大语言模型的适配与使用

从学习信号、参数和上下文到结果证据

学习目标

- **学习信号与SFT**：解释条件token、有效监督位置与平均损失，区分训练真实前缀与生成历史。
- **参数更新与上下文**：比较全量微调、LoRA与ICL改变的对象；解释低秩形状、缩放及参数计数条件。
- **偏好优化**：区分示范与偏好信号，解释奖励模型、RLHF代表流程与DPO。
- **结果依据及外部接口**：区分可见推理、答案与证据，说明稳定性、成本及RAG和工具的作用。

预训练能力与具体任务需求

预训练模型已学习根据上下文预测后续token。实际应用还需要模型按任务要求回答，并让结果具有可检查的依据。



本讲先把期望回答变成学习信号，再区分参数、输入与偏好，最后检查具体结果。各方法可以组合，不要求按授课顺序依次使用。

1 指令监督与参数适配

怎样把期望回答变成监督，又允许哪些参数改变？

期望回答提供学习信号

指令微调（Supervised Fine-Tuning，SFT）用任务输入与期望回答提供监督，以目标回答的预测损失训练可适配参数。

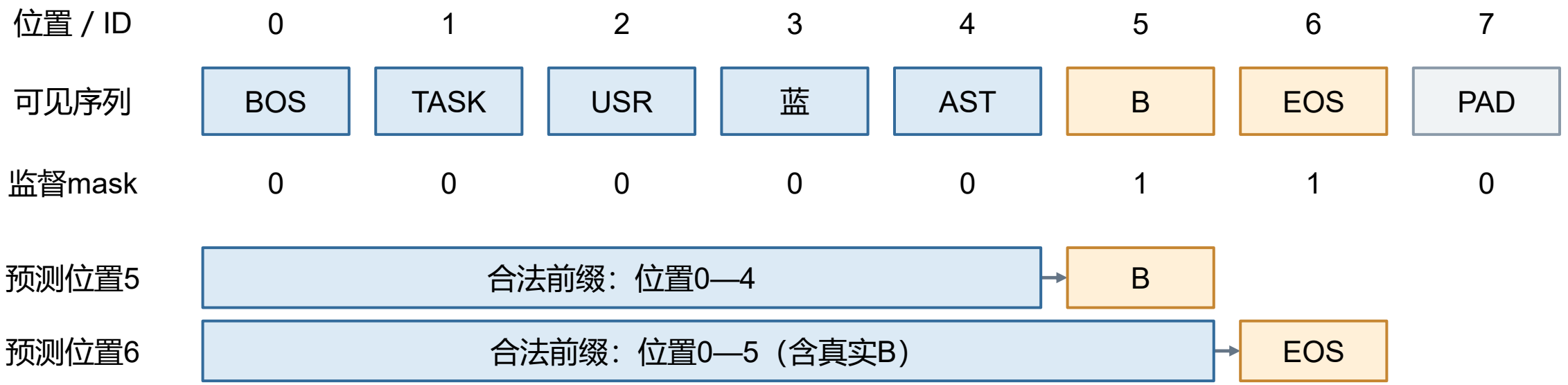
构造任务	数据内容
任务要求	根据色词输出约定代码，只输出代码
训练约定	红对应R；蓝对应B
当前输入	蓝
期望回答	B

训练约定用于构造标签，不自动代表它会完整写入每次推理Prompt。对话角色说明谁说了什么；监督选择决定哪些预测误差计入损失。

基础模型是适配起点的预训练模型；**指令模型**经过面向指令响应的训练适配。它们的区别涉及训练与用途，不要求更换架构，也不保证每次回答正确。

条件token与监督位置

人工序列为（BOS，TASK，USR，蓝，AST，B，EOS，PAD）。TASK、USR、AST是本例抽象符号，每个符号算一个token；本例ID恰与位置相等，不是实际模型的一般规则。



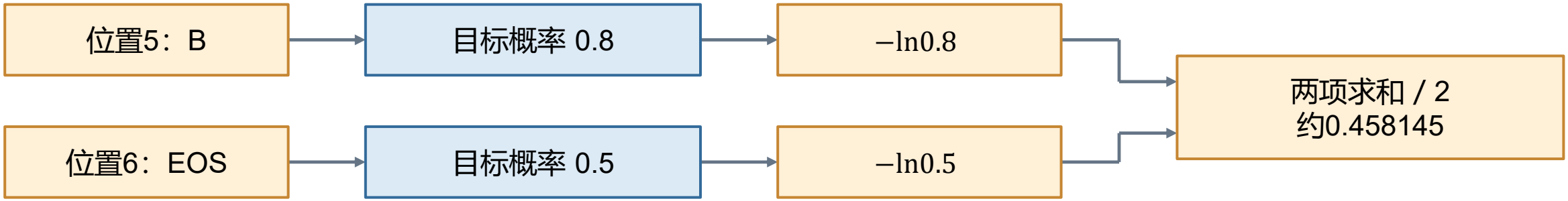
参与条件预测的token，不一定参与损失计算。本例有效监督位置$\mathcal{I} = \{5,6\}$：只对B和EOS计损失。条件token可见但不计本例损失；PAD既不是有效目标，也不是有效历史。预测均不能读取未来目标。

SFT损失计算

监督标签是真实目标token。**负对数似然损失**使低概率的真实目标产生较大代价。本例采用有效token平均与自然对数：

$$L_{\mathrm{SFT}}(\theta) = -\frac{1}{|\mathcal{I}|}\sum_{t\in\mathcal{I}} \ln p_\theta(x_t \mid x_{0:t-1}).$$

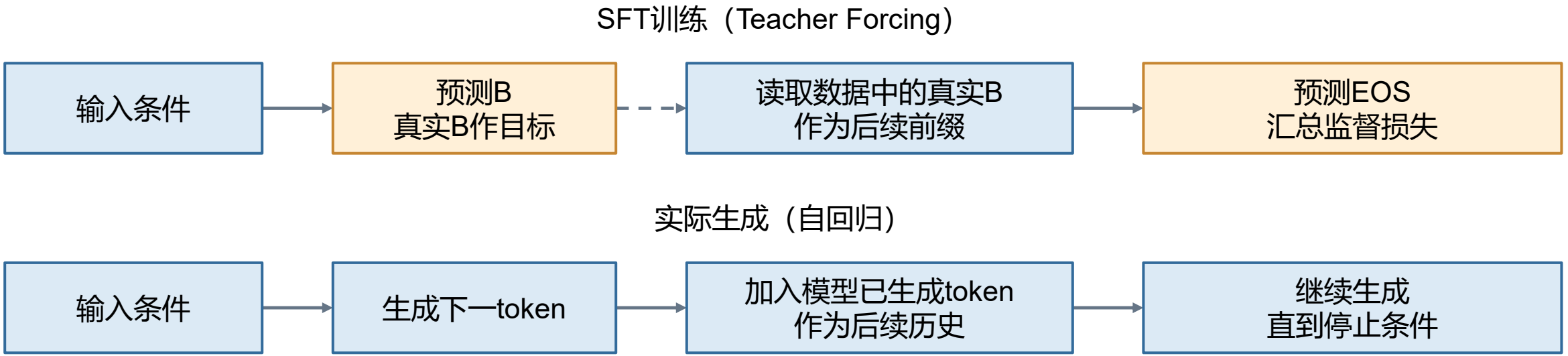
θ为模型参数，x_t为真实目标，$x_{0:t-1}$为合法前缀；分母为有效目标数，不是序列长度。



以上概率为构造值，$L = -(\ln 0.8 + \ln 0.5)/2 = -\ln 0.4/2 \approx 0.458145$。条件与PAD不贡献此平均。合并多样本token平均时，分母使用合并后的有效目标数；按样本平均是另一种汇总方式。

训练与生成的历史差异

两者都只读取合法因果前缀，差别在于后续回答历史来自哪里。



上轨是逻辑依赖：后续前缀取自训练数据，不使用模型先前猜出的B；通常可并行计算多个位置的预测。下轨用已生成输出，可能产生B以外的token；通常推理不为当前任务执行梯度更新。

条件、监督位置、参数更新范围是三个维度。损失已确定，接下来需要决定允许哪些参数更新。

SFT目标与参数更新范围

SFT已经规定希望预测哪些回答；降低同一个损失时，还要决定允许哪些变量改变。

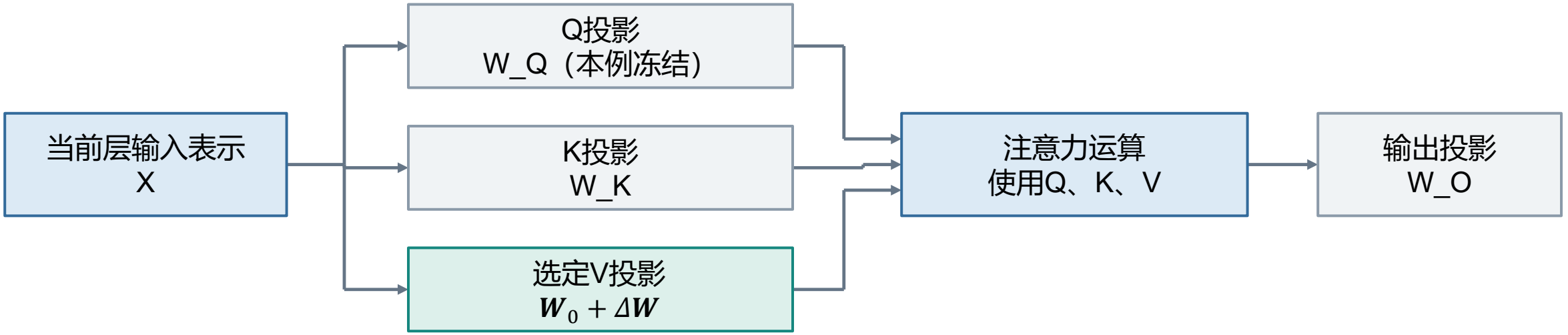
全量微调允许更新基座全部参数；可训练表示允许更新，不表示每个数值每步都必然改变。**LoRA**（Low-Rank Adaptation）冻结基座，以新增低秩因子表示选定线性映射的权重增量。

同一个SFT损失	全量微调	本讲LoRA设置
学习信号	指令输入与期望回答	同样的指令输入与期望回答
可训练对象	基座全部参数	选定映射的新增低秩因子
冻结对象	此设置不冻结基座	基座及未声明适配的部分

SFT决定学习信号，LoRA决定更新范围与参数表示。两者可以组合；LoRA本身不规定另一种损失。

Transformer中的LoRA适配位置

Transformer含有许多线性映射。本单元连续放大同一层的V投影：原权重记为$\boldsymbol{W}_0$，适配后为$\boldsymbol{W}_0 + \boldsymbol{\Delta W}$。

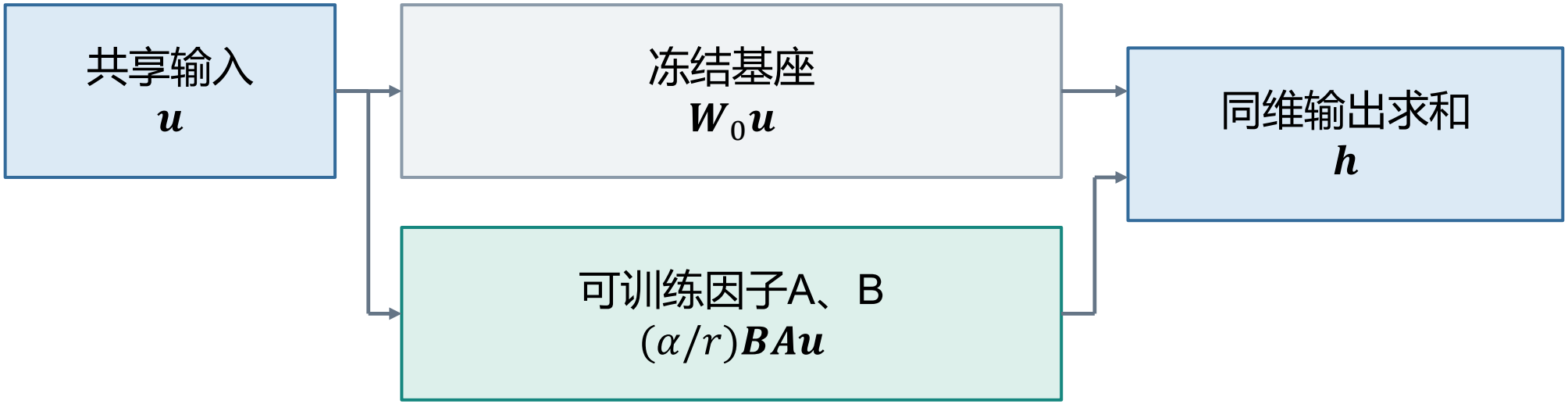


连续放大V投影：u是当前token的输入列向量，h是其输出。出处：Hu等，LoRA，2021，§4.1—4.2

绿色表示含可训练增量，原权重仍冻结。V仅是本例选择，Q等其他映射也可适配，不要求修改全部投影/层。图省略分头、残差、归一化与前馈层；后两页沿用同一个$\boldsymbol{W}_0, \boldsymbol{u}, \boldsymbol{h}$。

冻结基座与低秩增量

放大前页同一个V投影。两路共享输入$\boldsymbol{u}$，冻结的$\boldsymbol{W}_0$与新增低秩分支各给出同维结果，再逐分量相加。

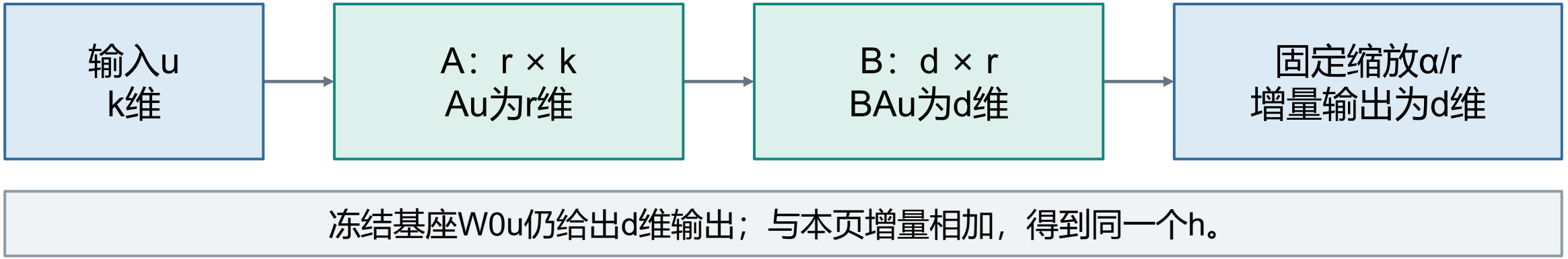


训练只更新声明的$\boldsymbol{A}, \boldsymbol{B}$，$\boldsymbol{W}_0$保持不变。**冻结不等于不参与前向计算**：基座仍提供输出，其计算也可参与对可训练部分的梯度计算。实线表示前向依赖，更新对象由“可训练”标签标明。

出处：Hu等，LoRA，2021，§4.1；本图延续前页选定V投影。

低秩因子的矩阵计算

延续同一个V投影：$\boldsymbol{u} \in \mathbb{R}^k$，冻结基座$\boldsymbol{W}_0 \in \mathbb{R}^{d\times k}$。省略偏置，先沿绿色增量分支检查形状：



$$\boldsymbol{A} \in \mathbb{R}^{r\times k}, \boldsymbol{B} \in \mathbb{R}^{d\times r}, \Delta\boldsymbol{W} = \frac{\alpha}{r}\boldsymbol{BA}. \boldsymbol{h} = \boldsymbol{W}_0\boldsymbol{u} + \frac{\alpha}{r}\boldsymbol{BAu} = (\boldsymbol{W}_0 + \Delta\boldsymbol{W})\boldsymbol{u} \in \mathbb{R}^d.$$

A降到r维，B恢复到d维，与基座d维输出相加。r是较小的秩参数，增量秩至多为r；$\alpha > 0$为固定缩放。仅A、B可训练，$\Delta\boldsymbol{W}$的定义已包含缩放。

可训练参数的计数条件

取构造维度$k = 8, d = 6, r = 2, \alpha = 4$。仅适配一个线性权重，基座全部冻结，无额外可训练头。

对象	形状或计算	参数个数
直接训练该权重$\boldsymbol{W}_0$	6×8	48
因子$\boldsymbol{A}$	2×8	16
因子$\boldsymbol{B}$	6×2	12
LoRA可训练因子合计	$16 + 12$	28

形状链为$8 \to 2 \to 6$，$\alpha/r = 2$；基座48个权重仍存在，可训练数减少20。一般比较为$r(k + d)$与dk，前者更少需满足$r(k + d) < dk$。

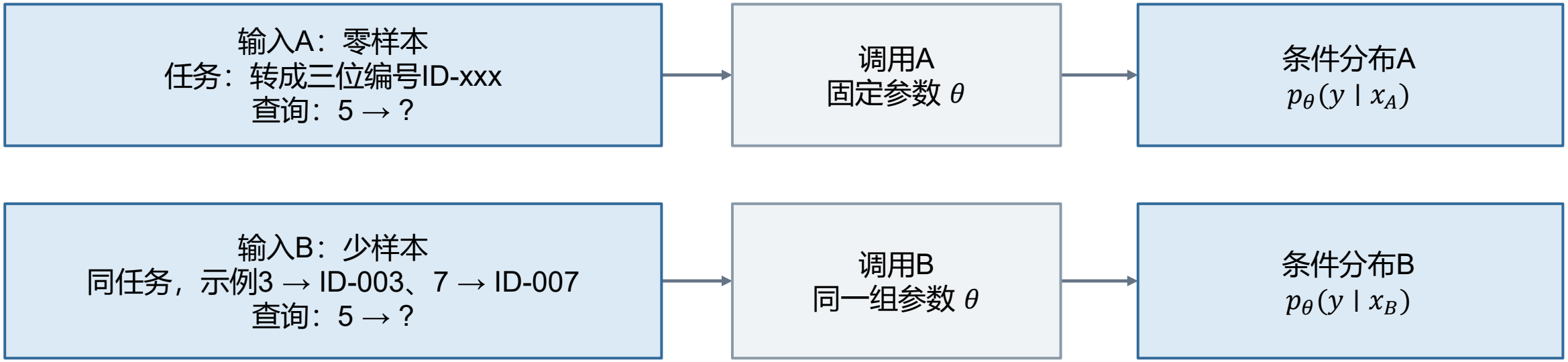
这是单映射教学计数，未测量整模型。参数减少不能直接推出同比显存下降、训练加速、推理成本下降或效果提升。

2 上下文与偏好信号

权重不变时输入怎样起作用？比较回答又提供什么信号？

固定权重下的上下文学习

全量微调与LoRA改变可训练参数；使用时也可把任务信息写进当前输入。**上下文学习**（In-Context Learning，ICL）通过任务说明或示例影响当前预测，通常不执行当前梯度更新。

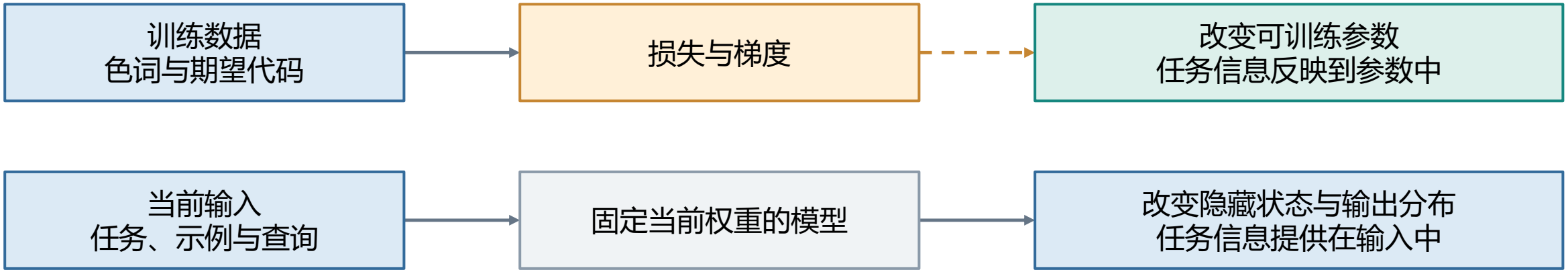


两次调用共享θ：条件不同，输出分布可能不同（不是两套参数）

零样本不含任务示范，少样本提供少量示范。按本例规则期望输出ID-005；图示未运行模型，也未比较实际分布。两路共享同一组权重，不是两套独立参数；不同条件不保证分布必然不同，也不保证示例一定改善结果。

输入变化与参数更新的区别

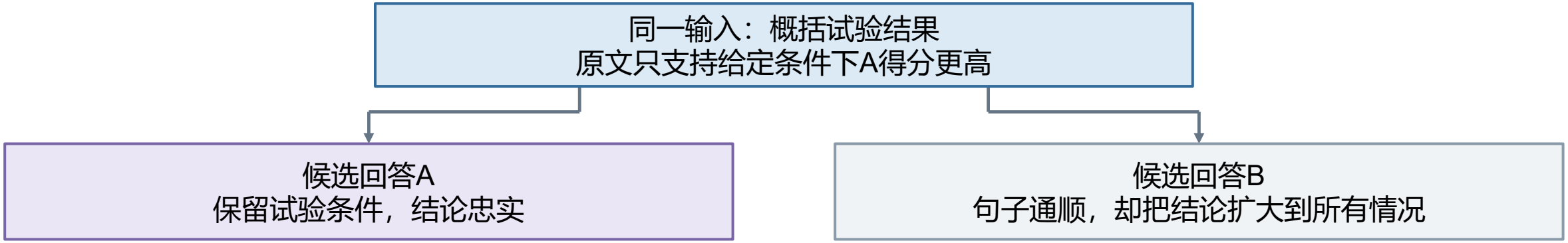
保留色词任务作对照：示例：红 → R；示例：蓝 → B；输入：蓝，期望B。这里查询已见，作用是对齐信息位置；前页编号例才说明不同查询的模式使用。



参数学习把任务信息反映到可训练参数中；上下文学习把任务信息提供在当前输入中。指令模型或LoRA模型也能结合ICL。输出行为变化本身不能证明权重变化；无当前更新也不代表模型从未训练。

从示范监督到偏好信号

SFT已能通过期望回答学习质量要求；当标注者能比较两个回答时，还可以直接表达相对偏好。两种数据表达的监督信息不同。



比较标准：忠实概括原文；相对更偏好A

示范监督提供“希望模仿的回答”；偏好监督提供“同一输入下哪个回答更符合给定标准”。这不是SFT无法学习质量的证明，也不是ICL的后续步骤；下面把比较写成明确的数据对象。

偏好数据与奖励信号

把上一页的比较记录为一个训练样本：提示x、较偏好回答y_w和较不偏好回答y_l。Chosen/Rejected是比较标签，不是模型评分。

Prompt：概括“在本试验条件下，A的得分高于B。”

Chosen：y_w
在给定试验条件下，A的得分更高。

Rejected：y_l
A在所有情况下都比B好。

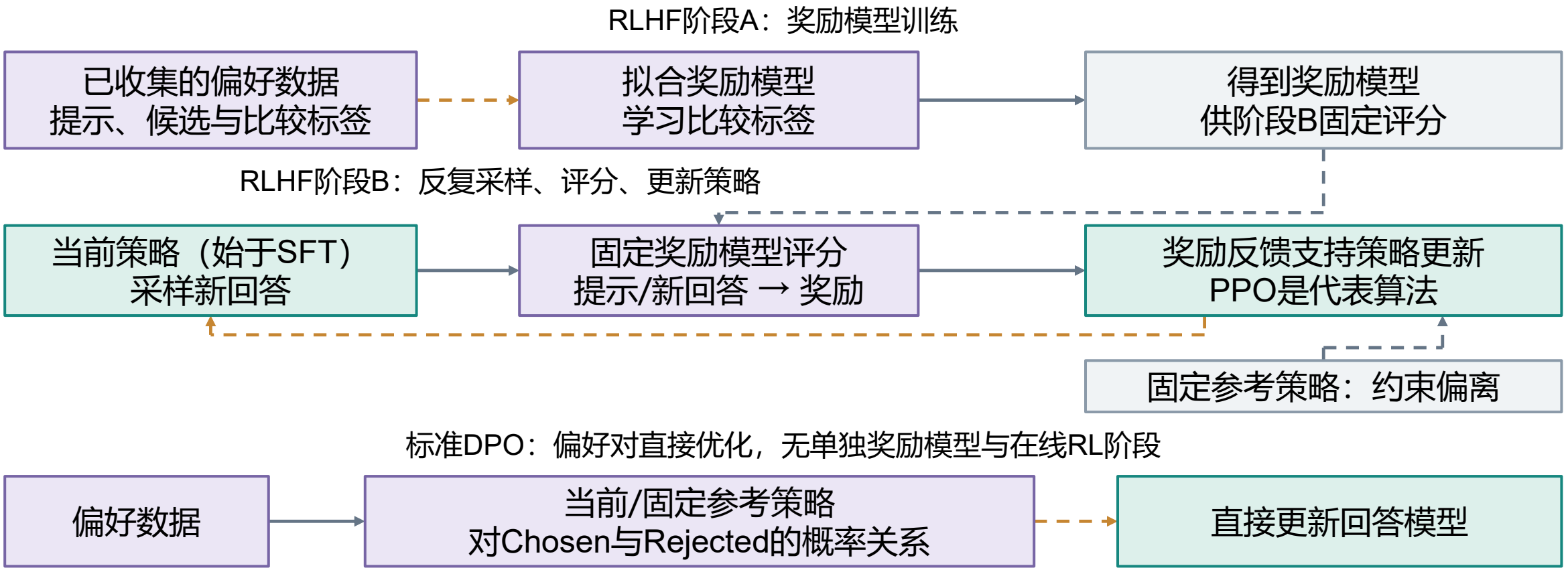
偏好训练样本
(x, y_w, y_l)

用比较信息拟合奖励模型
评分时输入(x,y)，输出标量反馈

偏好标准仍是“忠实概括”，原试验真伪另需证据。**奖励模型**由比较数据学习估计偏好；其分数是代理信号，不是事实正确率、客观真值、生成概率或通用质量评分。

RLHF与DPO

回答策略生成回答；**参考策略**固定不变，用于约束策略偏离。下面分开奖励模型训练与策略优化。



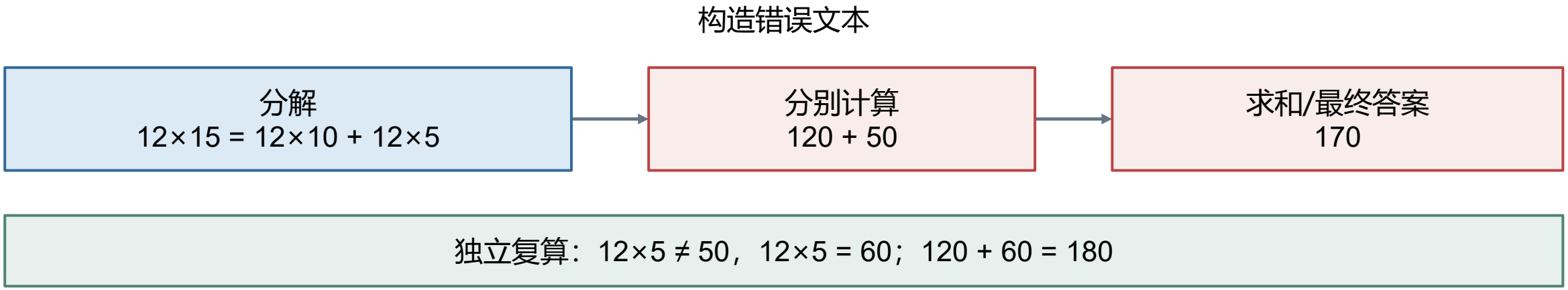
DPO仍有隐式奖励解释，并非Chosen-only SFT。**优化偏好不保证具体输出的事实正确性。**不作普遍优劣排序。出处：Ouyang等，InstructGPT，2022，§3.1；Rafailov等，DPO，2023，§3—4。

3 结果依据与外部接口

行为更符合要求，具体答案是否正确？缺信息或执行时怎样补充？

推理步骤与最终答案检查

偏好优化引导回答行为，不能单独确认某份输出正确。**可见推理文本**展示步骤与解释；最终答案给出结论，两者分别需要检查。



原文本为$12\times15 = 12\times10 + 12\times5 = 120 + 50 = 170$；复算为$12\times15 = 120 + 60 = 180$。分配关系成立，局部乘法错误向后传播。

- **可见步骤**：12×5算错；**最终答案**：170与复算180不一致。
- **内部忠实性**：仅凭这段文本不能确定它是否反映内部真实计算过程。流畅或写出步骤不能证明正确，最终答对也不能反推出每步成立。

证据、稳定性与成本

验证证据是支持或反驳具体结果的来源、复算、测试或实际观察，须与主张及适用条件对应。“输出正确”是可用性的一个条件，实际采用还取决于任务质量、稳定性与成本。

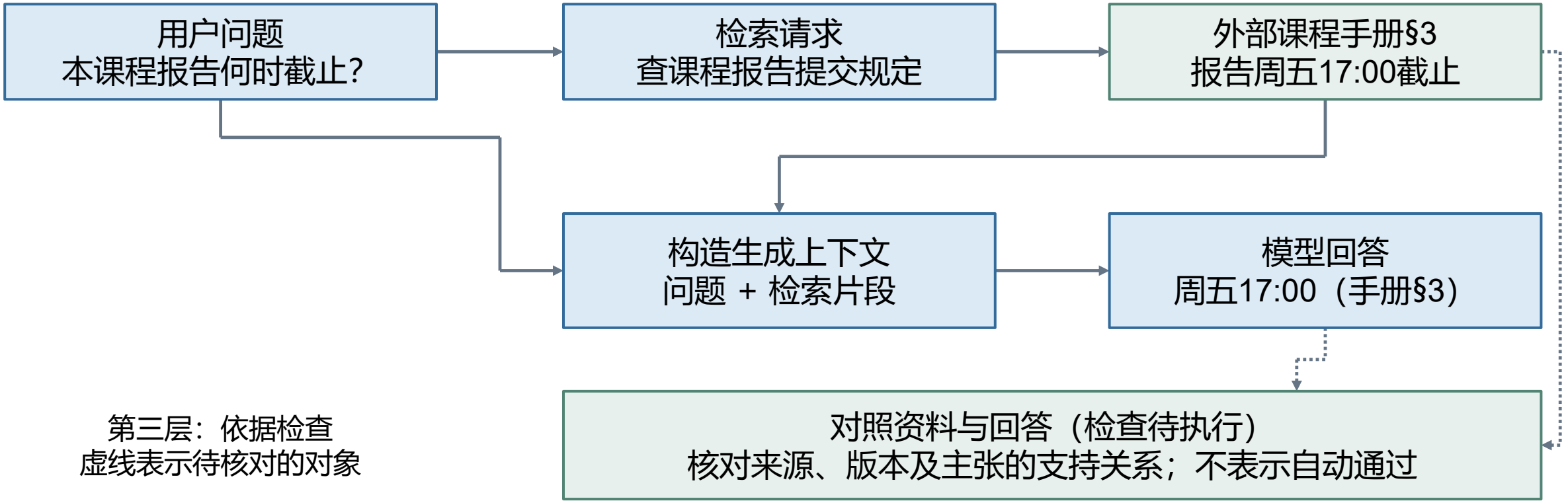
检查对象	需要明确的关系
事实依据	来源可靠、相关，内容确实支持该主张；仅有链接不够
复算与测试	针对同一输入和要求，并有实际计算/执行依据；模型自评仍待核查
稳定性	明确任务、输入、模型及解码条件，比较多次输出；单次成功不代表总体表现
成本	输入输出长度、调用/计算量、延迟和验证负担，与质量要求共同权衡

重复调用可能增加验证信息，也消耗资源；更长步骤不自动更可用。温度等解码条件须按第3讲接口记录；证据不足时保留不确定性。

下一步：若缺当前文档的信息，可检索资料；若缺实际运算或执行，可调用工具。取得资料与执行观察以后，仍要检查支持关系。

RAG外部资料

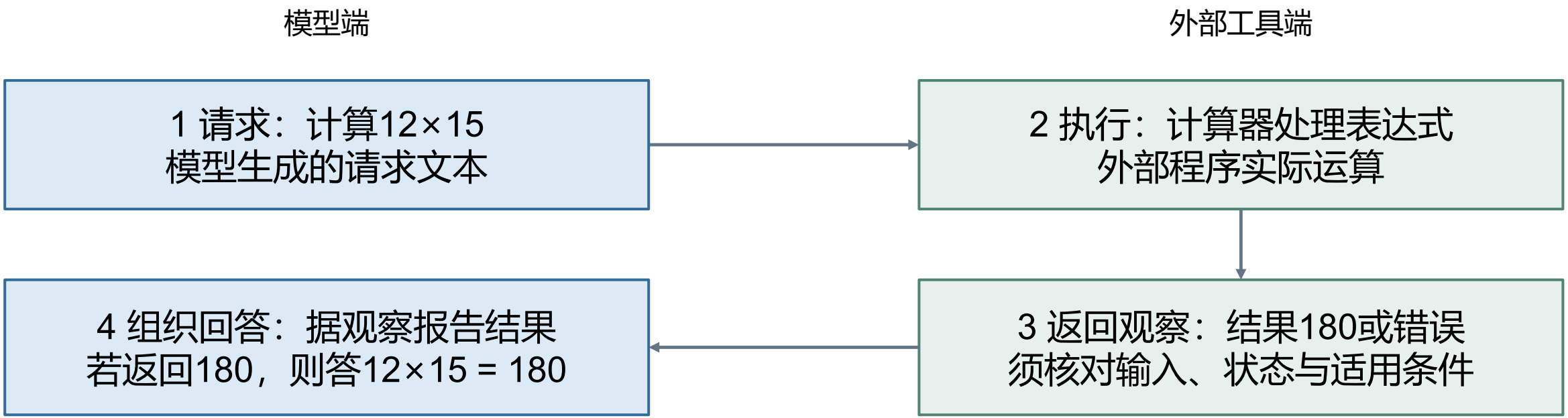
检索增强生成（Retrieval-Augmented Generation，RAG）检索相关资料，将资料与问题共同提供为生成条件；资料通常不直接更新模型权重。



手册、时间及流程均为构造示意，未运行检索或核验服务。检索命中不等于资料正确；依据检查还要判断资料是否支持回答，取得资料不代表验证成功。

工具调用与外部执行

工具调用把模型提出的请求交给外部程序执行，再把结果或错误作为观察返回，模型可以据此组织回答。



这是接口示意，未实际调用外部服务，非执行日志。模型说“已经计算”不等于工具执行；只有实际记录才支持执行发生。结果180与前面复算一致，但工具也可能收到错误输入、执行失败或返回不适用结果，采用前仍需核验。

本讲总结

各方法分属不同维度，可以组合。以下是职责框架，不是必须逐项执行的技术流水线。

训练与监督 SFT：期望回答与有效位置	可训练参数 全量微调 / LoRA	偏好信号 RLHF / DPO
推理上下文 Prompt / ICL	外部资料与执行 RAG / 工具调用	结果检查 依据 / 稳定性 / 成本

可选组合示例（无性能排名；各项均可按任务选择）



- 不同适配方式改变的对象不同，而且可以组合。
- 优化模型行为与验证具体结果是两个不同问题。
- 方法选择与结果采用须结合任务要求、质量、稳定性和成本；不作未经验证的通用性能排名。