 RL 流水线

数据收集
Prompts

→

偏好标注
Pairs

→

~~奖励模型训练~~
~~RM Train~~

→

~~PPO 强化学习~~
~~PPO Loop~~

→

对齐模型
Aligned LLM

✂ DPO 砍掉中间两步 训练流程立刻简单很多

DPO 极简流程 · 更像分类训练

偏好数据
Pairs

→

DPO 损失
Classification-like

→

唐国梁Tommy | TGLTommy 原创出品 © 版权所有

◆ 数学等价性 · 完整推导 ◆

DPO 为何能绕开奖励模型?

DPO LOSS · 分类训练形式

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

y_w = 偏好选项（chosen） · y_l = 拒绝选项（rejected）

唐国梁Tommy | TGLTommy 原创出品 © 版权所有

◆ DEMOCRATIZATION ◆

偏好对齐 · 走向平民化

BEFORE · RLHF

大团队工程
Big-Team Engineering

- 10+ 人 ML 团队 · 数月协调
- GPU 集群 · 多副本 RM + Actor + Critic
- PPO 调参 · KL 控制 · 奖励黑客

DPO → shift

AFTER · DPO

更多团队都能尝试
Accessible to More Teams

- 1~2 人小团队 · 数日完成
- 单卡或少卡 · 仅前向 + 反向
- 类分类训练 · 调参直觉化

💡 DPO 把偏好对齐从 「大团队工程」 变成 「更多团队都能尝试的训练方法」

唐国梁Tommy | TGLTommy 原创出品 © 版权所有

2024

方法多元化阶段

DPO 的成功 · 打开了一道闸门

唐国梁Tommy | TGLTommy 原创出品 © 版权所有

研究者们开始追问

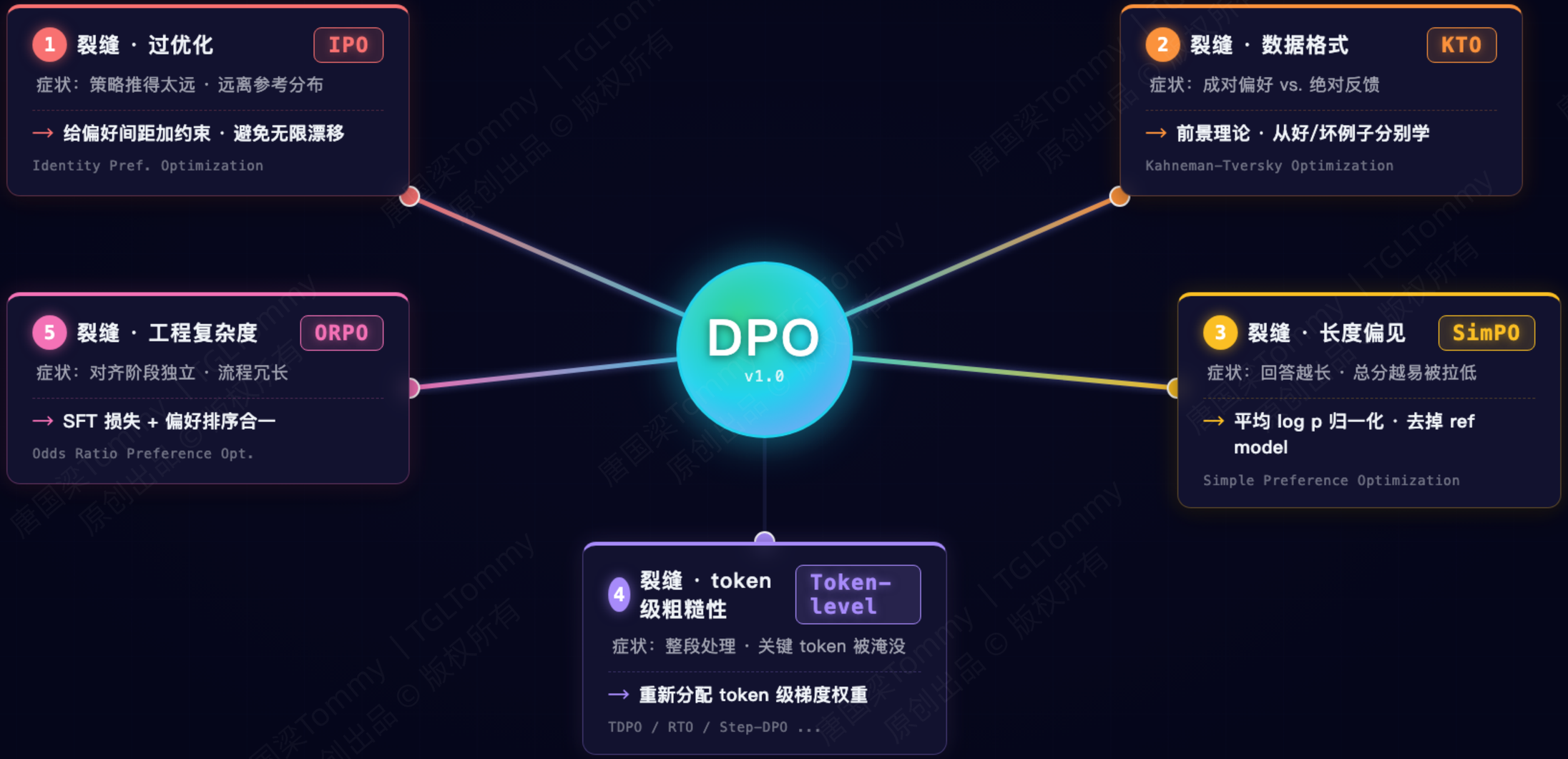


DPO 虽好，但它留下了哪些裂缝？

5 cracks · 5 patches

唐国梁Tommy | TGLTommy 原创出品 © 版权所有

★ DPO 家族 · 5 条裂缝 5 把扳手 ★



唐国梁Tommy | TGLTommy | 版权所有 © 原创出品



⚖ TRADE-OFF **每次修补都有效，每次也带来新权衡**

流程越极简 · **参考模型提供的稳定性锚点就越弱**

唐国梁Tommy | TGLTommy 原创出品 © 版权所有

2025

★ 第二个、更剧烈的转折 ★

RLVR

可验证奖励强化学习

唐国梁Tommy | TGLTommy 原创出品 © 版权所有

◆ ACRONYM · 全称展开 ◆

R · L · V · R

R

Reinforcement

强化

L

Learning

学习

V

Verifiable

可验证

R

Rewards

奖励

🎯 **可验证奖励强化学习**

唐国梁Tommy | TGLTommy 原创出品 © 版权所有

◆ CORE IDEA · 任务本身能自动验证 ◆

那为什么还要训练奖励模型?

结果可被程序验证 → 任务结果本身就可以变成奖励信号

数学题

Math · 标准答案

2 + 3 = 5 ✓ ✓



答案正确率（%）

代码任务

Code · 单元测试

def add(a,b): … ✓ Pass ✓



测试通过率（%）

搜索任务

Search · 事实核对

Q → A · 事实对齐 ✓ ✓



事实准确率（%）

唐国梁Tommy 原创出品 © 版权所有 | TGLTommy

◆ PARADIGM SHIFT · 奖励来源 ◆

这不是小优化 · 是范式变化

奖励信号的来源 · 三种范式对比

RLHF

贵

r = f(人类偏好)

Reward Model 拟合

成本	$$$ / sample
规模化	需大规模标注团队

DPO

隐式

r ≈ f(偏好数据)

隐含在概率比里

成本	$$ / sample
规模化	仍依赖偏好数据

RLVR

自动

r = f(任务验证器)

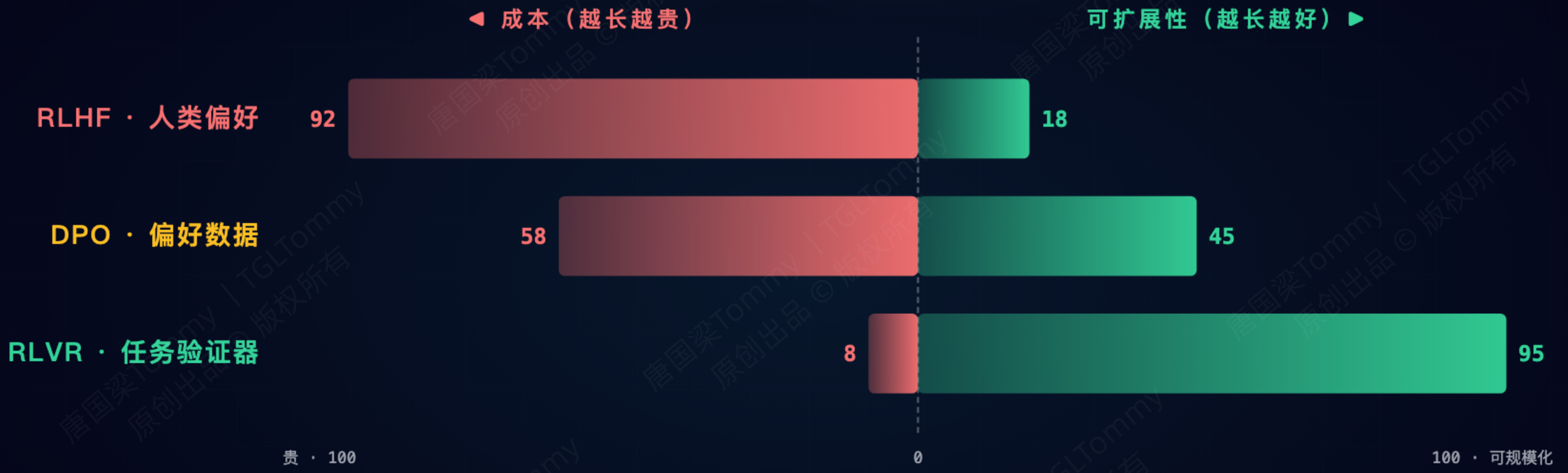
程序自动判定

成本	≈ 0 / sample
规模化	免费 · 可规模化

唐国梁Tommy | TGLTommy 原创出品 © 版权所有

◆ COST × SCALABILITY · 双维度对比 ◆

人类判断 ↘ 自动验证 ↗



🚀 奖励信号从 昂贵的人类判断 → 可规模化的自动验证

唐国梁Tommy | TGLTommy 原创出品 © 版权所有

◆ 2025 · POST-TRAINING ◆

DeepSeek-R1 ——— 可见样板

把可验证奖励路线，做成了实用范本

R1-Zero
纯 RL · 涌现推理

DeepSeek-R1
RL + 冷启动 + 多阶段 + 蒸馏

唐国梁Tommy | TGLTommy 原创出品 © 版权所有

PATH 01 · R1-ZERO

纯 RL 冷启动，也能诱发推理行为



可验证任务（数学 / 代码）+ 规则奖励 = 零 SFT 涌现长链推理

唐国梁Tommy | TGLTommy © 版权所有 原创出品

R1-ZERO 基础模型 → 纯 RL · 规则奖励 → R1-Zero ✦ 涌现推理

PATH 02 · DEEPSEEK-R1

RL + 冷启动 + 多阶段 + 蒸馏

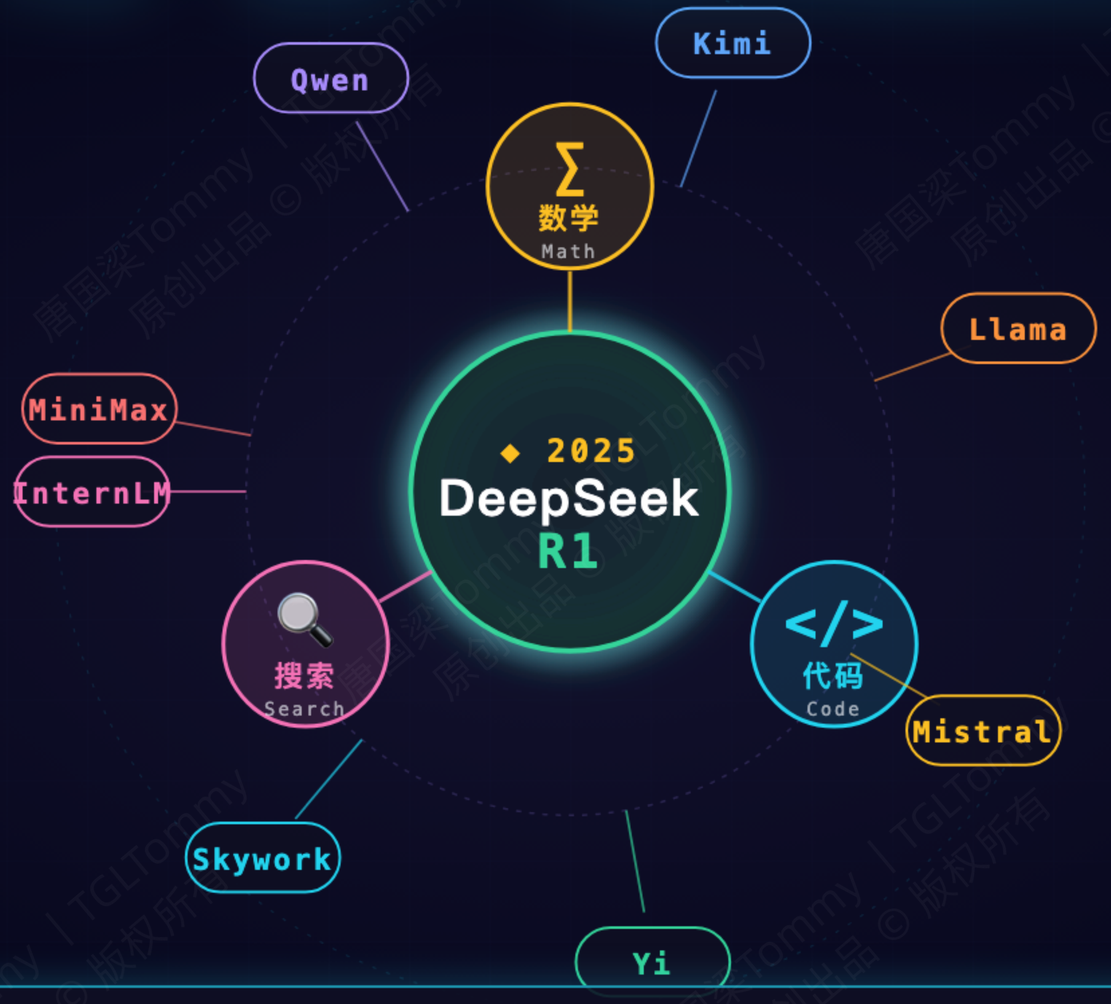


🛠️ RL + 冷启动 + 多阶段 + 蒸馏 ⇒ 实用推理模型

唐国梁Tommy | TGLTommy 原创出品 © 版权所有

OPEN-SOURCE FAST FOLLOWUP

数学 · 代码 · 搜索 — 可验证奖励路线全面铺开



开源社区迅速跟进 —— 可验证奖励 = 推理模型的新通用配方

唐国梁Tommy | TGLTommy 原创出品 © 版权所有

✦ 2025 年之后 · ANNO DOMINI POST-TRAINING ✦

RL 后训练

= 推理模型能力提升的 核心关键词

#可验证奖励 #RL 路线 #多阶段配方 #开源生态 #推理涌现

唐国梁Tommy | TGLTommy 原创出品 © 版权所有

◆ OPEN QUESTION ◆

深层争议

RL 到底是「教」新能力，还是「引出」已有能力?

TEACH · 教论

RL 教模型
新能力

?

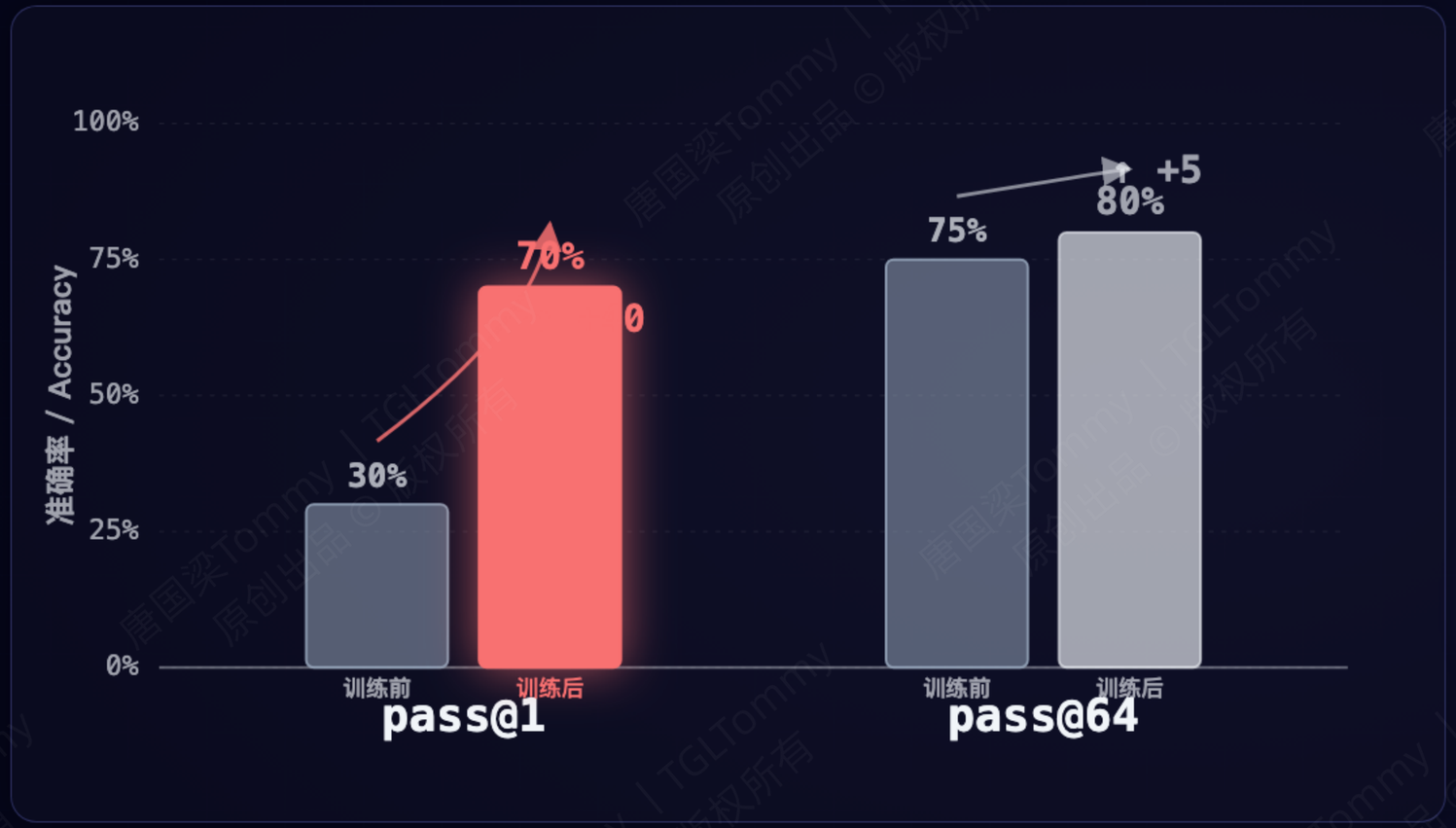
ELICIT · 引出论

RL 引出预训练
已有能力

唐国梁Tommy | TGLTommy 原创出品 © 版权所有

EVIDENCE · 引出论

答案早就在采样空间里



▲ pass@1 大幅提升

RL 教会模型「第一次就选对路径」

≈ pass@k 几乎不变

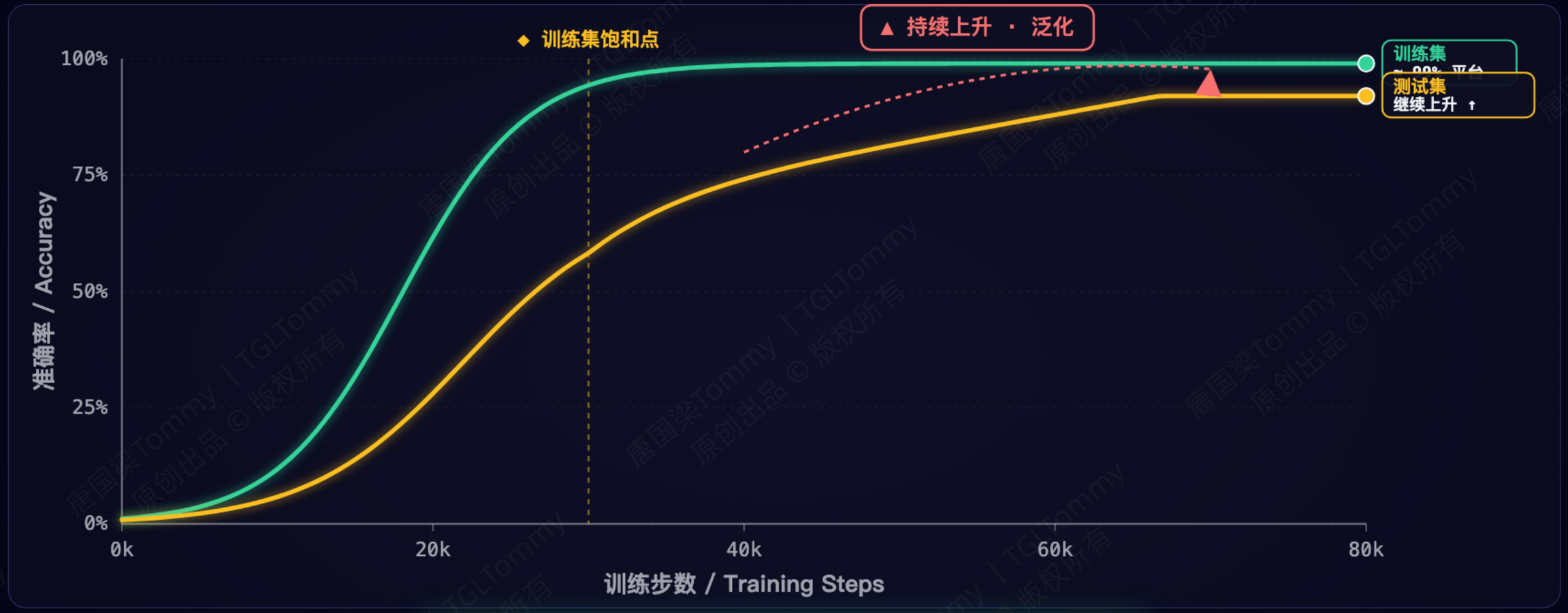
正确答案本就在模型采样空间内

→ 结论：支持「引出论」

唐国梁Tommy | TGLTommy | 版权所有 © 原创出品

EVIDENCE · 教论

后饱和泛化现象



→ 结论：支持「教论」 — RL 可能带来真正的泛化

唐国梁Tommy | TGLTommy 原创出品 © 版权所有

◆ TO BE CONTINUED ◆

争议未决

教 vs 引出 → EPISODE 03 第 3 集 · 深入讨论

我们将专门拆解：什么样的设置下 RL 真的「教」了新东西？

唐国梁Tommy | TGLTommy 原创出品 © 版权所有

◆ 推理 RL · 必须单独提 ◆

GRPO

Group Relative Policy Optimization

★ DeepSeek 系列 · DeepSeekMath / R1

唐国梁Tommy | TGLTommy 原创出品 © 版权所有

—— PPO · 重资产 ——

PPO 需要价值网络估计优势函数



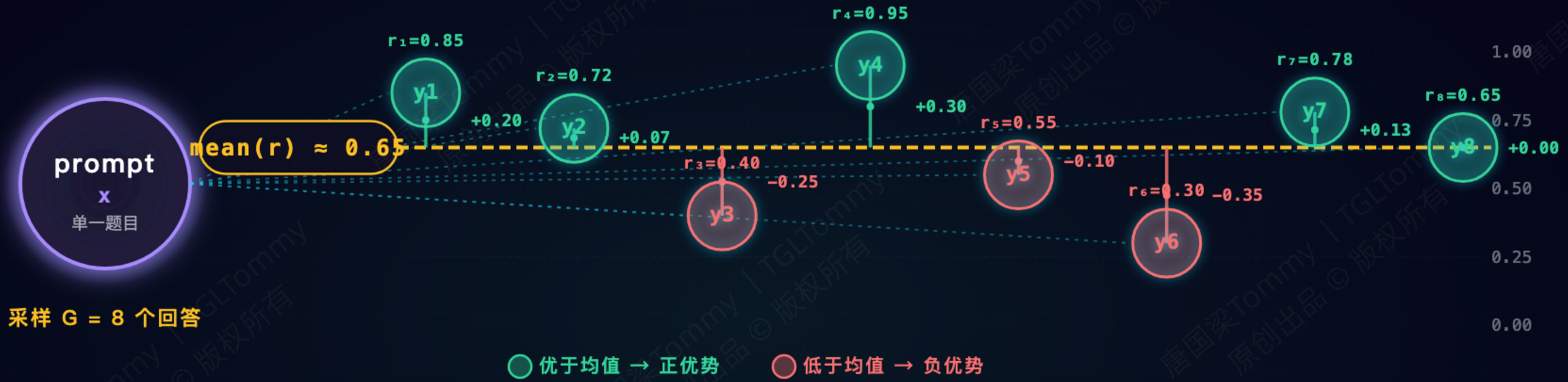
$$A^{\text{PPO}}(s,a) = Q(s,a) - V_\phi(s)$$

💸 训练成本高

唐国梁Tommy | TGITommy © 版权所有 原创出品

GRPO · 更轻

组内平均 = 基线 · 砍掉 critic



相对优势 $$A_i^{\mathrm{GRPO}} = \frac{r_i - \mathrm{mean}(r_1, \ldots, r_G)}{\mathrm{std}(r_1, \ldots, r_G)}$$

唐国梁Tommy | TGLTommy 原创出品 © 版权所有

ARCHITECTURE DIFF

砍掉价值网络 — 显存 ↓ 计算 ↓ 效果 保持

PPO

4 模型架构

- π_θ 策略 (Actor)
- ~~V_φ 价值 (Critic)~~ × removed in GRPO
- r_φ 奖励模型 (RM)
- π_{ref} 参考策略 (Ref)

优势: $A = Q - V_\varphi$ · 价值网络与策略同规模 → 显存翻倍

GRPO

3 模型架构

- π_θ 策略 (Actor)
- $r(\cdot)$ 奖励 / 验证器
- π_{ref} 参考策略 (Ref)
- ✓ 组内均值 = 基线 免费 · 无需训练

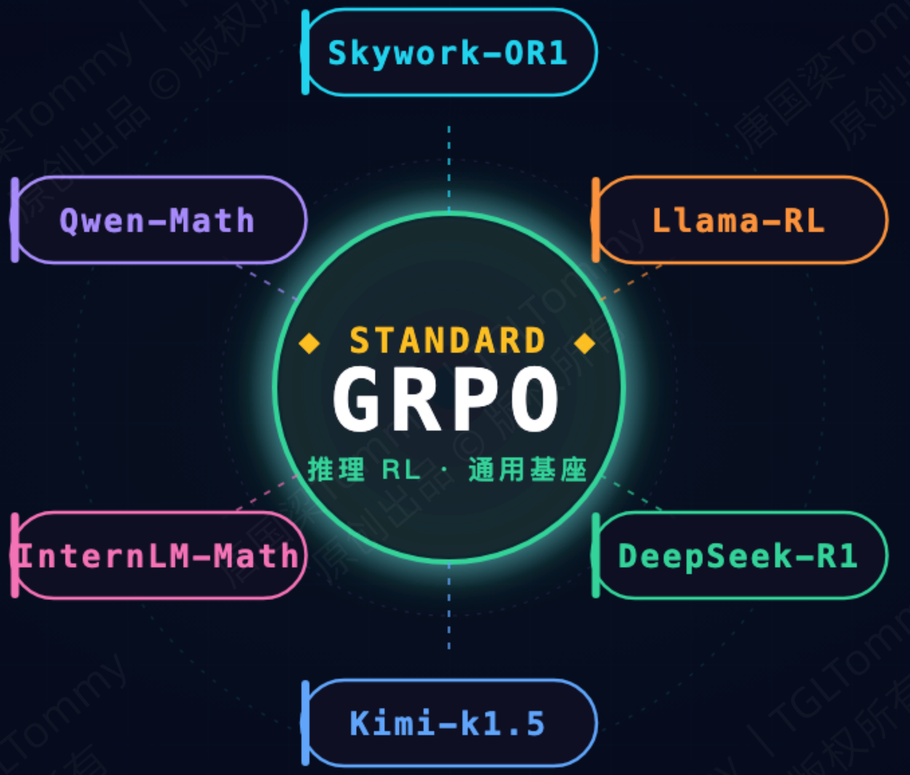
优势: $A_i = (r_i - \text{mean}) / \text{std}$ · 仅需采样 → 显存大降

💾 显存 ↓ ⚡ 计算 ↓ 🎯 推理任务效果 保持

唐国梁Tommy | TGITommy © 版权所有 原创出品

★ DeepSeek-R1 之后 · POST 2025-01 ★

推理模型训练 · 最常被讨论的优化算法之一



推理模型训练 —— GRPO 已成事实标准

唐国梁Tommy | TGLTommy 原创出品 © 版权所有



算法工具箱

四种工具 · 各有所长

RLHF + PPO · DPO · GRPO · RLVR

唐国梁Tommy丨TGLTommy 原创出品 © 版权所有

◆ 四类算法 · 一张地图 ◆

RLHF + PPO

01

Reinforcement Learning from Human Feedback

最传统 · 最重 · 最灵活

适合场景

- 奖励信号复杂
- 质量要求很高
- 资源比较充裕

`+ 灵活` `− 重量级` `− 高资源`

DPO

Direct Preference Optimization

02

砍掉显式奖励模型 · 简化偏好优化

适合场景

- 已有偏好数据
- 希望稳定低成本
- 对齐任务

`+ 简单` `+ 低成本` `· 需偏好数据`

GRPO

Group Relative Policy Optimization

03

砍掉价值网络 · 群组采样估优势

适合场景

- 数学 · 代码 · 搜索
- 可批量采样比较
- 推理任务

`+ 推理强` `+ 省显存` `· 群组采样`

RLVR

Reinforcement Learning with Verifiable Rewards

04

不是优化算法 · 是一种奖励来源

适合场景

- 任务能自动验证
- 数学题 · 单元测试
- 形式化判定

`· 奖励来源` `+ 无需人标` `− 需可验证`

唐国梁Tommy | TGLTommy | 版权所有 © 原创出品

任务特征 → 推荐工具



可验证任务
数学 · 代码 · 形式化
优先考虑 → **RLVR + GRPO**

离线偏好数据充足
已有 chosen / rejected 对
很有吸引力 → **DPO**

复杂奖励 + 精细控制
需要奖励塑形 · 多目标
仍有价值 → **RLHF + PPO**

唐国梁Tommy | TGLTommy | 版权所有 © 原创出品



课程地图

接下来 10 章 · 我们要讲什么

Ch.2–3 基础 | Ch.4–5 核心 | Ch.6 Agentic | Ch.7 算法 | Ch.8–10 实战 | Ch.11 论文

唐国梁Tommy | TGLTommy 原创出品 © 版权所有

Ch.2–3

RL 基础

- MDP
- 价值函数
- 贝尔曼
- 策略梯度

Ch.4

RLHF & RLAIF

- 偏好收集
- 奖励建模
- PPO 优化

Ch.5

奖励函数设计

- RLVR · GenRM · PRM
- 奖励黑客 · 头号敌人

理论奠基 · 第 2–5 章

唐国梁Tommy | TGLTommy 原创出品 © 版权所有

Ch.2–3

RL 基础

- MDP
- 价值函数
- 贝尔曼
- 策略梯度

Ch.4

RLHF & RLAIF

- 偏好收集
- 奖励建模
- PPO 优化

Ch.5

奖励函数设计

- RLVR · GenRM · PRM
- 奖励黑客 · 头号敌人

Ch.6

Agentic RL

- 信用分配难题
- 长程决策

Ch.7

算法实战

- PPO 四模型
- GRPO 群组采样
- DPO 数学等价

Agentic 与算法实战 · 第 6–7 章

唐国梁Tommy | TGLTommy 原创出品 © 版权所有

Ch.2–3
RL 基础
- MDP
- 价值函数
- 贝尔曼
- 策略梯度

Ch.4
RLHF & RLAIF
- 偏好收集
- 奖励建模
- PPO 优化

Ch.5
奖励函数设计
- RLVR · GenRM · PRM
- 奖励黑客 · 头号敌人

Ch.6
Agentic RL
- 信用分配难题
- 长程决策

Ch.7
算法实战
- PPO 四模型
- GRPO 群组采样
- DPO 数学等价

Ch.8
DeepAnalyze 实战
- 端到端项目
- 分析智能体

Ch.9
OpenClaw-RL
- 训练系统
- 工业级架构

Ch.10
Agent Skill
- 技能注入
- 自进化 Agent

工业落地与论文 · 第 8–11 章

唐国梁Tommy | TGLTommy © 版权所有 原创出品

Ch.2–3
RL 基础
- MDP
- 价值函数
- 贝尔曼
- 策略梯度

Ch.4
RLHF & RLAIF
- 偏好收集
- 奖励建模
- PPO 优化

Ch.5
奖励函数设计
- RLVR · GenRM · PRM
- 奖励黑客 · 头号敌人

Ch.6
Agentic RL
- 信用分配难题
- 长程决策

Ch.7
算法实战
- PPO 四模型
- GRPO 群组采样
- DPO 数学等价

Ch.8
DeepAnalyze 实战
- 端到端项目
- 分析智能体

Ch.9
OpenClaw-RL
- 训练系统
- 工业级架构

Ch.10
Agent Skill
- 技能注入
- 自进化 Agent

Ch.11
论文精讲
- 前沿追踪
- 代码复现

理论 → 算法 → 实战 → 论文 · 完整闭环

唐国梁Tommy | TGLTommy 原创出品 © 版权所有

EPISODE 01 · RECAP

本集回顾

对齐演化简史 · 一张地图

2017 → 2022 → 2023 → 2025

唐国梁Tommy | TGLTommy 原创出品 © 版权所有

ALIGNMENT EVOLUTION

对齐演化时间轴 · 四个里程碑

2017
起源

1

偏好学习

Preference

Deep RL · Pref

Christiano et al.

2022
主舞台

2

InstructGPT

RLHF + PPO

Reward + Policy

ChatGPT 横空出世

2023
效率革命

3

DPO

DPO

Direct Pref Opt.

免奖励模型

2025
推理 RL

4

RLVR + GRPO

RLVR · GRPO

可验证奖励

DeepSeek-R1 引爆

⚖ 各有适用场景 · 各有代价

唐国梁Tommy | TGLTommy 原创出品 © 版权所有

BEYOND THE SAFE ZONE

地图画好了，但还有一个问题

当 RL 走出"可自动判对错"的安全区 ……



会遇到什么新的挑战？

唐国梁Tommy | TGLTommy 原创出品 © 版权所有

◆ TO BE CONTINUED ◆

下一集

RL 的边界



唐国梁Tommy丨TGLTommy 原创出品 © 版权所有

版权声明

© 2026 唐国梁（Tommy）/ 熵智未来（深圳）科技有限公司。保留所有权利。

本课程及其配套资料，包括但不限于课程视频、PDF 文档、课件、代码、图表、示例项目、知识库文档及其他学习材料，仅供购买本课程的学习者本人学习使用。除法律法规另有规定或经权利人书面许可外，不得复制、录屏、翻录、截图传播、转载、上传、分发、转售、改编、翻译、节选、汇编，或用于商业培训、企业内训、知识付费产品、AI 模型训练、微调、数据集构建、商业化知识库建设及其他商业用途。

本课程及配套资料中涉及的第三方开源软件、模型、代码、数据集、商标、图片或其他材料，相关权利归原权利人所有，并遵循其相应许可证或授权条款。

对于侵犯本课程及配套资料相关权益的行为，权利人将依法采取包括但不限于投诉、要求下架、要求停止侵害、索赔、追究法律责任等措施；构成违法犯罪的，将依法追究相应责任。

侵权举报 / 授权合作：legal@tgltommyai.com