条件语言模型与序列生成

从输入条件到输出质量

学习目标

- 在给定文本任务和必要解释支持下，辨认源序列、目标序列与预测条件，解释编码器和解码器的职责。
- 跟踪目标移位、教师强制、有效位置损失和逐步生成，说明真实前缀与生成前缀的区别及生成终止条件。
- 根据给定候选和概率，解释贪婪、束搜索与采样怎样产生输出，并区分模型概率、搜索分数与任务质量。
- 理解RNN Seq2seq、注意力与Transformer在源表示接口上的联系；说明BLEU的用途与比较边界。
- 初步认识温度、top-k与top-p改变分布或候选集合的方式。

1 条件模型与信息接口

明确源、目标与预测条件，解释条件模型的职责。

翻译同时依赖源文本和目标历史

只生成通顺英文还不能完成翻译，输出须保留源义。**条件生成**在给定外部输入下产生输出；翻译、摘要都是文本条件生成任务。



本例按空格分词、英文小写，词恰好对应token；真实分词也可用子词。源$s_{1:S}$与目标$y_{1:T}$分别编号，长度不要求相等。主要目标是i like cats EOS；完整目标用于训练监督与评价，生成时只用源与已生成历史。

目标序列概率由条件预测相乘

条件自回归模型按源和目标前缀给出下一token分布；参数θ决定分布，输出选择规则另定。下图四次预测均以同一完整源s为条件。



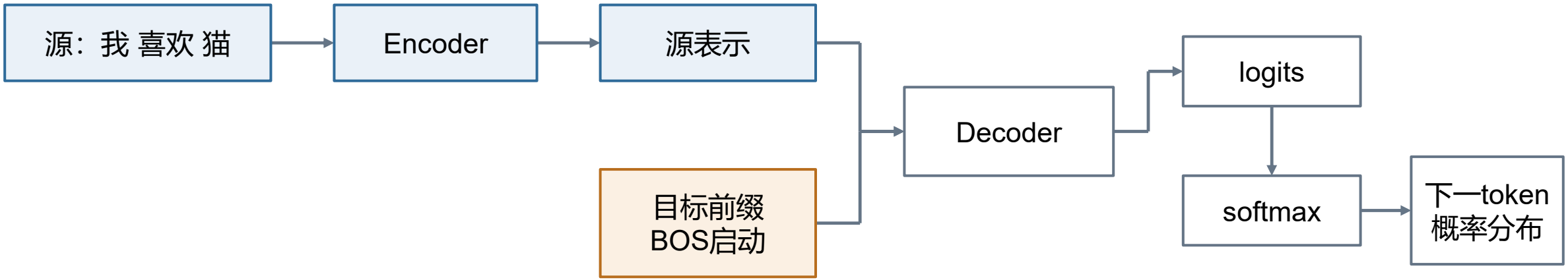
$p(i \mid BOS, s)$　　$p(like \mid i, s)$　　$p(cats \mid i\ like, s)$　　$p(EOS \mid i\ like\ cats, s)$

四次预测的条件概率相乘，形成含EOS的完整目标概率

$p_\theta(y_{1:T} \mid s_{1:S}) = \prod_{t=1}^{T} p_\theta\left(y_t \mid y_{<t}, s_{1:S}\right)$. $y_{<t}$是t之前的历史；每个分布在目标词表上归一化。BOS启动且不作标签，EOS计入完整概率，PAD只补齐。公式描述同一个逐步预测过程；具体网络实现另定，本讲自回归设定不代表所有条件生成形式。

编码与解码组织两类条件信息

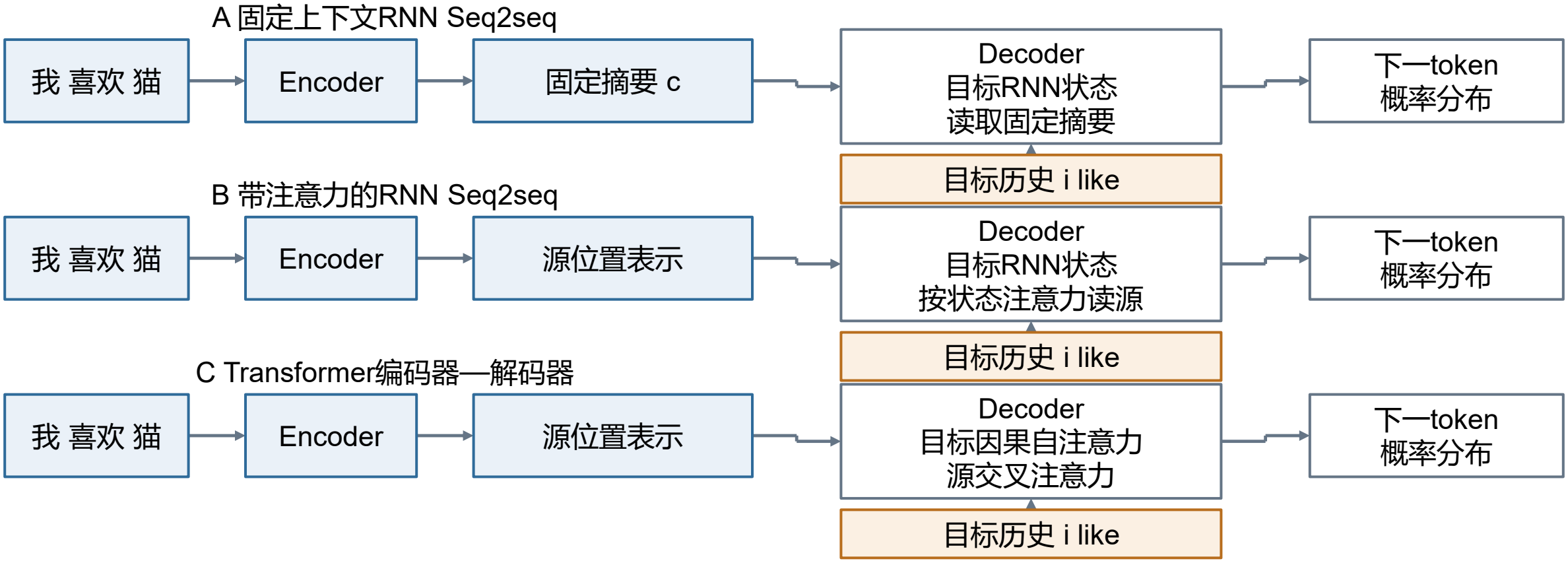
编码器—解码器：Encoder组织源表示，Decoder结合源表示与目标历史预测下一token。两类条件共同进入Decoder。



logits是未归一化得分$\boldsymbol{z}_t \in \mathbb{R}^{V_{\text{tgt}}}$，每项对应一个目标token，$V_{\text{tgt}} = |\mathcal{V}_{\text{tgt}}|$。softmax将其转为概率。模型计算止于分布，怎样选择输出留给解码策略。编码器—解码器是本讲主要接口，并非所有条件模型都必须采用。

不同实现保留条件生成的共同职责

同一输入、目标历史和输出职责，比较源信息的读取方式。



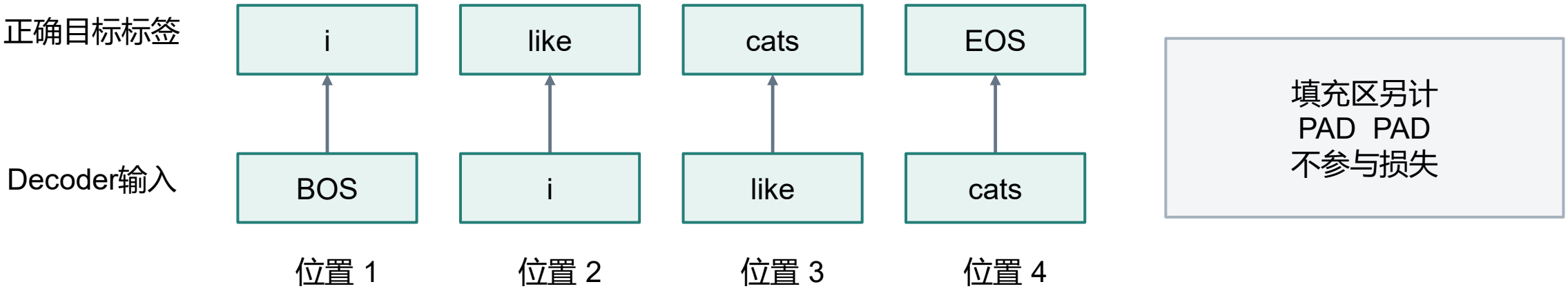
三行都保留Decoder预测职责，差别是源信息的读取方式。源有效内容可完整可见，目标未来受限，PAD须排除；内部运算沿用DL基础。

2 训练监督与生成反馈

区分训练与生成的输入、监督及可见信息。

右移输入与原位标签形成监督

源“我 喜欢 猫”固定。**教师强制**用真实目标历史预测下一token；四个有效位置按下图对齐。



预测cats：输入历史为BOS、i、like；cats是当前监督标签

有效位置满足$d_1 = \mathrm{BOS}$、$d_t = y_{t-1}$ $(2 \le t \le 4)$，EOS计入监督。

当前正确标签不作为当前位置的输入。填充区仅作批量对齐示意；本页重点是四个有效位置的输入与标签对应。

可读信息与损失位置由不同约束控制

移位输入位置t预测y_t；因果自注意力只允许读取$j \le t$的输入列。

A 因果可见性：行是预测位置，列是移位输入

	BOS	i	like	cats
预测i	可读	禁读	禁读	禁读
预测like	可读	可读	禁读	禁读
预测cats	可读	可读	可读	禁读
预测EOS	可读	可读	可读	可读

第3行可读BOS、i、like
输入列4是cats，当前仍不可读

键PAD掩码：源 / 目标PAD键排除读取。
目标键还要满足因果条件。

B 目标损失掩码

i：计入	like：计入	cats：计入	EOS：计入	PAD：排除	PAD：排除

三种掩码限制不同对象：信息不可见与标签不计损失须分别设置，损失分母也须另定。Transformer可并行计算受因果约束的训练位置；RNN仍依赖前一递归状态，教师强制不消除递归依赖。

有效位置上的负对数概率形成训练损失

平均负对数似然在非PAD目标位置集合$\mathcal{I}$上平均（$|\mathcal{I}| > 0$），n为批内样本。正确概率越低，负对数损失越大；图中为构造概率。

i	like	cats	EOS	PAD	PAD
正确概率 0.6	正确概率 1	正确概率 0.5	正确概率 1	不计损失	不计损失
−ln p ≈ 0.5108	−ln p ≈ 0.0000	−ln p ≈ 0.6931	−ln p ≈ 0.0000	不计分母	不计分母

$$L(\theta) = -\frac{1}{|\mathcal{I}|}\sum_{(n,t)\in\mathcal{I}} \ln p_\theta(y_t^{[n]} \mid y_{<t}^{[n]}, s_{1:S_n}^{[n]}).$$

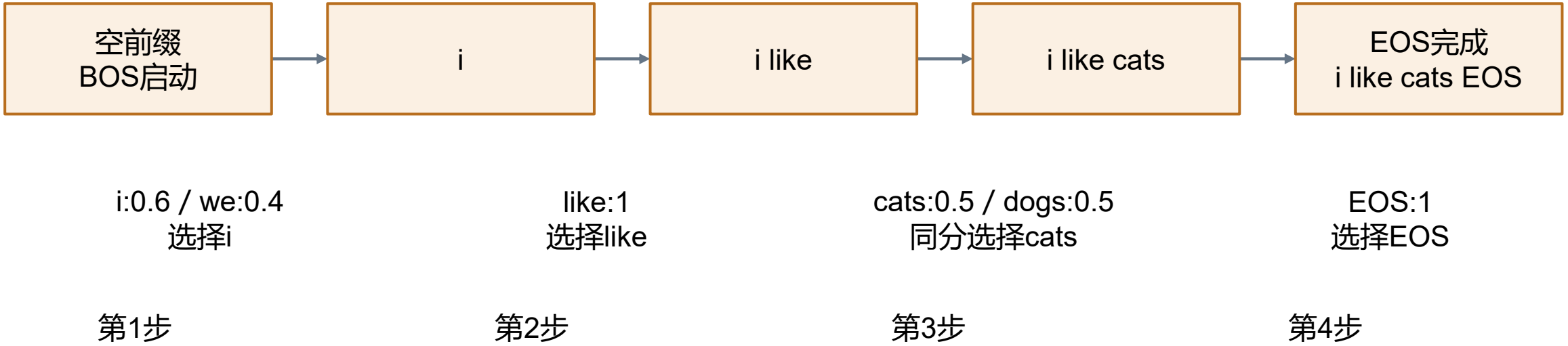
$L = -\ln(0.6\times1\times0.5\times1)/4 \approx 0.300993.$

分母核对：EOS计入、PAD排除；错误地除以补齐长度6会得到0.200662。按token平均通常不同于逐句先平均。

训练用此损失联合调整编码器与解码器，无需另造源表示标签。

生成把自己的输出反馈为下一步历史

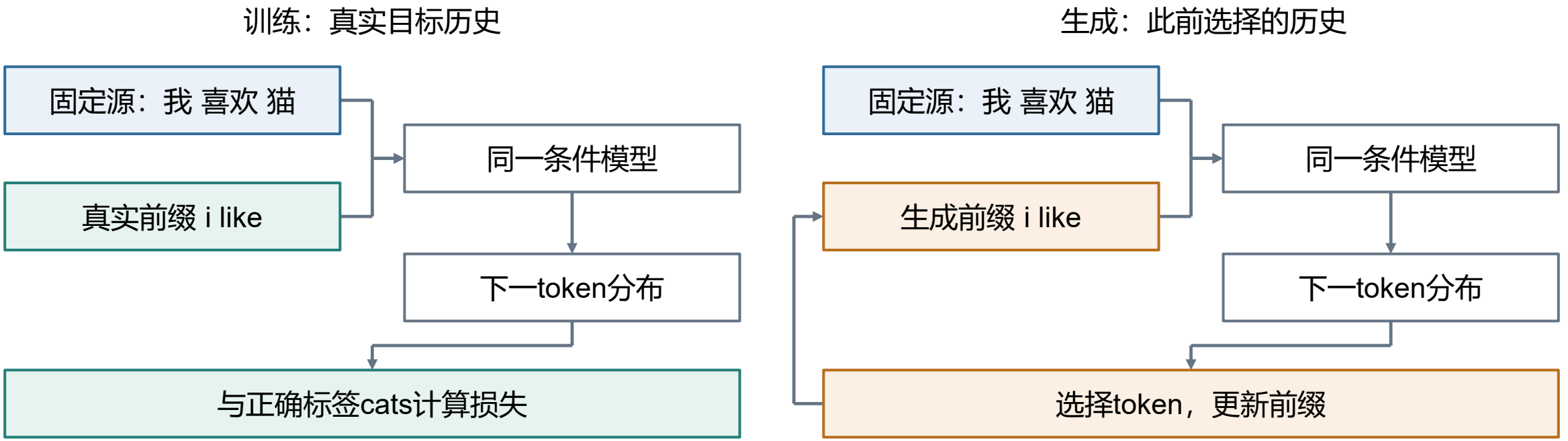
训练用真实历史；生成使用自己的**生成前缀**。源固定，模型参数在一次生成中不更新。本例采用**贪婪选择**，每步选最大条件概率，并列按字母顺序。



每步依次取得“当前前缀 → 模型条件分布 → 选择后的新前缀”。分布来自同一构造模型，未列token概率为0。真实参考用于训练或评价，生成时不会作为未知答案输入。

选择的token改变下一次预测条件

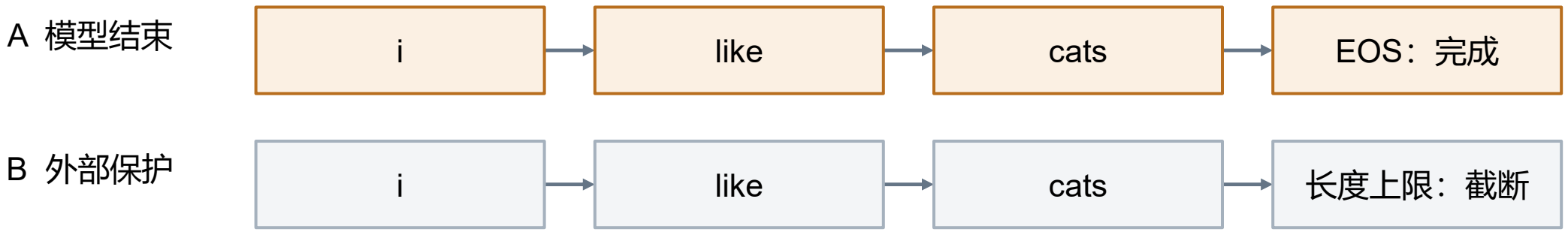
两侧共享模型职责和源条件。训练更新参数；一次实际生成中参数保持固定。



训练由真实历史提供下一位置条件；生成把已选token反馈为下一步历史。选择i或we后，后续分布可以不同。一次局部选择因而会影响整句，而非只改变当前输出。

EOS完成与长度截断保留不同含义

EOS是模型预测的结束token；T_{max}是外部长度保护。本例EOS计为一个生成token，路径A上限4，路径B上限3。



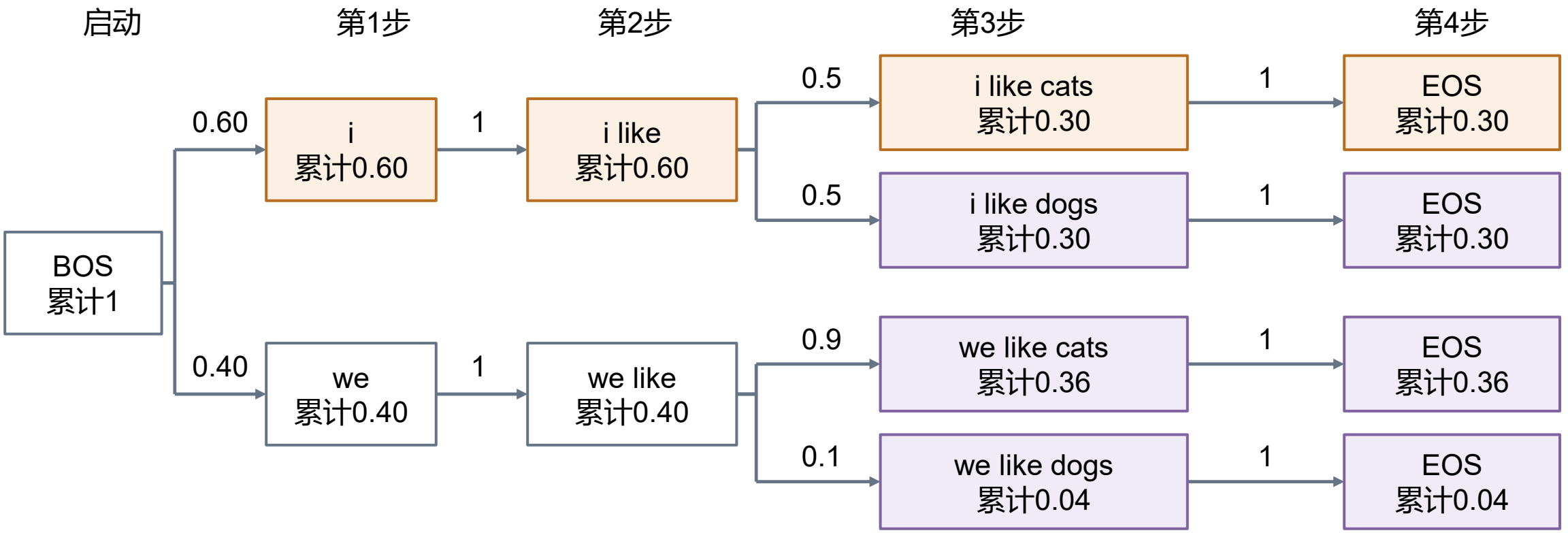
两条路径可显示同样的正文i like cats，内部候选状态不同。完整概率保留EOS因子：$p_\theta(\text{i like cats EOS} \mid s) = 0.6\times1\times0.5\times1 = 0.30$. 本例EOS概率恰为1，截断前缀与完整序列概率数值相同；一般不能推广。显示正文可去除BOS、EOS、PAD，内部状态和评分仍须保留各自职责。

3 从概率分布到输出选择

比较贪婪、束搜索与采样怎样使用条件分布。

后续分布依赖已经选择的前缀

同一源“我 喜欢 猫”。边旁标条件概率，节点内标累计路径概率；琥珀节点标出贪婪路径（同分按字母顺序）。



$p(\text{cats} \mid \text{i like}, s) = 0.5$，$p(\text{cats} \mid \text{we like}, s) = 0.9$：它们属于不同前缀的分布。每个父前缀下概率和为1，未列项为0；EOS因子也计入完整概率。

束搜索保留多条前缀继续比较

束搜索每步扩展当前保留前缀，按累计得分排序并剪枝。本例束宽$b = 2$，按原始累计概率比较，同分按完整候选的字母顺序。



第1步保留i(0.6)、we(0.4)，第2步分别接like(条件1)。图示第3步四条扩展；第4步仅给两条保留候选接EOS(条件1)，完成后不再扩展。贪婪：i like cats EOS，0.30；束搜索：we like cats EOS，0.36。有限束通常不保证全局最优。

搜索得分说明选择标准，质量还须另判

计算中通常累加自然对数概率，避免直接连乘很小的数：

$$\ln p_\theta(y_{1:T} \mid s) = \sum_{t=1}^{T} \ln p_\theta(y_t \mid y_{<t}, s).$$

完成候选	原始模型概率	原始对数概率
i like cats EOS	0.30	-1.203973
we like cats EOS	0.36	-1.021651

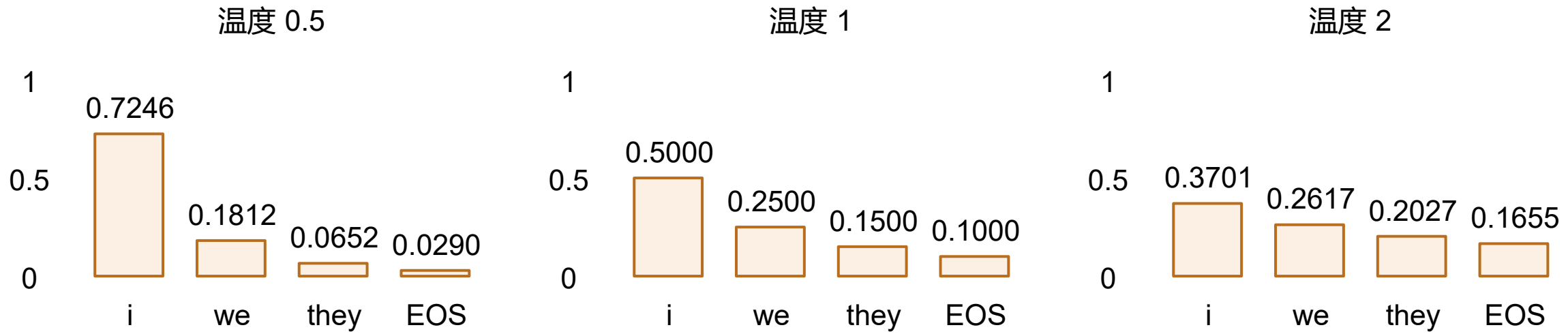
自然对数单调递增，因此按原始对数概率与按原始概率得到相同排序。若为处理长度差异而调整得分，调整后的**搜索得分**是新的比较标准，不能继续称为原始序列对数概率。

- 束搜索只保留有限候选，评分、束宽和终止处理共同影响结果；扩大束宽没有任务质量必然改善的保证。
- 长度截断候选仍是前缀，EOS完成候选包含结束因子；比较时须保留它们的身份。

高概率候选的源义留待第四节检查。下一步：除了按分数选择，也可以按分布随机采样。

温度改变采样分布的集中程度

采样按分布随机抽取token并反馈。独立构造分布q(i,we,they,EOS)=(0.50,0.25,0.15,0.10)，与前面候选树分开。三图纵轴均为0到1。



低温更集中，温度1恢复原分布，高温更平坦。

$$q_\tau(v) = \frac{q(v)^{1/\tau}}{\sum_u q(u)^{1/\tau}}.$$

正温度$\tau > 0$将logits除以τ，等价于上式；u、v取自目标词表。显示值四舍五入。温度控制集中度，不能保证任务正确。

top-k固定候选数，top-p固定累计概率门槛

top-k按概率降序保留前k项；**top-p**取累计质量至少达到p的最短前缀集合。$1 \le k \le V_{\text{tgt}}$、$0 < p \le 1$，并列按声明规则排序。

top-k：k=2

i	we	they	EOS
0.50	0.25	0.15	0.10

保留质量 0.75

归一化：0.6667、0.3333、0、0

top-p：p=0.8

i	we	they	EOS
0.50	0.25	0.15	0.10

保留质量 0.90

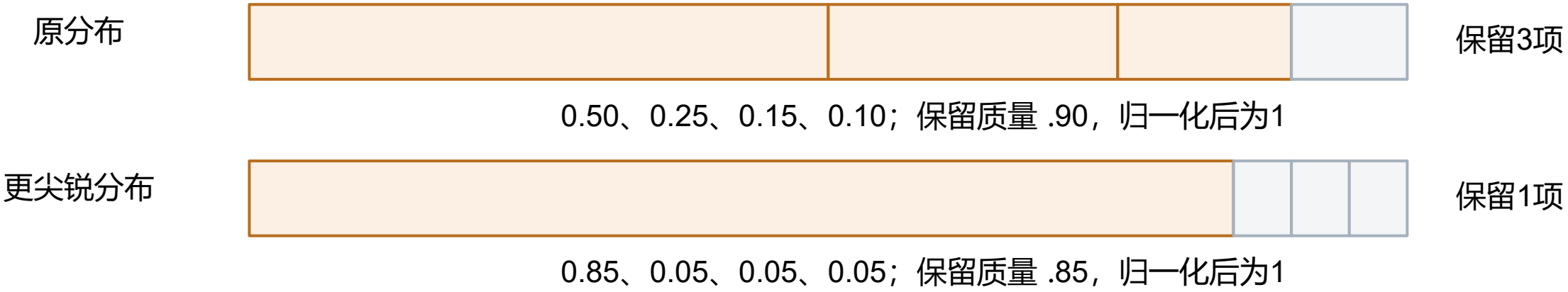
累计：0.50 → 0.75 < 0.8 → 0.90 ≥ 0.8

归一化：0.5556、0.2778、0.1667、0

两种设置分别从原分布q出发，不先作温度变换。top-k固定数量；top-p固定累计门槛，前两项质量0.75不足0.8，故须加入they。截断后均除以保留质量，再采样；本次排除EOS不表示整次生成始终排除EOS。

截断后的归一化与动态候选集合

截断只留下部分概率质量，各保留项除以同一保留质量，才能重新构成总和为1的采样分布。两行都使用top-p=0.8。



对保留集合A，且 $\sum_{u\in A} q(u) > 0$： $q_A(v) = \frac{\mathbf{1}[v\in A]q(v)}{\sum_{u\in A} q(u)}$. 指示函数在v属于A时为1，否则为0。原分布需3项，尖锐分布只需1项：top-p集合随当前分布而变。每个生成位置都可重新确定集合。温度与截断联用须声明顺序和同分处理，规则不自动保证正确。

4 任务质量与评价证据

回到源与任务判断质量，说明参考指标的证据。

翻译质量须回到源文本和任务要求

源为“我 喜欢 猫”，参考为i like cats。翻译要求保留源义，并形成可理解、自然的目标表达。

候选	对源义与表达的判断
i like cats	保留本例的主语、喜欢关系与对象
we like cats	把“我”变为“我们”；它虽是本构造模型的最高概率序列，仍改变了源义
i like **dogs**	对象由猫变为狗
i like cats **cats**	主干含义接近参考，但重复降低表达质量
i am fond of cats	可表达与源义接近的意思，词面与单一参考不同

模型偏好、搜索标准与任务质量各有依据

同一构造模型的最高概率候选把“我”变为“我们”。输出怎样被选中，与输出是否完成翻译任务，需要分别解释。

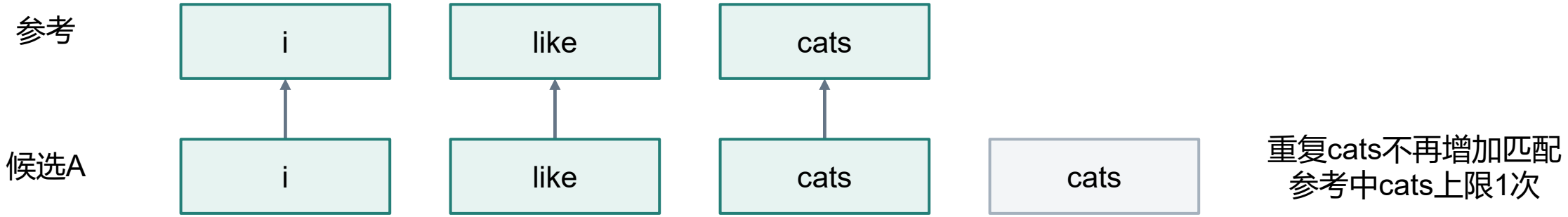
- **模型概率**：模型认为这段输出有多可能？we like cats在本例为0.36。
- **搜索得分**：解码策略怎样比较候选？按原始概率，0.36排在0.30前；长度调整可改变标准。
- **任务质量**：是否满足源与任务？we把“我”改成“我们”，仍须检查意义、完整性和表达。

三者可以有关联，却没有由单个分数推出完整任务质量的普遍保证。更多搜索候选也不能自动修正模型偏好与任务要求之间的差异。

参考文本可提供可重复的匹配证据，但只代表可接受表达的样本；i am fond of cats与单一参考词面不同，仍可能表达接近的意思。

BLEU裁剪重复匹配并汇总语料计数

BLEU通过裁剪n-gram匹配精确率、加权几何平均与过短惩罚，提供参考匹配证据。n-gram是连续n个token片段；先看单词组件怎样裁剪重复。



A：3/4　B（i like cats）：3/3　语料累计：6/7 ≈ 0.857143

裁剪按参考中出现次数限制匹配，不是任意挑某一个重复词。跨句累加匹配数与候选总数得到$6/7$，不同于逐句精确率平均$(0.75+1)/2=0.875$。图中两记录仅解释单词组件，不是完整BLEU分数或正式测试集。

长度惩罚与参考条件限定BLEU的解释

参考i like cats；候选i like，遗漏对象cats。

短候选可能有较高局部匹配精确率，故需过短惩罚。语料级c、r分别表示候选与有效参考总长。

$$\mathrm{BP}=\begin{cases}1, & c\geq r,\\ \exp(1-r/c), & 0<c<r.\end{cases}$$

本例$c=2$、$r=3$，代入得$\exp(1-3/2)=\exp(-0.5)$：

$$\mathrm{BP}\approx 0.606531.$$

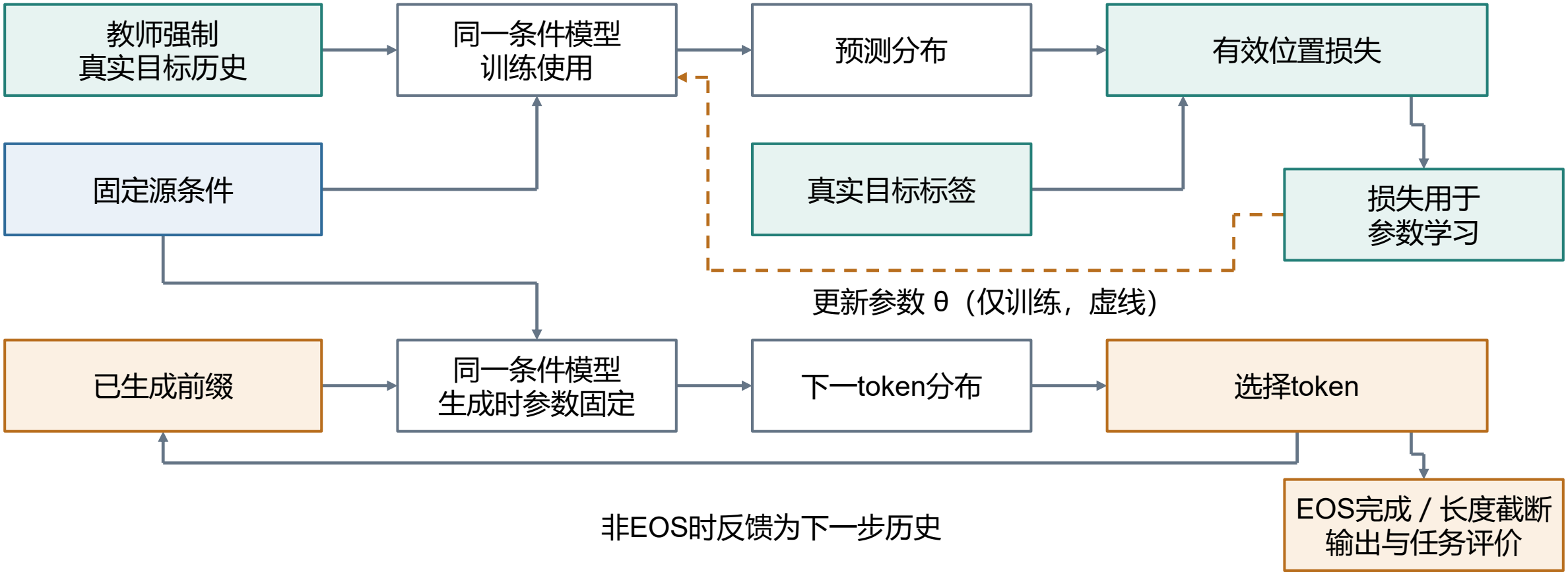
指标组成：BP是长度因子；完整BLEU还包含多阶裁剪精确率的加权几何平均，主要用于语料级比较。

比较条件：分词、大小写、参考集合、数据与实现设置须可比。多参考时取最接近候选长度的参考，等距并列处理须说明。

解释边界：无平滑时，某阶零匹配可使BLEU为零，短句尤其容易发生。BLEU不能独立保证源义，也不能穷尽合法改写。

本讲总结

训练与生成是同一模型的不同使用阶段；训练损失同时读取预测分布和真实标签。



训练用真实历史作条件、真实标签作监督；生成用已选token反馈历史。模型概率、输出选择与任务评价分别承担不同职责。