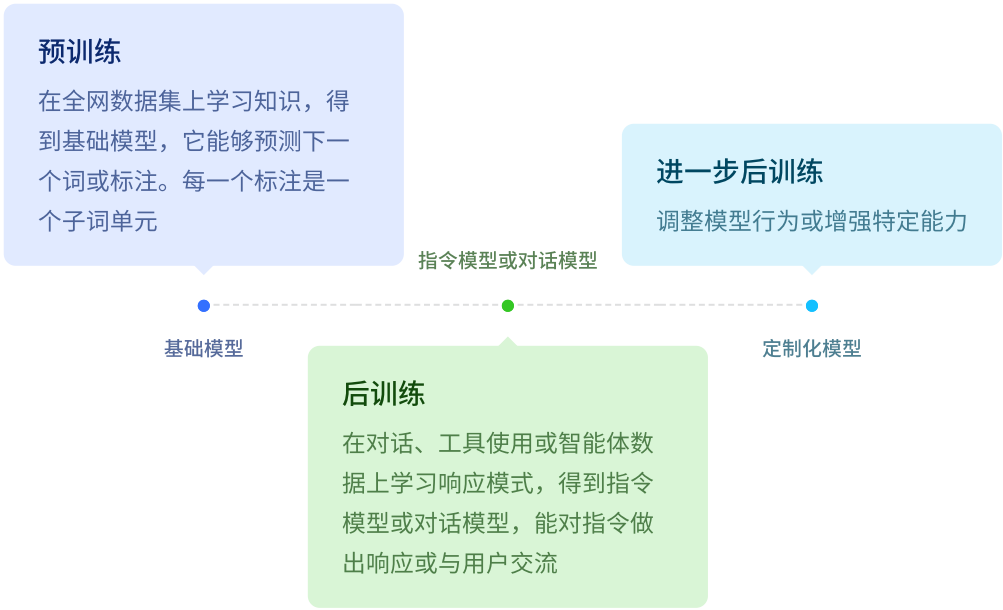
Post-Training of LLMs



1.预训练：无监督学习，从大规模无标注文本语料中提取超2万亿个标注

当输入"我喜欢猫"这样的句子时，模型会基于前面所有标记来最小化每个标记的**负对数概率**：

首先最小化"我"的**负对数概率**——>"我"时"喜欢"的**负对数似然**——>"我喜欢"时"猫"的**概率**

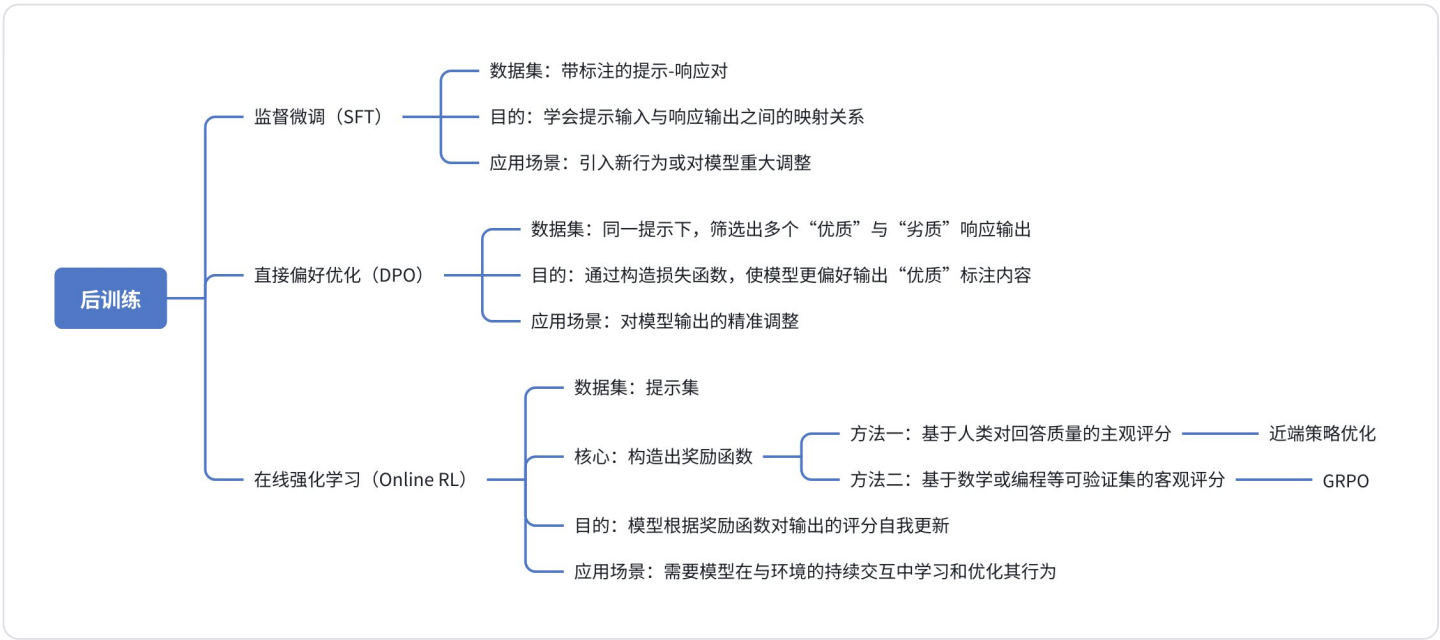
通过这种方式，模型被训练成能根据已见标记预测下一个标记，如图所示

$$\text{“I like cats”}$$

$$\min_{\pi} -\log \pi(\mathrm{I}) - \log \pi(\mathrm{like} \mid \mathrm{I}) - \log \pi(\mathrm{cats} \mid \mathrm{I\ like})$$

2.后训练与预训练的区别：预训练是大语言模型的训练主体，在数万亿文本标注上训练；后训练是针对具体任务的训练，利用较小数据集快速实现目的

3.后训练的主要方法：



（1）监督微调（Supervised Fine-tuning）：监督学习/模仿学习。提示：给模型的指令；响应：模型应有的回答。仅需1000至10亿个标记。训练损失关键：仅对响应标记训练，不涉及提示标记。

$$\min_{\pi} -\log \pi\left(\text{Response} \mid \text{Prompt}\right)$$

（2）直接偏好优化（Direct Preference Optimization）：仅需1000至10亿个标记，采用更复杂的损失函数。

$$\min_{\pi} -\log \sigma\left(\beta\left(\log \frac{\pi(\text{Good R} \mid \text{Prompt})}{\pi_{\text{ref}}(\text{Good R} \mid \text{Prompt})} - \log \frac{\pi(\text{Bad R} \mid \text{Prompt})}{\pi_{\text{ref}}(\text{Bad R} \mid \text{Prompt})}\right)\right)$$

（3）在线强化学习（Online Reinforcement Learning）：提示——>语言模型生成响应——>奖励函数对该响应评分——>利用该信号更新模型。1,000至1,000万（或更多）个提示。目标是通过模型自身生成的响应来最大化奖励值

$$\max_{\pi} \ \text{Reward}(\text{Prompt}, \text{Response}(\pi))$$

4.后训练成功的三大要素

Post-training Requires Getting 3 Elements Right

Data & algorithm co-design

- SFT
- DPO
- Reinforce / RLOO
- GRPO
- PPO
- …

Reliable and efficient library

- Huggingface TRL
- OpenRLHF
- veRL
- Nemo RL

Appropriate evaluation suite

(An Incomplete List of) Popular LLM Evals

Human Preferences for chat	**Chatbot Arena**
LLM as a judge for chat	Alpaca Eval MT Bench **Arena Hard V1 / V2**
Static Benchmarks for Instruct LLM	**LivecodeBench** **AIME 2024 / 2025** GPQA MMLU Pro IFEval
Function Calling & Agent	BFCL V2 / V3 NexusBench V1 / V2 **TauBench** **ToolSandbox**

It's easy to improve any one of the benchmarks.

It's much harder to improve **without degrading other domains**.

提升单一基准成绩容易，但要在不损害其他领域能力的前提下改进特定行为更难

5.针对不同需求的解决方案，明确后训练的应用场景

Do you really need post-training?

Use Cases	Methods	Characteristics
Follow a few instructions (do not discuss XXX)	Prompting	Simple yet brittle: models may not always follow all instructions
Query real-time database or knowledgebase	Retrieval- Augmented Generation (RAG) or Search	Adapt to rapidly-changing knowledgebase
Create a medical LLM / Cybersecurity LLM	Continual Pre-training + Post-training	Inject large-scale domain knowledge (>1B tokens) not seen during pre-training
Follow 20+ instructions tightly; Improve targeted capabilities (“Create a strong SQL / function calling / reasoning model”)	Post-training	Reliably change model behavior & improve targeted capabilities; May degrade other capabilities if not done right