预训练语言模型

机器学习 · 语言模型专题

学习目标

本讲问题：如何从大量文本中学习能被不同任务复用的语言模型？

- 解释文本怎样构造监督信号，区分参数、上下文表示与输出接口
- 跟踪GPT与BERT的输入、可见信息、目标及损失位置，解释续写与分类迁移
- 读懂T5的去噪源、哨兵和目标序列，按四个维度比较代表模型
- 说明数据、参数和计算的协调关系，以及具体任务仍需评价

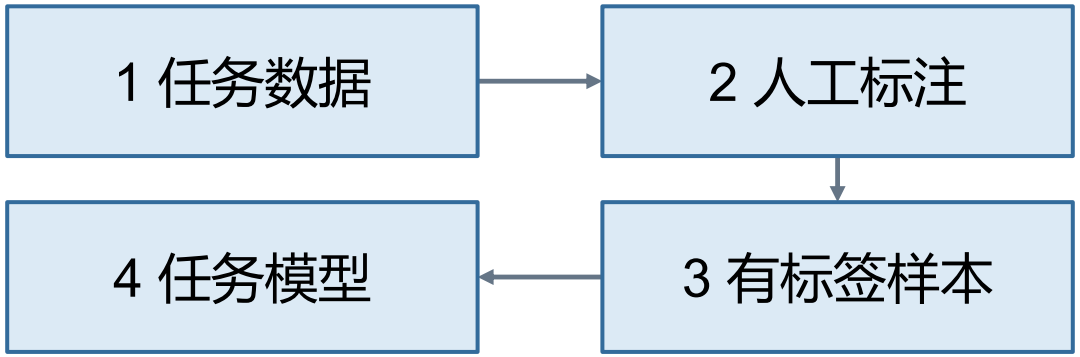
结论：重点是读懂训练机制与迁移接口，完整示范和必要提示提供支持。

先看标注不足时，大量文本能怎样参与学习。

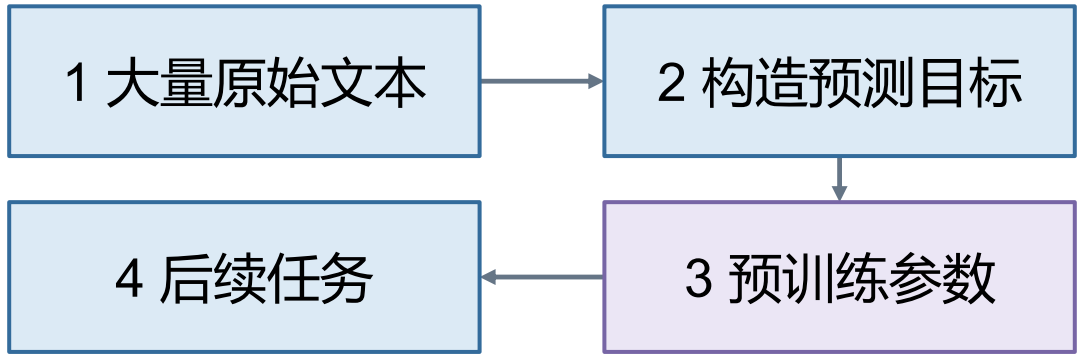
大量文本没有人工标签，如何用于训练？

文本相对丰富，针对具体任务的人工标注可能不足。可以先让文本提供预测目标，再把学到的参数用于后续任务。

具体任务的监督学习



预训练再使用或适配

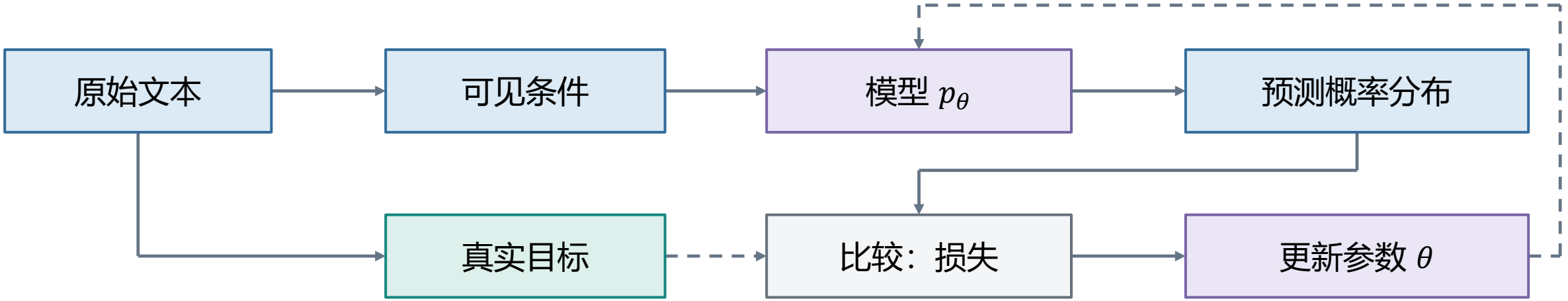


结论：预训练先学习可复用的参数，再根据具体任务使用或适配。

没有人工任务标签，预测目标从哪里来？

文本自身提供预测目标

自监督学习从文本本身构造标签。原始文本、可见条件、真实目标和有效监督位置具有不同身份。



依据什么：可见条件　预测什么：原始标签
哪些位置计入损失：由有效监督集合决定

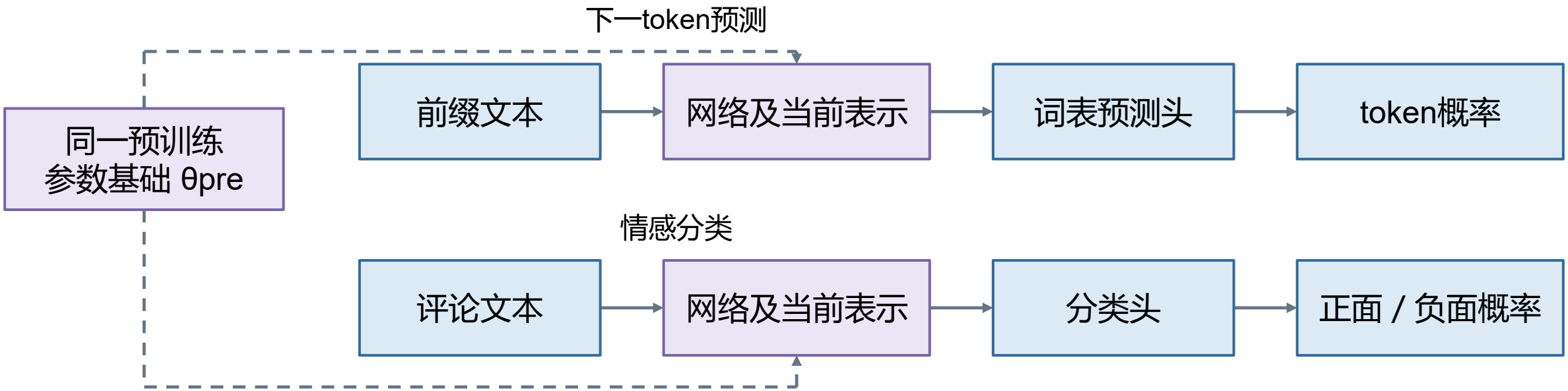
虚线反馈：更新参数
用于下一次预测

结论：平均负对数似然 $L(\theta) = -\frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}} \ln p_\theta(y_i \mid c_i)$：只对有效位置集合$\mathcal{I}$平均。

训练留下的参数，怎样进入不同任务？

迁移复用参数，任务决定输出接口

可复用的是参数。上下文表示由当前输入经过网络计算得到，再交给任务输出接口。



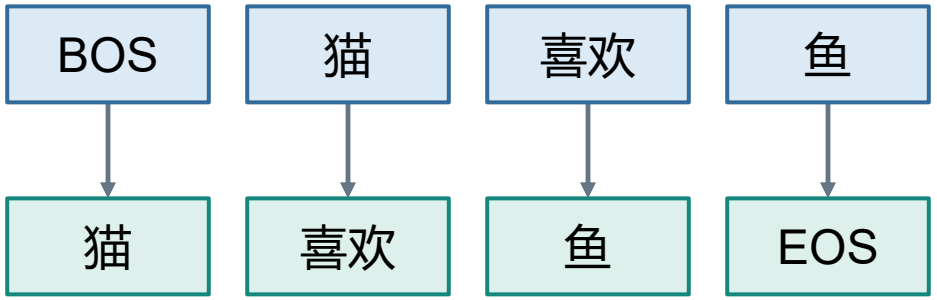
结论：词表头输出token概率，情感分类头输出正面 / 负面概率；任务决定表示的使用和输出形式。

网络可以冻结作为特征提取器，也可以与任务头一起微调。先看GPT怎样预训练。

因果预测限制每个位置可读的前缀

GPT依据已有前缀预测下一token：$p_\theta(x_t \mid x_{0:t-1})$。Decoder-only网络的当前表示不能依赖待预测token及后续位置。

输入移位：确定输入与标签对应



因果掩码：行查询，列为输入位置

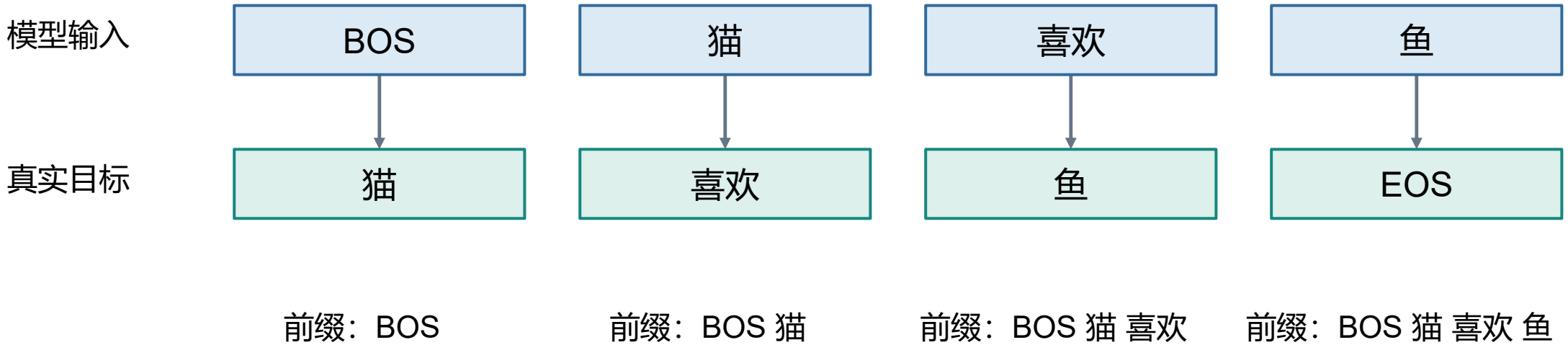
	BOS	猫	喜欢	鱼
BOS	可见	×	×	×
猫	可见	可见	×	×
喜欢	可见	可见	可见	×
鱼	可见	可见	可见	可见

结论：移位决定“哪个位置预测哪个标签”，掩码决定“当前位置可以读取哪些输入”。

位置可并行计算，仍各自受因果限制。下一页逐列展开四个前缀。

同一短序列对应四个因果预测

固定token边界。完整目标序列为猫 / 喜欢 / 鱼 / EOS，输入为BOS / 猫 / 喜欢 / 鱼。每列是一个有效预测位置。



结论：真实目标只用于损失；当前预测位置实际读取的是图下方对应前缀。

BOS仅作输入，EOS是本例第4个目标，PAD不计损失；截断窗口不能自动补EOS。

损失平均对应真实目标的概率

模型在每个位置输出完整词表分布。下表只取真实标签对应的概率，数值为课程构造。

目标	猫	喜欢	鱼	EOS
真实标签概率	0.5	0.25	0.8	0.4
负自然对数	0.693147	1.386294	0.223144	0.916291

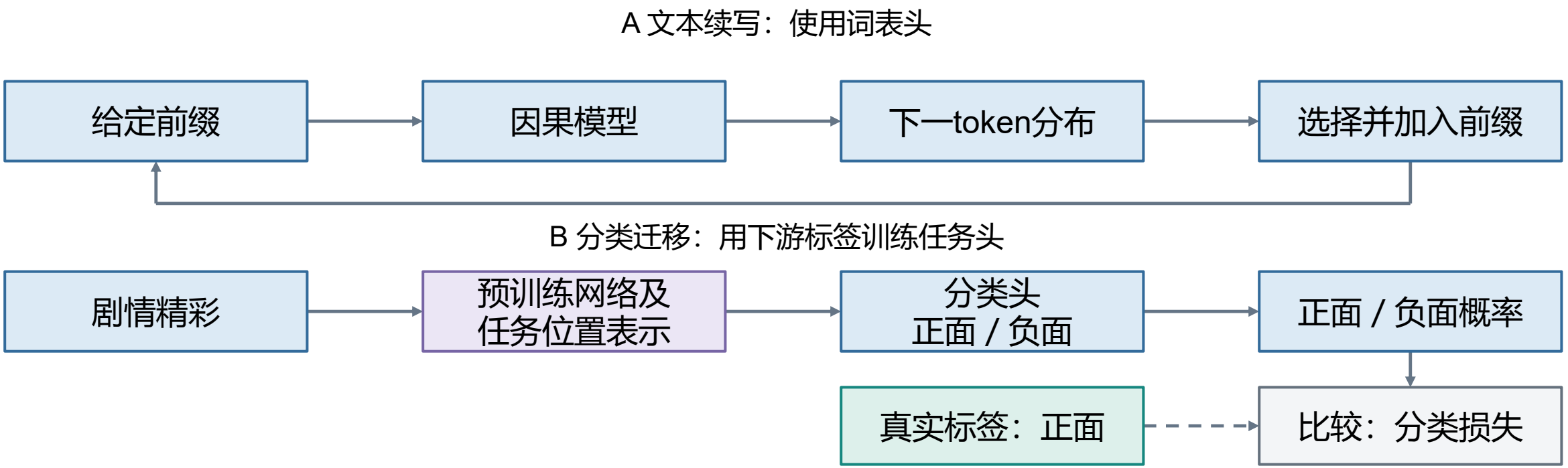
$$L_{\mathrm{AR}} = -\frac{\ln 0.5+\ln 0.25+\ln 0.8+\ln 0.4}{4} \approx 0.804719.$$

结论：标签概率越低，负对数损失越大；分母为4个有效目标，EOS计入、PAD排除。

损失更新嵌入、上下文网络及输出参数。这些参数怎样用于续写与分类?

预训练参数可以支持续写与任务迁移

同一参数基础可以支持两种使用方式，输出对象和后续训练条件不同。

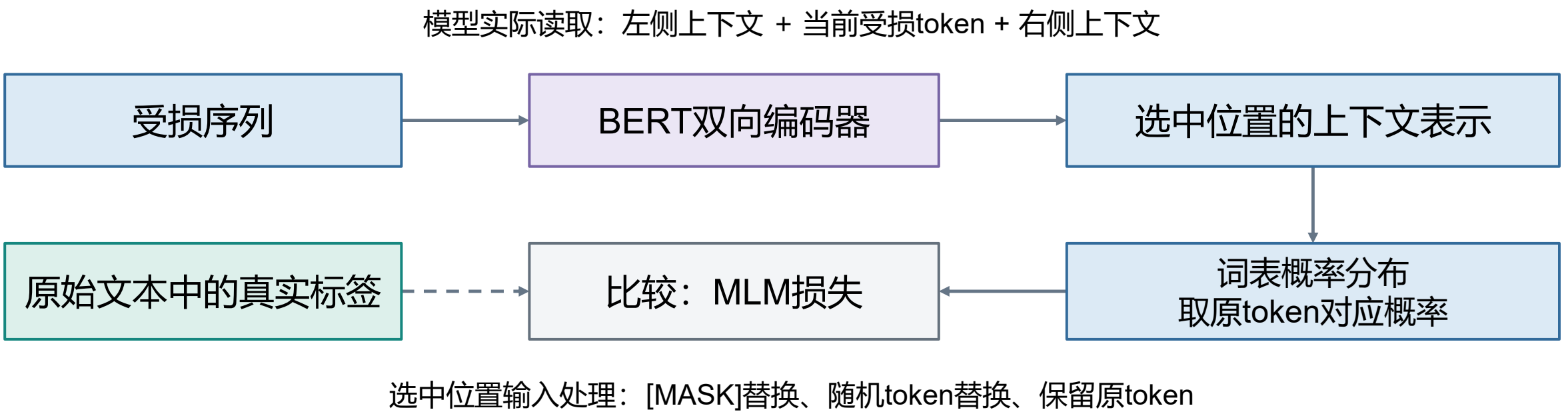


结论：续写循环使用当前前缀；原始GPT分类迁移通过任务头和下游标签学习类别。

如果希望预测某位置时利用左右两侧上下文，该怎样构造训练？

受损输入支持双向上下文预测

为了利用左右信息且保留预测任务，BERT从原文构造受损输入，以双向编码器预测选中位置原来的token。

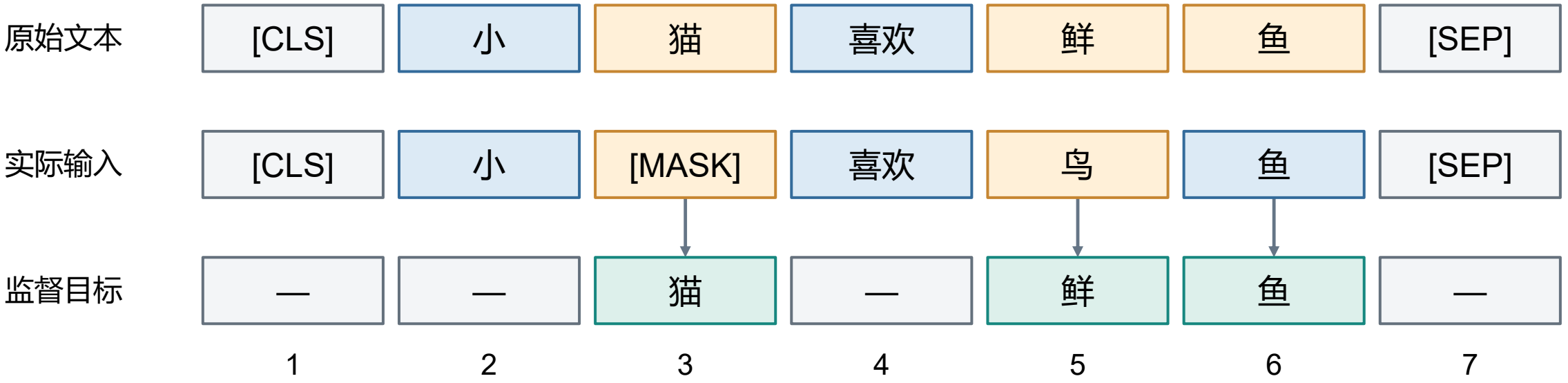


结论：选中监督集合$\mathcal{M}$包含三种处理，实际MASK集合$\mathcal{A} \subseteq \mathcal{M}$，一般不能视作相等。

三种处理怎样在同一序列中定位?

选中位置与MASK位置在同一序列中区分

按位置1—7对齐：$\mathcal{M} = \{3,5,6\}$，$\mathcal{A} = \{3\}$。原始token与受损token逐列对应。



结论：位置3读[MASK]预测猫；位置5读鸟预测鲜；位置6保留鱼，仍预测鱼。

[CLS]是序列级接口，[SEP]用于分隔。其余位置不计本例MLM损失。

MLM只对选中监督位置平均

位置3、5、6预测原标签猫、鲜、鱼。真实标签概率为0.5、0.25、0.8，均为构造数值。

监督位置	3：猫	5：鲜	6：鱼
负自然对数贡献	0.693147	1.386294	0.223144

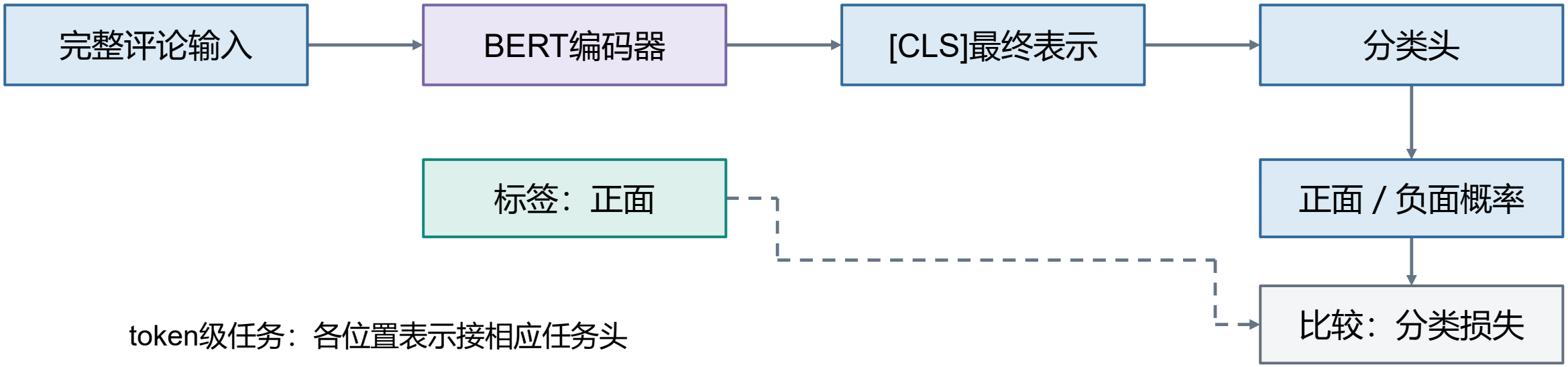
$$L_{\mathrm{MLM}} = -\frac{1}{|\mathcal{M}|}\sum_{i\in\mathcal{M}} \ln p_\theta(x_i \mid \tilde{x}_{1:T}) = -\frac{\ln 0.5+\ln 0.25+\ln 0.8}{3} \approx 0.767528.$$

结论：分母是3个选中监督位置；只算MASK会漏掉5、6，对全部7个位置平均也不成立。

不同目标的损失不能直接比较优劣。训练得到的双向表示怎样接入分类?

双向表示接入任务头，NSP保留原始背景

输入[CLS] 剧情精彩 [SEP]，分类头输出正面 / 负面的类别概率；用真实标签正面训练。下游输入无需继续MLM损坏。



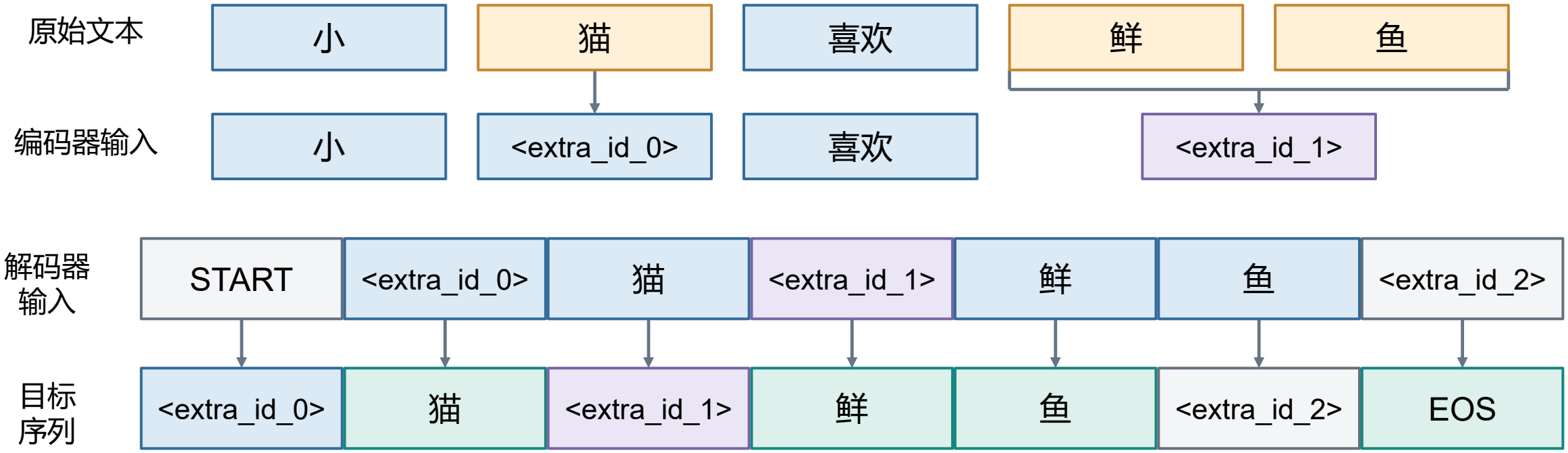
原始BERT还含NSP文本段配对二分类；它是独立预训练目标，并非所有MLM模型的必要机制。

结论：任务头与MLM词表头职责不同，微调可以共同更新预训练参数与新增参数。

若任务既要利用输入上下文，又要生成长度可变的新文本，可怎样组织训练？

缺失片段按哨兵对应为目标序列

T5以编码器—解码器恢复缺失片段。哨兵标记片段身份：第一段猫，第二段鲜 / 鱼。

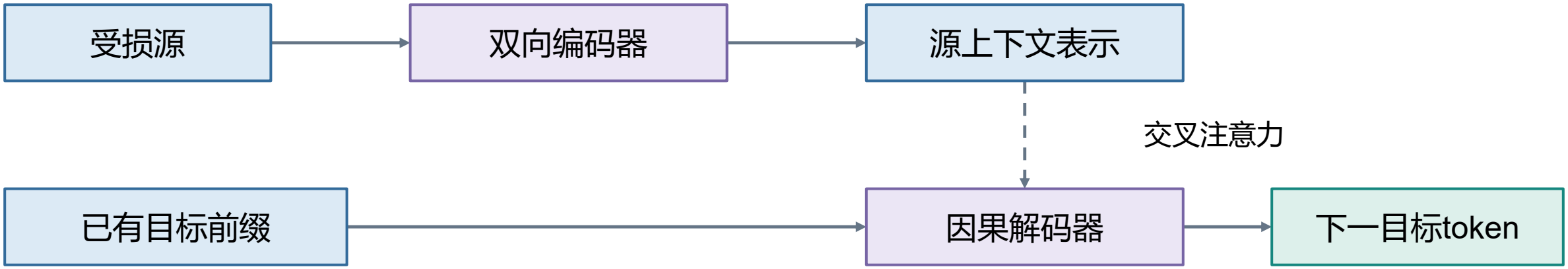


结论：<extra_id_0>对应猫，<extra_id_1>对应鲜鱼；<extra_id_2>结束缺失片段目标串，EOS终止本例目标序列。

7个有效目标只含缺失内容及标记，PAD排除。START仅为解码起始位置的说明性名称。

同一文本接口可以表达不同任务目标

解码器预测$p_\theta(y_t \mid s_{1:S}, y_{<t})$，同时使用源表示与已有目标前缀。



结论：编码器使用受损源的双向信息，解码器按目标前缀预测，并通过交叉注意力读取源表示。

文本到文本例：输入“情感：我喜欢这本书”，目标“正面”。统一接口仍需相应任务训练或适配。

三类代表模型按四个维度比较

先比较预测目标和可见条件，再看结构与任务使用。三种路线是不同设计，不能排列为简单性能递进。

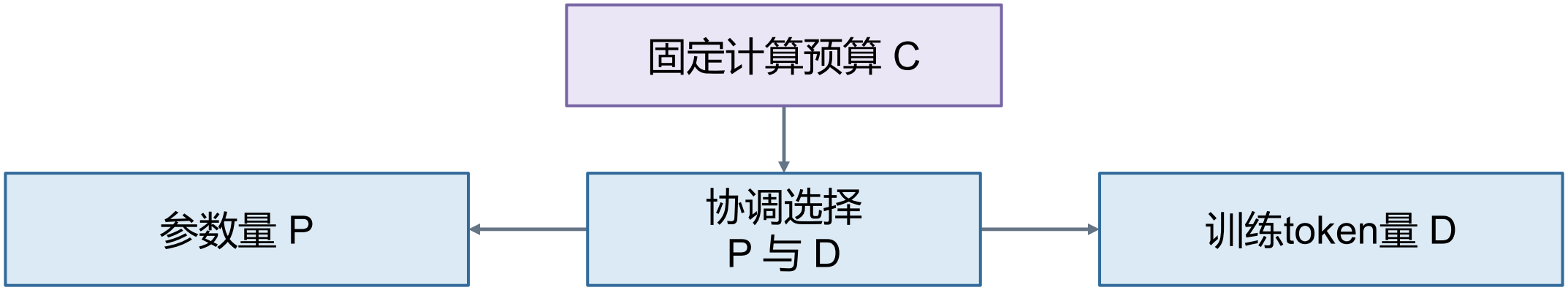
维度	GPT代表设置	BERT代表设置	T5代表设置
训练目标	前缀预测下一token	受损输入预测选中位置原token；原始另含NSP	受损源和目标前缀预测下一目标token
可见信息	当前目标之前的前缀	实际受损输入的双向信息	受损源与已有目标前缀
网络结构	因果Decoder-only	双向Transformer编码器	编码器—解码器
使用方式	续写；经适配用于分类等	表示接任务头，用于分类或位置任务	文本输入与文本输出，按任务适配

结论：方法差异首先体现在预测目标及可见条件，再反映到结构与使用方式。

跨不同预测目标的PPL未必可比；还须对齐数据、分词、上下文与评估协议。确定目标后，能力还受哪些资源条件约束？

数据、参数与预算共同约束预训练

训练目标确定后，参数规模和实际处理的token量需要在计算预算下协调。



固定预算、相近训练设置：模型更大时，可承受的训练token量通常更少。

协调还取决于数据质量与适用性；本图只表达定性关系。

结论：相同D不等于相同独立信息量，参数更多不等于训练充分；数据的重复、噪声和任务适用性同样重要。

C是运算量预算，区别于峰值算力和墙钟时间。预训练预测能力仍需经具体任务评价，回到本讲主线。

本讲总结

统一问题：文本怎样提供可复用的学习基础?

- **文本提供监督**：先确定预测目标、可见条件和有效位置，再用损失更新参数
- **三种训练路线**
  - GPT依据前缀预测下一token
  - BERT依据受损输入预测选中位置原token
  - T5依据受损源和目标前缀生成缺失片段及其标记
- **预训练之后**：复用参数，针对当前输入产生表示或生成输出，以相应接口适配任务并评价
- **资源边界**：数据质量、训练量、参数量和计算预算共同约束训练，预测能力仍需任务证据

结论：沿“目标与条件—参数学习—任务使用—评价边界”恢复全讲，而非用结构名称直接推断能力。