大模型训练阶段的关键方法简介

在大语言模型（Large Language Models，LLMs）的训练流程中，这些方法通常按顺序应用：从预训练（PT）开始，到监督微调（SFT），再到使用奖励模型（Reward Model）和强化学习或偏好优化（如PPO、DPO、KTO）进行对齐。这些方法帮助模型从海量数据中学习基础知识，并逐步与人类偏好对齐，提高安全性、帮助性和任务适应性。下面我将分别解释每个术语的含义、核心作用、在训练阶段的位置，以及典型应用方向。信息基于当前（2025年）的LLM训练实践。

方法	全称与简介	在训练阶段的作用	应用方向
PT	Pre-Training (预训练)：模型在海量无标签文本数据上从零训练，学习语言模式、语法和世界知识。通常使用自监督学习，如预测下一个词（Next Token Prediction）。这是LLM训练的起点，构建基础参数。	作为初始阶段，建立模型的核心表示和泛化能力。消耗大量计算资源，但无需人工标签。后续方法（如SFT）在此基础上微调。	基础模型构建，如GPT系列的初始训练。用于通用语言理解、知识表示学习，常用大规模数据处理（如万亿级token）。在多态模型中扩展到图像-文本对齐。 youtube.com
SFT	Supervised Fine-Tuning	在PT后，作为中间阶段桥接基础模型到应用模型，提升任务特定性能，减少无	指令跟随和任务适应，如对话系统、摘要成。广泛用于ChatGPT-like模型的初步

	Fine-Tuning (监督微调)：使用有标签的高质量数据集（如指令-响应对）微调预训练模型，使其适应特定任务。数据通常由人类或合成生成，优化损失函数如交叉熵。	应用模型。提升任务特定性能，减少泛化问题，但可能引入过拟合。常作为PPO或DPO的前置步骤，确保数据分布一致。 huggingface.co +更多 2	成。广泛用于ChatGPT-like模型的初步对齐，也在多模态LLM中微调视觉-语言任务。 blog.gopenai.com cs224r.stanford.e
Reward	Reward Model (奖励模型)：一个分类器模型，使用人类偏好数据（如成对比较）训练，评估生成的响应质量（如帮助性、无害性）。它输出分数，作为RL信号。	在RLHF（Reinforcement Learning from Human Feedback）中，作为中间模型提供反馈。桥接人类偏好到优化过程，避免直接使用稀疏奖励。常与PPO结合使用。 arxiv.org	对齐人类偏好，如评估响应安全性。应用RLHF管道中，例如InstructGPT的奖励阶段，也用于自举训练（如DPO后的隐励）。 openreview.net proceedings.iclr.cc
PPO	Proximal Policy	在SFT后，作为RLHF的核心优化阶段。使用代理损失函数迭代更新模型，使其	人类反馈强化学习，如ChatGPT的对齐。用于复杂任务如对话生成、代码编写、

	Policy Optimization (近端策略优化)：一种强化学习算法，使用奖励模型指导策略模型更新。通过KL散度约束限制更新幅度，确保训练稳定。涉及优势函数和价值模型。	使用代理损失函数迭代更新模型，以最大化累积奖励。防止训练崩溃，但计算密集。 arxiv.org bobrupakroy.medium.com	用于复杂任务如对话生成、代码辅写，但率较低，常被DPO替代。 blog.gopenai.com youtube.com
DPO	Direct Preference Optimization (直接偏好优化)：从偏好数据对直接优化模型，无需显式奖励模型。通过重新参数化，将问题转化为分类损失，简化RLHF。	在SFT后，作为高效对齐阶段。直接最大化偏好响应概率，同时最小化与参考模型的偏差。计算成本低，稳定性高。 arxiv.org +更多 6	偏好对齐，如提升模型帮助性和无害性。用于开源模型如LLaMA，常与SFT结合，于摘要、翻译等任务。 blog.gopenai.com +更多 5
KTO	Kahneman-Tversky	在SFT后，作为对齐阶段的替代或补充。利用非成对标签提升泛化，减少数据收	人类偏好对齐，如使用单标签数据训练。用于RLHF变体，如多模态模型或开源工

Optimization (卡内曼-特沃斯基优化)：一种偏好优化方法，基于人类决策理论，使用期望/非期望示例（而非成对）优化模型。避免了DPO的成对数据需求。	利用非成对标签提升泛化，减少数据收集成本。适用于数据稀缺场景。 github.com docs.unsloth.ai	用于RLHF变体，如多模态模型或开源工链（如Unsloth），适合情感分析或泛化任务。 github.com docs.unsloth.ai