arXiv:2602.01395v1 [cs.CL] 1 Feb 2026

重新思考选择性知识蒸馏

Almog Tavor [1] **Itay Ebenspanger** [1*] **Neil Cnaan** [1*] **Mor Geva** [1]

摘要

为提升大语言模型（LLM）的知识蒸馏（KD）效果，人们日益倾向于用选择性蒸馏取代密集教师监督，即仅使用部分词元位置、词汇类别或训练样本进行监督。然而，哪些重要性信号和选择策略最有效，以及它们如何相互作用，仍不明确。在这项工作中，我们重新探讨自回归大语言模型中在哪里以及如何进行蒸馏。我们沿着位置、类别和样本轴拆解选择性知识蒸馏，并系统比较重要性信号与选择策略。随后，在这项分析的指导下，我们找出尚未充分探索的机会，并提出学生熵引导的位置选择（SE-KD）。在一系列基准测试中，与密集蒸馏相比，SE-KD 通常能提升准确率、下游任务遵循度和内存效率。将这一方法扩展至类别轴和样本轴（SE-KD$_{3X}$），可带来互补的效率提升，使教师模型离线缓存成为可行方案。在实践中，该方法将运行时间缩短 70%，峰值内存降低 18%，存储用量较以往方法减少 80%，同时不牺牲性能。

1. 引言

大型语言模型（LLMs）在各种任务中都取得了最先进的成果，但其规模庞大，导致服务成本高昂且难以适配。知识蒸馏（KD；Hinton 等人，2015年）通过训练较小的学生模型模仿较大的教师模型来解决这一问题。对于自回归大语言模型，通常的做法是在训练序列的每个位置上匹配教师模型的下一个词元分布。

然而，由于对所有位置采用统一监督，在每个词元位置应用知识蒸馏往往并非最优。近期研究表明，可以根据学生模型交叉熵等信号（Wang 等人，2021）、教师不确定性（Zhong 等人，2024；Huang 等人，2025）以及师生差异（Xie 等人，2025）来选择或重新加权位置，从而提升性能。然而，对于哪些词元重要性信号最能可靠地识别出受益于大语言模型基于 logit 的蒸馏的位置，以及不同的位置选择策略如何与这些信号相互作用以形成有效的蒸馏课程，目前仍不清楚。

在这项工作中，我们重新审视在知识蒸馏中以及如何为自回归大语言模型应用教师监督。我们首先将选择性知识蒸馏拆解为五个设计轴：对齐准则、位置、类别、样本和特征。在此框架下，我们聚焦于三个关键选择轴（图1）——位置、类别和样本，并系统分析：（i）位置重要性信号的选择，比较不确定性和基于差异的度量，例如熵和教师–学生 KL 散度；（ii）将这些信号转化为选择性监督所采用的策略，例如 top-k 选择、课程学习，以及固定预算下的随机分配；（iii）位置选择如何与沿类别轴和样本轴进行的稀疏化相互作用。

受本分析所揭示的空白启发，我们识别出两个尚未充分探索的机会：（1）使用学生熵作为位置重要性信号，以及（2）跨多个轴的联合选择。我们通过引入一种学生熵引导的位置选择式知识蒸馏方法（称为 SE-KD）及其三轴变体 SE-KD$_{3X}$来应对这些空白，该方法对样本、位置和类别进行选择（图 1D）。

通过在一系列广泛的基准测试上开展实验（涵盖 9 种重要性信号、5 种选择策略和 6 种知识蒸馏基线），我们发现基于学生不确定性的位置选择能够可靠地识别出对蒸馏具有高价值的词元。基于学生熵选择前 20% 的位置，可使平均评估准确率（64.8，而完整知识蒸馏为 64.4）和困惑度（6.9，而完整知识蒸馏为 7.3）获得稳定提升，同时仅需对一小部分词元位置进行监督。这些收益伴随着校准的一定下降（0.273 → 0.276），但显著降低了计算与内存开销，因为需要计算的教师逻辑值和学生模型逻辑值更少。

接下来，我们在另外两种

[*]共同贡献 [1]特拉维夫大学布拉瓦特尼克计算机科学与人工智能学院。联系作者：Almog Tavor<almogt@mail.tau.ac.il>。

预印本。*2026*年*2*月*3*日。

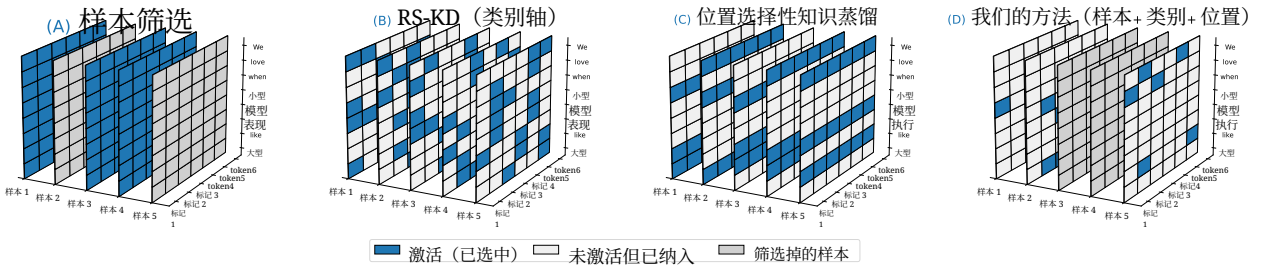

图 *1.* 知识蒸馏三种关键选择轴的示意图：**(A)** 样本选择、**(B)** 类别采样（RS-KD）、**(C)** 位置选择性知识蒸馏，以及**(D)** 我们提出的组合方法 SE-KD$_{3X}$，该方法结合了样本、类别和位置选择。蓝色单元格表示激活（选中）的监督信号，浅灰色表示未激活但仍包含的元素，深灰色表示筛选掉的样本。

同策略蒸馏的设置（Agarwal 等人，2024）上评估 SE-KD 和 SE-KD，其中监督应用于学生模型生成的轨迹；以及特定任务蒸馏，重点关注数学推理。我们发现，我们的方法在这两种设置下仍具有竞争力，这表明我们的方法能够推广到单一蒸馏范式之外。

最后，我们分析了位置选择性知识蒸馏与样本选择和类别采样结合时的效率提升。在 8000 万 token 蒸馏中，基于学生熵的样本选择将总墙钟时间缩短了 70%；按类别采样的离线缓存也因此在实践中变得可行，存储量减少了 99.96%，同时保持了性能。

总之，我们的工作做出了以下贡献：

- 我们提出了一个通用的选择性知识蒸馏理论框架，对已有方法进行梳理，并指出了尚未探索的变体。
- 我们提出了由学生熵引导的新型选择性知识蒸馏变体 SE-KD 和 SE-KD*3X*。我们证明，这些变体为知识蒸馏提供了往往表现最佳的信号，在位置选择方面优于以往的位置重要性指标，并带来最显著的准确率提升，同时保持下游任务遵循度。
- 我们证明，将 SE-KD 与类别和样本选择相结合，并借助离线教师缓存和选择感知型实现（选择性语言模型头和分块熵计算），能够大幅提升蒸馏效率，降低运行时间、峰值内存和存储开销。

我们的代码已发布于：
https://github.com/almogtavor/SE-KD3x。

2. 相关工作

采用位置选择的知识蒸馏 一项重要研究方向探索了如何通过仅在序列的部分位置选择性地施加监督来改进知识蒸馏。 Wang 等人 (2021) 使用批次内选择和全局级选择（GLS），选取学生交叉熵最高的前 $k\%$ 位置。近期，Huang 等人 (2025) 提出降低教师不支持的学生模型提议所对应位置的权重。与此同时，令牌自适应框架会根据教师与学生模型之间的分布差异，动态调整令牌级监督（Xie 等人，2025）。这些方法聚焦于单一的选择启发式或设置，大体遵循“二八”直觉：少数高熵“分叉”位置可能承载了大部分蒸馏信号（Wang 等人，2025）。我们的工作在此基础上进一步推进，在统一的蒸馏设置下，对位置选择策略和指标进行全面比较，从而确定哪些信号能够在不同任务中可靠地识别信息丰富的位置。

课程学习对于思维链蒸馏，Feng 等人(2024) 学习了位置重要性权重，并在给定预算下采用一种课程学习策略，将监督从较简单的推理步骤逐步扩展到较困难的步骤。受这项工作的启发，我们以两种方式引入课程学习：(1) 我们的学生熵选择会随着训练过程中受监督位置的变化，形成一种隐式课程学习；(2) 我们评估了一种显式的课程式位置选择方法。

不确定性引导的位置加权知识蒸馏 以往研究表明，不确定性加权蒸馏能够提升可靠性和校准效果（Guo 等人，2024）。近期，适应性教学知识蒸馏（AT-KD；Zhong 等人，2024）在解耦知识蒸馏（Zhao 等人，2022年）的基础上，利用教师的真实标签概率对词元级监督进行路由，$1 - p_t\,(y_t)$，其中 $p_t\,(y_t)$ 表示教师分配给真实下一个令牌的概率。对于每个批次，AT-KD 根据该不确定性分数对位置进行排序，并将其划分为简单和困难令牌：对简单令牌跳过目标类别 KL 项，同时强调困难令牌的多样性。不同于以往通过逐位置损失重加权来纳入不确定性的方法，我们仅将不确定性用作排序信号，以进行显式选择。

带类别采样的知识蒸馏 另一类互补研究致力于通过稀疏化教师输出分布来降低蒸馏成本。确定性 top-k 或按百分位截断教师 logits（Raman 等人，2023年；

Shum 等人，2024）可降低计算和存储成本，但会丢弃尾部概率质量，导致梯度估计产生偏差，并使学生模型的校准不佳。随机采样知识蒸馏（RS-KD；Anshumann 等人，2025）用重要性采样取代截断，从而提供无偏梯度估计并改善校准。这些研究聚焦于类别采样，而类别采样正是我们研究的选择轴之一。

样本选择与加权知识蒸馏 减少教师查询次数也能提高知识蒸馏效率。例如，UNIX（Xu 等人，2023年）采用不确定性感知采样，将蒸馏重点放在信息丰富的样本上。其他研究则侧重于通过样本选择来提高准确率。基于熵的自适应知识蒸馏根据教师和学生模型的熵，对知识蒸馏损失重新加权，优先处理样本（Su 等人，2023年）。近期，难度感知知识蒸馏（DA-KD）（He 等人，2025）通过教师与学生交叉熵损失之间的差异，明确衡量样本难度，并将其定义为交叉熵比率，$\mathcal{L}^{\mathrm{CE}}_{\mathrm{student}}(x)/\mathcal{L}^{\mathrm{CE}}_{\mathrm{teacher}}(x)$，并利用该分数进行难度感知分层采样，使蒸馏聚焦于难度高但信息丰富的样本，同时保持数据多样性。我们的工作同时考虑教师与学生以及仅基于学生的样本选择。

3. 选择性知识蒸馏框架

我们提出了一个用于选择性知识蒸馏的通用框架，该框架囊括了现有方法，并凸显了扩展这些方法的机会。随后，我们概述了实现该框架所涉及的关键设计选择。

问题设定 在知识蒸馏中，学生模型通过最小化一组输入上的下一词元分布之间的散度，经过训练来模仿教师模型。设 $x=(x_1,\ \ldots\ldots,\ x_L)$ 是一个包含L 个词元的输入样本，$\mathcal{V}$ 是教师与学生共享的词汇表。在每个位置 $t\in\{1,\ldots,L-1\}$，教师和学生分别在 $\mathcal{V}$上定义下一词元分布 $p_t(\cdot)=p(\cdot\mid x_{\le t})$ 和 $q_t(\cdot)=q(\cdot\mid x_{\le t})$。

标准的非选择性形式称为完整知识蒸馏，它对教师与学生的 KL 散度和真实标签交叉熵（交叉熵）的混合项进行优化，并在词元位置和训练样本上取平均。对于给定样本 i，位置 t处的蒸馏损失 $\ell^{(i)}_{\text{知识蒸馏}}(t)$ 定义为

$$\ell_{\mathrm{KD}}(t)=\lambda\,\mathrm{KL}(p_t\,\|\,q_t)+(1-\lambda)\,\mathrm{CE}(y_t,q_t),\qquad(1)$$

对于训练集 $\mathcal{D}$，整体目标函数为

$$\mathcal{L}_{\mathrm{KD}}=\frac{1}{|\mathcal{D}|}\sum_{i=1}^{|\mathcal{D}|}\frac{1}{L_i-1}\sum_{t=1}^{L_i-1}\ell^{(i)}_{\mathrm{KD}}(t),\qquad(2)$$

其中，$\ell^{(i)}_{\text{知识蒸馏}}(t)$表示样本 i在位置 t 处的损失，L_i 表示样本 i的长度。

因此，选择可以沿三个不同的轴进行：类别（在特定位置上）、位置（给定样本中的位置），以及样本（训练集中的样本）。令 $\mathrm{KL}_{\mathcal{C}_t^{(i)}}$ 表示在类别子集上计算的 KL 散度（其中 $\mathcal{C}_t^{(i)}\subseteq\mathcal{V}$，而$\mathcal{C}_t^{(i)}=\mathcal{V}$ 表示完整知识蒸馏）。此外，令 $m_t^{(i)}\in\{0,1\}$ 表示第 i个样本的位置 t 是否接受监督，并令 $s_i\in\{0,1\}$ 表示样本 i 是否被选中进行蒸馏。选择性知识蒸馏（SKD）的目标函数可写为：

$$\ell^{(i)}_{\mathrm{SKD}}(t)=\lambda\,\mathrm{KL}_{\mathcal{C}_t^{(i)}}(p_t\,\|\,q_t)+(1-\lambda)\,\mathrm{CE}(y_t,q_t)\qquad(3)$$

$$\mathcal{L}^{(i)}_{\mathrm{SKD}}=\frac{1}{\sum_{t=1}^{L_i-1}m_t^{(i)}}\sum_{t=1}^{L_i-1}m_t^{(i)}\,\ell^{(i)}_{\mathrm{SKD}}(t)\qquad(4)$$

$$\mathcal{L}_{\mathrm{SKD}}=\frac{1}{\sum_{i=1}^{|\mathcal{D}|}s_i}\sum_{i=1}^{|\mathcal{D}|}s_i\,\mathcal{L}^{(i)}_{\mathrm{SKD}}\qquad(5)$$

核心问题是如何为蒸馏选择类别、位置和样本，也就是如何构造 $\mathcal{C}_t$、$m_t^{(i)}$ 和 s_i。

选择性蒸馏的关键设计我们将选择性知识蒸馏分解为五个相互正交的设计选择，它们决定了教师信息如何传递给学生模型：

1. 对齐准则：用于教师与学生模型对齐的目标函数，例如基于 KL 散度的方法或解耦知识蒸馏。2. 位置轴：哪些词元位置接受蒸馏，即如何选择 $m_t^{(i)}$。我们通过以下两个方面研究这一轴：(i) 位置重要性指标 $u(t)$，用于量化每个位置 t的重要性；(ii) 位置选择策略，即一条将给定样本的得分 $u(t)$ 映射为数值 $m_t^{(i)}$ 的规则。3. 类别轴：如何在每个位置对词汇表上的教师分布进行稀疏化，即选择 $\mathcal{C}_t^{(i)}$。4. 样本轴：哪些训练样本接受蒸馏，即如何选择 s_i。5. 特征轴（本文未探讨）：哪些教师和学生模型表征需要对齐。除了下一个词元分布之外，知识蒸馏还可以对齐中间特征，例如隐藏状态或注意力图（Romero 等，2015年；Jiao 等，2020年）。虽然也可以沿此轴进行选择（例如，选择层或注意力头），但我们将这一方向留待未来研究。

表 1 从我们的框架角度总结了以往的选择性知识蒸馏方法。值得注意的是，我们发现此前没有研究在多个维度上进行选择。此外，学生熵作为蒸馏信号的作用尚未得到充分探索，

表1. **选择轴归属方面的选择性知识蒸馏方法概览。** 列 **位置、类别和样本** 分别表示方法是否沿位置轴、类别轴或样本轴进行选择/稀疏化。 ✓ 表示该方法明确作用于相应轴，而✗ 表示该方法不作用于相应轴。我们用绿色突出显示提出的学生熵变体。

	方法	描述	Pos	Cls	Smp
对齐准则	完整知识蒸馏（Hinton 等人，2015年）	对所有位置使用 KL散度/交叉熵（公式 1）	✗	✗	✗
	解耦知识蒸馏（Zhao 等人，2022年）	重新调整 KL 损失中目标项与非目标项的权重	✗	✗	✗
	AT-KD（Zhong 等人，2024）	将位置划分为简单/困难两类，并分别使用 KL 散度	✓	✗	✗
	加权知识蒸馏 (Guo 等人，2024)	使用教师的真实标签（y_t）概率 $1-p_t(y_t)$ 的术语 使用 $w_t \propto u(t)$ 在损失函数中重新加权逐位置KL散度	✓	✗	✗
位置-重要性指标	学生交叉熵（Wang 等人，2021）	学生交叉熵 CE(y_t, q_t)	✓	✗	✗
	教师交叉熵（Zhong 等人，2024）	教师交叉熵 CE(y_t, p_t)	✓	✗	✗
	学生熵	学生熵 $H(q_t)$	✓	✗	✗
	教师熵	教师熵 $H(p_t)$	✓	✗	✗
	KL散度 / 反向KL散度	教师–学生差异 KL散度($p_t\|q_t$) / KL散度($q_t\|p_t$)	✓	✗	✗
	KL散度 +学生熵	组合 差异 and 不确定性 排序 KL散度($p_t\|q_t$) + $H(q_t)$	✓	✗	✗
	交叉熵比率（He 等人，2025）	教师–学生交叉熵比率 $r(t)$ = 交叉熵(y_t, q_t)/交叉熵(y_t, p_t)	✓	✗	✗
	交叉熵比率 + 学生熵	综合难度与不确定性 $r(t) + H(q_t)$	✓	✗	✗
位置-选择策略	前k%	根据 $u(t)$选择排名前-k%的位置	✓	✗	✗
	GLS（Wang 等人，2021）	使用全局阈值 τ，跨批次归一化排名前-k%的位置	✓	✗	✗
	课程学习（Feng 等人，2024）	将监督从简单位置转移到困难位置，采用预定的调度-滑动窗口	✓	✗	✗
	位置 RS-KD / 位置 RS-KD*	加权/完整知识蒸馏的随机估计量 $q(t) \propto w_t$	✓	✗	✗
类别采样	RS-KD（Anshumann 等人，2025）	在位置 t，对 U 个索引 $v_k \propto p_t(v)$进行有放回采样。令 $\mathcal{C}_t = \{v_k\}_{k=1}^{U}$ 为唯一采样索引。构建一个稀疏教师目标 $\tilde{p}_t$，其定义在 $\mathcal{C}_t$ 上（基于采样计数），并最小化 $\sum_{v\in\mathcal{C}_t} \tilde{p}_t(v)\log(\tilde{p}_t(v)/q_t(v))$.	✗	✓	✗
采样选择	前-ℓ% 平均学生熵（Xu 等人，2023年）	选择样本 $\frac{1}{L_i-1}\sum_t H(q_t)$ 使用 学生模型 熵 U_i =	✗	✗	✓

尽管已有证据表明它在训练中有效（Wang 等人，2025）。接下来，我们将着手解决这些不足。

4. 学生熵引导的选择性知识蒸馏

鉴于先前工作的不足，我们提出一种选择性知识蒸馏方法，利用学生熵作为位置重要性信号，并沿多个轴进行选择。

基于学生熵的位置选择（SE-KD）我们使用学生熵对位置重要性进行评分，即$u(t) = H(q_t)$。给定一个长度为 L的样本 i，SE-KD会选择不确定性最高的k%位置进行蒸馏：

$$m_t^{(i)} = \mathbb{I}[u(t) \geq \tau], \tag{6}$$

其中，τ的取值使得恰有 $\lceil k(L-1)\rceil$ 个位置满足 $m_t^{(i)}$=1。此外，我们还在损失中采用按序列归一化，以确保监督预算固定。

跨轴选择 除位置选择外，我们将 SE-KD 扩展到类别、位置和样本这三个轴上。具体而言，我们使用 RS-KD，通过逐词元类别采样（$\mathcal{C}_t^{(i)}$）进行类别选择，并且

通过按平均学生模型熵进行 top-ℓ% 排序来实现样本选择；该熵在使用冻结学生模型的单次前向传播预处理步骤中计算，随后在 top-ℓ% 的样本上进行蒸馏。我们将此变体称为 SE-KD$_{3X}$。这些扩展与位置选择正交，从而实现统一的多轴知识蒸馏，提升准确率、效率并改善存储成本。

选择性语言模型头与分块熵计算选择性知识蒸馏支持两种简单的、选择感知的优化，以降低与 logits 相关的内存占用。设 B 表示批次大小，L 表示序列长度，$V = |\mathcal{V}|$ 表示词汇表大小。

首先，分块熵计算在不物化完整 $[B, L, V]$ 尺寸 logits 张量的情况下计算逐位置熵：学生隐藏状态在禁用梯度的情况下以小型分块通过语言模型头进行投影，归约为 $O(BL)$ 个熵标量后丢弃。

其次，选择性语言模型头仅在整个批次的这些位置计算逻辑值 N_{select}：教师 logits 从$[B, L, V]$ 缩减至 $[N_{\text{select}}, V]$；对于学生模型，它仅在选定位置计算逻辑值，并启用梯度，因此 KL 损失会通过 N_{select} 个位置进行反向传播，而非

th全部 BL，从而减少前向和反向 成本。

5. 实验

我们开展全面实验，沿§3所定义的维度评估选择性知识蒸馏方法。值得注意的是，即便在保守选择下，设计空间仍然很大，涵盖位置重要性指标、位置选择策略以及类别/样本选择，从而产生数百种配置，使穷尽评估不可行。因此，我们采用受控评估方案，每次固定除一个维度以外的所有维度。这使我们能够分离每种设计选择的影响。

方法 我们评估表1中的全部位置重要性指标与选择策略。除GLS外，位置选择始终按序列长度进行归一化。除非另有说明，所有模型均使用§B中所述的相同超参数进行训练。以下是关于位置选择策略的更多细节：

• **GLS**：维护一个近期熵值队列，并将τ 设为该全局分布的经验第（$100-k$）百分位数，以稳定跨批次的top-k 选择。

• **Pos RS-KD**：一种受 RS-KD 启发的随机位置选择策略，按概率采样位置$q(t) = {}^{H}{}_{(q_t)\sum_j H(q_j)}$。尽管此处将其视为一种选择策略，重复采样会产生隐式损失重加权，从而得到加权知识蒸馏的无偏估计量（见 §A）。

• **Pos RS-KD***：重要性校正变体：按概率 $q(t)$ 采样位置后，对每个采样位置的损失按 $1/q(t)$重新加权，从而得到完整知识蒸馏的无偏估计量。

•**课程学习**：一种课程式位置选择方法，每个序列的位置预算固定为 k=20%，并在训练过程中逐渐将监督从学生熵较低的词元转向学生熵较高的词元。

基线与消融实验 我们与以下基线和组件消融实验进行比较：

• 未经蒸馏的现成学生模型，以及作为上界的教师模型。

• **完整知识蒸馏**：在所有类别、位置和样本上密集应用监督式知识蒸馏。
• **AT-KD**：一种具有代表性的不确定性引导的位置加权方法。
• **RS-KD**：通过对教师 logits 进行重要性采样的类别轴选择性蒸馏。
• **随机位置** k%：随机选择位置，对每个样本中固定比例 k%的位置进行监督。
• **TopSmp** ℓ%：基于学生熵的样本选择。这是 SE-KD$_{3X}$ 的消融实验，移除了类别采样（RS-KD）和位置选择。
• **随机样本** ℓ%：随机选择样本，对固定比例 ℓ% 的训练样本进行监督。

我们将全局配置选择与最终评估分开。所有方法采用相同的蒸馏设置：知识蒸馏超参数（例如温度 $T = 1.0$ 和损失权重 $\lambda = 1.0$，从而得到仅含 KL 散度的蒸馏目标）仅在验证数据上选择一次，随后固定，不针对具体方法进行调优（详情见 §C）。top-k% 位置选择和 top-ℓ% 样本选择的监督预算通过在验证集划分上对每种设置运行一次进行搜索确定（结果见 §F），随后在所有主要比较中固定。

评估 我们考虑两种蒸馏设置：

(1)通用蒸馏使用大型预训练风格语料库。我们在 FineWeb-Edu 的 8000 万个词元上训练所有模型（Penedo 等人，2024年），并在零样本设置下对其进行评估。文档被打包成长度最多为 512 个词元的序列。我们在 HellaSwag（Zellers 等人，2019）、PIQA（Bisk 等人，2019）和 Arc-E（Clark 等人，2018）（多项选择推理）、GSM8K（Cobbe 等人，2021）（数学推理）以及 LAMBADA（Paperno 等人，2016）（长程预测）上衡量性能，并报告平均准确率。对于 LAMBADA，我们还报告困惑度和期望校准误差（ECE；Guo 等人，2017）。我们还在 IFEval（Zhou 等人，2023年）上评估指令遵循能力，并依据官方验证器报告 Pass@1[1]。所有结果均为三个随机种子的平均值，标准差见 §G。

(2) 特定任务蒸馏应用于下游推理任务。我们直接在 GSM8K 训练集上进行知识蒸馏（Cobbe 等人，2021），并报告在 GSM8K 测试集上的精确匹配准确率。除标准离策略蒸馏外，我们还评估同策略蒸馏（Agarwal 等人，2024）。本次评估不包含 SE-KD$_{3X}$，因为类别级采样依赖离线教师缓存，而该缓存与动态学生文本生成不兼容。

模型我们遵循先前的研究（Chen 等人，2025；Lu &Lab，2025），使用 Qwen3-1.7B 作为学生模型，使用 Qwen3-8B 作为教师（Yang 等人，2025）。

6. 结果

比较位置重要性指标 我们首先固定选择策略和预算为前 20%，并比较位置重要性指标。表 2展示了结果，表明基于学生熵的信号和教师–学生差异指标（交叉熵比率、KL散度和反向

[1]https://github.com/google-research/google-research/tree/master/ifeval

表 2. 各种**位置重要性指标在 Top-20% 硬选择下的评估结果**。最佳学生方法以**粗体**标出，次佳结果加下划线，粗斜体表示教师（上界）。标准差数值见§G。

方法	准确率 ↑	IFEval ↑	困惑度 ↓	期望校准误差 ↓
Qwen3 1.7B	61.9	19.4	12.2	30.5
Qwen3 8B	***73.8***	***28.9***	***4.6***	***23.5***
完整知识蒸馏	64.4	20.5	7.3	27.3
随机位置 20%	64.2	20.2	7.7	27.2
AT-KD	63.8	19.8	7.3	**26.7**
位置选择策略：前*20%*				
学生熵（SE-KD）	64.8	21.4	6.9	27.6
教师熵	63.2	20.5	9.4	27.3
学生交叉熵	63.8	20.4	8.1	27.8
教师交叉熵	63.4	19.4	9.3	27.8
KL	64.5	21.0	7.2	**26.7**
反向 KL	64.7	20.7	6.8	27.0
交叉熵比率	64.6	**22.5**	**6.5**	27.7
交叉熵比率 + 学生熵	64.6	21.4	6.7	27.5
学生熵 + KL散度	**65.1**	20.9	6.8	26.9

表 3. 将**位置选择策略与学生熵**作为位置重要性指标，并采用 20% 的蒸馏预算，与基线进行比较的评估结果。

方法	准确率 ↑	IFEval ↑	困惑度 ↓	期望校准误差 ↓
Qwen3 1.7B	61.9	19.4	12.2	30.5
Qwen3 8B	***73.8***	***28.9***	***4.6***	***23.5***
完整知识蒸馏	64.4	20.5	7.3	27.3
随机位置 20%	64.2	20.2	7.7	27.2
AT-KD	63.8	19.8	7.3	**26.7**
位置重要性指标：学生熵，$k = 20\%$				
前20% (SE-KD)	**64.8**	**21.4**	**6.9**	27.6
前20% GLS 30K	64.5	20.7	7.5	27.6
课程学习 20%	64.6	20.7	**6.9**	27.7
Pos RS-KD* 20%	63.6	20.6	8.3	27.6
Pos RS-KD 20%	63.0	20.1	9.9	27.0

KL）最可靠地识别出信息丰富的位置：选择学生熵最高的前 20% 位置，表现强劲（准确率 64。8，困惑度 6。9），优于完整知识蒸馏和随机位置；而选择 KL散度/反向KL散度/交叉熵比率最高的前 20% 位置也具有竞争力（准确率 64–64。7，最佳困惑度为 6。5）。相比之下，按教师熵/交叉熵排序，在准确率和困惑度两方面表现都较差。值得注意的是，校准差异很小；尽管 AT-KD、KL散度和反向 KL 的期望校准误差最低，但差距有限，这表明性能提升主要源于更合理的监督分配，而非置信度行为的变化。

比较位置选择策略 我们在固定重要性指标和预算下比较位置选择策略。如表3所示，按学生熵进行前 20% 选择（SE-KD）可带来最强的整体表现，提升准确率（64.4 → 64.8）、困惑度（7.3 → 6.9）和指令遵循（20.5 → 21.4）。在准确率和 IFEval 方面，它优于完整知识蒸馏、随机选择、GLS、课程调度

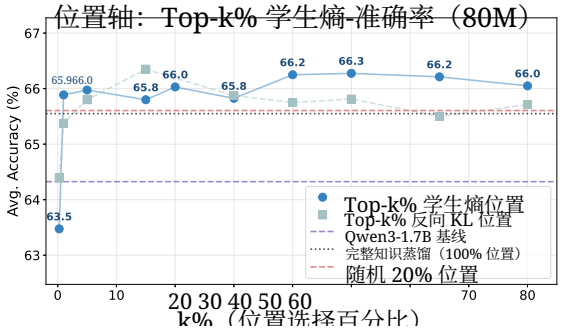

图 2. **位置轴预算扫描。**在 80M FineWeb-Edu 词元上进行蒸馏后，平均验证准确率随监督位置预算 k% 的变化。我们比较学生熵最高的 Top-k%（SE-KD）和反向KL散度最高的 Top-k%，并以完整知识蒸馏和 RandomPos 作为参照。教师准确率为 77.0。

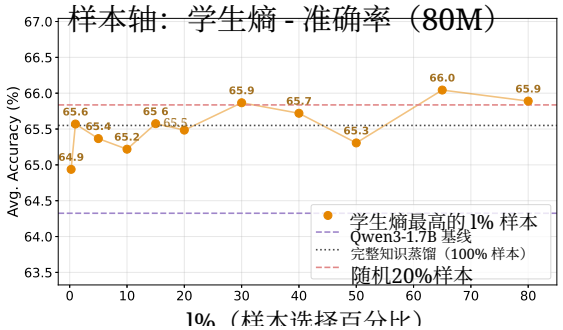

图 3. **样本轴预算扫描。** 在 80M FineWeb-Edu 词元上进行蒸馏后，平均验证准确率随样本选择预算 ℓ% 的变化。仅对按平均学生熵排序后位居前-ℓ% 的样本进行蒸馏；完整知识蒸馏和 RandomSmp 作为参照。教师准确率为 77.0。

和 AT-KD；不过，AT-KD 的校准效果最佳，其次是 Pos RS-KD，SE-KD 则排在最后。Pos RS-KD 和 Pos RS-KD* 的表现不如前 k%，这表明若不辅以额外平滑或覆盖约束，朴素的按熵比例采样可能并非最优（见 §D）。总体而言，在 k=20%的条件下，学生熵引导的选择是最可靠的位置选择策略，这支持了密集监督并非最优的观点。

蒸馏预算对性能的影响 图2和3 报告了验证集（ArcEasy、GSM8K、HellaSwag 和 PIQA）上经多次运行平均后的平均准确率。我们展示了 SE-KD 与反向KL散度（作为一种代表性的学生-教师差异度量）在不同选择预算下的性能。两种方法的最佳性能均在 k=20%（图2）处取得，这与近期发现一致，即约 20% 的高熵词元不成比例地驱动学习（Wang 等人，2025）。两种方法在较宽范围的 k 取值上均表现稳健，并在中等预算处呈现平缓最优；值得注意的是，监督比例少至 ∼ 1% 的位置

表4. **多轴选择性知识蒸馏**，将 SE-KD$_{3X}$ 与基线以及位置选择（SE-KD）、类别采样（RS-KD）和样本选择（TopSmp）的组合在通用蒸馏（测试集划分，80M 词元）上进行比较。我们报告各基准测试的平均准确率（准确率）、指令遵循性能（IFEval）、LAMBADA 困惑度（困惑度）和期望校准误差（期望校准误差）。三个随机种子上的标准差见§G。

方法	准确率 ↑	IFEval ↑	困惑度 ↓	期望校准误差 ↓
Qwen3 1.7B	61.9	19.4	12.2	30.5
Qwen3 8B	73.8	29.0	4.6	23.5
AT-KD	63.8	20.7	7.4	**26.7**
完整知识蒸馏	64.4	20.5	7.3	27.3
随机位置 20%	64.1	19.8	7.6	27.1
随机样本 20%	64.0	20.6	8.2	27.5
SE-KD	**64.8**	21.4	**6.9**	27.6
RS-KD	64.7	20.9	7.4	27.3
TopSmp 20%	64.2	20.8	7.4	27.8
TopSmp 20% +RS-KD	64.1	20.9	7.4	27.7
SE-KD +TopSmp 20%	64.6	**22.0**	6.9	28.0
SE-KD$_{3X}$	64.4	20.7	7.3	27.9

便已达到或超过完整知识蒸馏，而极小预算（例如 $\sim 0.25\%$）仍更接近未蒸馏基线。我们使用 k=20% 进行后续实验，作为较强的准确率–计算量权衡（接近平台期），并与先前“小比例”的发现保持一致（Wang 等人，2025）。我们还改变样本轴预算，仅蒸馏前ℓ% 按平均学生熵排序的样本（图3）。准确率随 ℓ 变化很小，而计算量大致呈线性增长，因此我们使用 ℓ=20% 进行多轴实验。

跨位置、类别和样本的选择 表 4 比较了沿位置、类别和样本轴进行的选择性知识蒸馏。位置选择是性能的主要贡献因素；我们的学生熵 SE-KD 将平均准确率从 64.4（完整知识蒸馏）提升至 64.8，提高了指令遵循能力（21.4 对比 20.5），并降低了 PPL（6.9 对比 7.3），同时期望校准误差略有增加（27.6 对比 27.3）。RS-KD 提高了准确率并保持了校准效果，而 TopSmp 的整体表现仍接近完整知识蒸馏，但会降低校准效果。结合所有轴后，SE-KD$_{3X}$ 实现了具有竞争力的性能（准确率 64.4、IFEval 20.7、PPL 7.3），校准效果略差，但显著降低了运行时间、内存和存储开销（见 §7）。

通用蒸馏与特定任务蒸馏 表 5报告了 GSM8K 上的特定任务蒸馏结果，其定性表现与 FineWeb-Edu 上的通用蒸馏不同。在离策略设置下，完整知识蒸馏取得了最高的 GSM8K 准确率（71.6），而基于熵的 Top-20%位置选择则导致性能下降（69.5）。我们表现最强的方法 SE-KD +TopSmp 在监督大幅减少的情况下，仍接近完整知识蒸馏（70.9）。在同策略设置下，SE-KD + TopSmp 达到了最高的

表 5. **GSM8K 上特定任务蒸馏的结果**。我们比较了离策略与同策略知识蒸馏方法，报告了 GSM8K 精确匹配准确率、评测套件平均准确率以及 LAMBADA OAI 困惑度。对于同策略蒸馏，我们使用了反向 KL 对齐准则。标准差见 §G。

	方法	GSM8K 准确率 ↑	准确率 ↑	困惑度 ↓
	Qwen3 1.7B	68.2	61.9	12.2
	Qwen3 8B	***87.8***	***73.8***	***4.6***
Off Policy Distill.	完整知识蒸馏	**71.6**	**64.5**	**7.8**
	随机位置 20%	70.2	64.0	8.0
	SE-KD	69.5	63.9	8.0
	Pos RS-KD 20%	70.5	63.5	8.9
	Pos RS-KD* 20%	69.1	63.3	9.2
	TopSmp 20%	69.0	63.6	8.6
	SE-KD +TopSmp 20%	70.9	64.0	8.6
	SE-KD$_{3X}$	70.2	63.9	8.6
On Policy Distill.	完整知识蒸馏	70.6	**63.7**	10.0
	随机位置 20%	69.3	63.3	10.0
	SE-KD	70.0	**63.7**	**9.5**
	Pos RS-KD 20%	70.5	63.2	10.5
	Pos RS-KD* 20%	69.7	63.3	10.0
	TopSmp 20%	70.4	**63.7**	10.1
	SE-KD +TopSmp 20%	**71.2**	63.4	10.4

GSM8K 准确率（71.2），优于完整知识蒸馏（70.6），而平均准确率差异仍然很小。

在在线策略设置中，将熵引导的位置选择与样本筛选相结合可取得最佳结果。然而，在离策略设置中，与通用蒸馏不同，单独使用熵引导的位置选择并不能始终在 GSM8K 上优于完整知识蒸馏。相反，尽管使用的监督数据少得多，我们的方法在单轮训练后仍与完整知识蒸馏表现相近。我们认为，这在一定程度上归因于 GSM8K 的规模有限，这可能会限制选择性蒸馏的收益；而在更大型的数据集或多轮次训练中，这些收益或许会更明显。我们将这一假设留待未来工作验证。

7. 知识蒸馏效率

选择性知识蒸馏的一项主要动机是降低计算成本。因此，我们从教师监督的离线存储 和蒸馏过程中的运行时计算 两个方面分析蒸馏效率。我们表明，位置选择主要提升准确率，样本级选择则能显著提高效率，而类别级采样可将存储需求降低数个数量级。

7.1. 存储效率

我们遵循Anshumann 等人（2025）的表述，重点关注类别选择和样本选择带来的节省。我们不考虑位置选择，因为它需要存储动态不确定性掩码（见§E）。

表 6. **离线缓存占用**（太字节，TB）：针对 N=100B 个训练词元和词汇表大小为 $|\mathcal{V}|$=100,000。RS-KD 对类别进行重要性采样；SE-KD$_{3X}$ 则通过以 ℓ=20% 进行样本级选择，进一步减少存储占用。

方法	类别	TB (U=12)	TB (U=64)
完整知识蒸馏	$\lvert\mathcal{V}\rvert = 100K$	10,000.0	10,000.0
RS-KD	U	3.6	19.2
SE-KD$_{3X}$	$U \times 0.2$	0.72	3.84
原始交叉熵	1	0.3	0.3

表7. **样本选择方法的运行时间和测试准确率**（8000万词元，前20%，单次运行），使用 GeForce RTX 3090。

方法	样本选择	总墙钟时间	Acc.
完整知识蒸馏（100% 位置）	0小时00分	22h52m	**64.6**
随机位置 20%	0h00m	18h38m	64.1
TopSmp 交叉熵比例	8h50m	13小时36分钟	64.3
TopSmp KL	9小时37分钟	14小时42分钟	64.2
TopSmp 学生熵（我们的方法）	**2小时01分钟**	**7小时01分**	64.2
SE-KD$_{3X}$ (缓存构建)	2小时11分	8小时46分	64.4
SE-KD$_{3X}$ (离线缓存复用)	0小时00分	3小时58分	64.4

存储量以每个词元的字节数衡量，并以十进制太字节（TB）报告，针对包含 N=100B 个词元的数据集。存储单个采样的教师类别需要 24 位（3 字节）：17 位用于词汇表索引，7 位用于量化概率，因此缓存 $U = |\mathcal{C}_t|$ 个采样类别，每个位置需要 $3U$ 字节。

表 6 总结了完整知识蒸馏、RS-KD 和 SE-KD$_{3X}$ 的存储占用。作为基线，我们加入了不使用教师 logits 的原始交叉熵训练。与使用 U=12 的 Anshumann 等（2025）不同，我们使用 U=64 以提升稳定性，得到 $64 \times 3 = 192$ 字节 / 位置。对于大小为 $|\mathcal{V}|$=100,000 的词汇表，以 float16 缓存完整教师 logits 每个位置需要 200 kB，因此使用 U=64 的 RS-KD 存储效率约高出 10^3–2×10^3 倍；对于 N=100B 个词元，则需要 19.2 TB。采用样本选择后，我们仅在蒸馏前对冻结学生模型进行一次前向传播，并根据平均学生熵排序，选取排名前 ℓ% 的样本进行蒸馏。这样，存储量会随 ℓ 线性减少。对于 ℓ=0.2，所需存储为：$\ell \cdot U \cdot 3 = 38.4$ 字节/位置，总计 3.84 TB。

总体而言，RS-KD 将存储量从 10,000 TB 降至 19.2 TB（减少 99.8%，$\sim 520\times$），而 SE-KD$_{3X}$ 进一步将其降至 3.84 TB（相比完整知识蒸馏减少 99.96%，相比 RS-KD 减少 80%）。样本索引也会缓存，但所需存储量可以忽略不计。

7.2. 运行效率

运行时间加速比 SE-KD$_{3X}$ 通过样本选择实现了显著的效率提升，直接减少了需要教师监督的序列数量。如表 7所示，这显著缩短了总实际运行时间。样本选择需要预先进行评分：教师–学生指标需要完整遍历数据（交叉熵比率需 8 小时 50 分钟，KL散度需 9 小时 37 分钟），而仅使用学生模型熵的成本较低（2 小时 01 分钟）。

重用选定索引的离线缓存可省去此步骤，将 SE-KD$_{3X}$的运行时间缩短至 3 小时 58 分钟（表 7）。对于包含 N 个样本、平均长度为 L 的训练集，完整知识蒸馏会监督 $\mathcal{O}(NL)$ 个位置；采用前 $\mathcal{O}(\ell NL)$ % ℓ的样本选择后，减少至 。因此，样本选择带来了主要的效率提升，而位置选择则进一步加快了运行速度（使用选择性语言模型头和分块熵时，最高可达 $\sim$30%；见 §H）。

SE-KD 的内存节省 位置选择主要是在序列内 重新分配知识蒸馏信号。尽管学生模型和教师仍然会处理完整上下文，但选择会减少需要进行 logit 计算和携带梯度的知识蒸馏项计算的位置数量。这使得面向内存优化的实现成为可能，大幅降低与 logits 相关的峰值内存占用。

在我们的设置中（B=2，L=512），在 k=20% 下使用分块熵的选择性语言模型头，使每个 GPU 的峰值内存分配总和减少 18.3%（33.18 GB $\rightarrow$ 27.10 GB）：学生模型峰值降低 28.1%（15.88 GB $\rightarrow$ 11.42 GB），教师峰值降低 9.4%（17.30 GB $\rightarrow$ 15.68GB）。这些收益来自避免在选择过程中物化完整 logits，并将知识蒸馏 logits 和反向传播限制在 $[B, L, V]$ 选定位置。消融实验和内存轨迹见 §H。

8. 结论与讨论

我们通过一个统一框架重新审视自回归大语言模型的选择性知识蒸馏，将教师监督应用于何处以及如何应用这两个问题加以区分。通过系统研究，我们发现，密集且均匀的 logit 监督往往并非必要：对于通用蒸馏，将监督集中于一小部分高不确定性位置，其效果始终可与完整知识蒸馏相当，甚至更胜一筹。

学生熵引导的前20%选择是整体上最可靠的策略，而课程学习、交叉熵比率排序和教师–学生 KL 散度则是很有前景的替代方案。我们还表明，位置选择可以与类别级和样本级稀疏化有效结合，从而实现有利的准确率–效率权衡；特别是，SE-KD$_{3X}$可通过样本筛选和离线教师缓存显著加速，并可借助选择性语言模型头降低峰值内存。

局限性与未来工作 选择性知识蒸馏的设计空间很大；为控制比较条件，我们研究了一组单一且广泛使用的教师–学生对，并采用固定的监督预算。下一步的重要工作是，在更多模型家族、不同规模和更长的上下文中验证这些趋势。选择性策略也可能与其他对齐标准（例如基于特征的知识蒸馏）相互作用；而我们观察到任务特定蒸馏中的性能下降幅度较小，这表明仍需进一步优化。

影响声明

本文旨在推动大语言模型的知识蒸馏研究。除训练和部署语言模型通常涉及的考量外，我们未发现本研究特有的社会影响。

参考文献

Agarwal, R., Vieillard, N., Zhou, Y., Stanczyk, P., Garea, S. R., Geist, M. 和 Bachem, O. 大语言模型的同策略蒸馏：从自生成错误中学习。载于 第十二届国际学习表征会议，2024 年。网址：https://openreview.net/forum?id=3zKtaqxLhW。

Anshumann, A., Zaidi, M. A., Kedia, A., Ahn, J., Kwon, T., Lee, K., Lee, H. 和 Lee, J. 稀疏 logit 采样：加速大语言模型中的知识蒸馏。*arXiv* 预印本 *arXiv:2503.16870*，2025 年。网址：https://arxiv.org/abs/2503.16870。

Bisk, Y., Zellers, R., Bras, R. L., Gao, J. 和 Choi, Y. PIQA：在自然语言中推理物理常识，2019 年。网址：https://arxiv.org/abs/1911.11641。

Chen, W.-R., Kothapalli, V., Fatahibaarzi, A., Sang, H., Tang, S., Song, Q., Wang, Z. 和 Abdul-Mageed, M. 提炼精髓：通过序列截断实现高效推理蒸馏，2025 年。网址：https://arxiv.org/abs/2512.21002。

Clark, P., Cowhey, I., Etzioni, O., Khot, T., Sabharwal, A., Schoenick, C. 和 Tafjord, O. 你以为自己解决了问答问题? 来试试 ARC，即 AI2推理挑战。载于 第 *56* 届计算语言学协会年会论文集（*ACL*），2018 年。网址：https://aclanthology.org/P18-1260。

Cobbe, K., Kosaraju, V., Bavarian, M., Chen, M., Jun, H., Kaiser, L., Plappert, M., Tworek, J., Hilton, J., Nakano, R., Hesse, C. 和 Schulman, J. 训练验证器以解决数学文字题。*CoRR*，abs/2110.14168，2021 年。网址：https://arxiv.org/abs/2110.14168。

Feng, K., Li, C., Zhang, X., Zhou, J., Yuan, Y. 和 Wang, G. 面向大语言模型的基于关键点的渐进式思维链蒸馏。*arXiv* 预印本 *arXiv:2405.16064*，2024 年。网址：https://arxiv.org/abs/2405.16064。

Guo, C., Pleiss, G., Sun, Y. 和 Weinberger, K. Q. 现代神经网络的校准。载于第 *34* 届国际机器学习会议论文集（*ICML*），第 70 卷，收录于机器学习研究论文集，第 1321–1330 页。PMLR，2017 年。网址：https://proceedings.mlr.press/v70/guo17a.html。

Guo, Z., Wang, D., He, Q. 和 Zhang, P. 利用 logit 不确定性提升知识蒸馏效果。*ScientificReports*，14(31249)，2024。doi：10.1038/ s41598-024-82647-6。网址：https://www.nature.com/articles/s41598-024-82647-6。

He, C., Ding, Y., Guo, J., Gong, R., Qin, H. 和 Liu, X. DA-KD：面向高效大语言模型的难度感知知识蒸馏。载于第四十二届国际机器学习会议，2025。网址：https://openreview.net/forum?id=NCYBdRCpw1。

Hinton, G., Vinyals, O. 和 Dean, J. 神经网络中的知识蒸馏。*arXiv*预印本 *arXiv:1503.02531*，2015。网址：https://arxiv.org/abs/1503.02531。

Huang, H., Song, J., Zhang, Y. 和 Ren, P. SelectKD：面向大语言模型的选择性词元加权知识蒸馏，2025。网址：https://arxiv.org/abs/2510.24021。

Jiao, X., Yin, Y., Shang, L., Jiang, X., Chen, X., Li, L., Wang, F. 和 Liu, Q. TinyBERT：面向自然语言理解的 BERT 蒸馏。载于 Cohn, T., He, Y. 和 Liu, Y.（编），计算语言学协会论文集：*EMNLP 2020 Findings*，第 4163–4174 页，线上举行，2020年11月。计算语言学协会。doi：10.18653/v1/2020.findings-emnlp.372。网址：https://aclanthology.org/2020.findings-emnlp.372/。

Lu, K. 和 Lab, T. M. 同策略蒸馏。*Thinking Machines Lab*：联结主义，2025。doi：10.64434/tml. 20251026。https://thinkingmachines.ai/blog/on-policy- distillation。

Paperno, D., Kruszewski, G., Lazaridou, A., Pham, Q., Bernardi, R., Pezzelle, S., Baroni, M., Boleda, G. 和 Fernández, R. LAMBADA数据集：需要广泛篇章上下文的词语预测。载于第*54*届计算语言学协会年会论文集（*ACL*），2016。网址：https://aclanthology.org/P16-1144。

Penedo, G., Kydlíˇcek, H., Ben Allal, L., Lozhkov, A., Mitchell, M., Raffel, C., Von Werra, L. 和 Wolf, T. Fine Web数据集：大规模筛选网络内容以获得优质文本数据。*arXiv*预印本 *arXiv:2406.17557*，2024。网址：https://arxiv.org/abs/2406.17557。

Raman, N., Vare, S., Srinivasan, A., Chandra, V. 和 Khandelwal, K. 对于蒸馏，词元并非全部所需。OpenReview，2023。网址：https://openreview.net/pdf?id=2fc5GOPYip。

Romero, A., Ballas, N., Kahou, S. E., Chassang, A., Gatta, C. 和 Bengio, Y. FitNets：瘦深度网络的提示。

载于 Bengio, Y. 和 LeCun, Y.（编），第三届国际学习表征会议（*ICLR 2015*），美国加利福尼亚州圣迭戈，*2015*年*5*月*7*日至*9*日，会议论文集，2015年。网址：http://arxiv.org/abs/1412.6550。

Shum, K., Xu, M., Zhang, J., Chen, Z., Diao, S., Dong, H., Zhang, J., and Raza, M. O. FIRST：通过高效的可信蒸馏，训练可靠的大语言模型。载于 *2024*年自然语言处理实证方法会议（*EMNLP*）论文集，第 12646–12659 页。计算语言学协会，2024。网址 https://aclanthology.org/2024.emnlp-main.703.pdf。

Su, C.-W., Tseng, S.-H., Martins, J.V., Ichimura,N., Seiji, Y., and Chou, C.-H. EA-KD：基于熵的自适应知识蒸馏。*arXiv* 预印本 *arXiv:2311.13621*，2023年。网址 https://arxiv.org/abs/2311.13621。

Wang, F., Yan, J., Meng, F. 和 Zhou, J. 面向神经机器翻译的选择性知识蒸馏。*arXiv* 预印本 *arXiv:2105.12967*，2021。网址：https://arxiv.org/abs/2105.12967。

Wang, S., Yu, L., Gao, C., Zheng, C., Liu, S., Lu, R., Dang, K., Chen, X., Yang, J., Zhang, Z., Liu, Y., Yang, A., Zhao, A., Yue, Y., Song, S., Yu, B., Huang, G. 和 Lin, J. 超越二八法则：高熵少数令牌推动大语言模型推理的有效强化学习。*arXiv* 预印本 *arXiv:2506.01939*，2025。网址：https://arxiv.org/abs/2506.01939。

Xie, X., Xue, Z., Wu, J., Li, J., Wang, Y., Hu, X., Liu, Y. 和 Zhang, J. 面向大语言模型的令牌自适应知识蒸馏，2025。网址https://arxiv.org/abs/2510.11615。

Xu, G., Liu, Z.和Loy, C. C. 通过不确定性感知混合实现计算高效的知识蒸馏。模式识别，138:109338，2023年。ISSN 0031- 3203。doi:https://doi.org/10.1016/j.patcog.2023.109338。URL
h
t
t
ps://www.sciencedirect.com/science/article/pii/S0031320323000390。

Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., Zheng, C., Liu, D., Zhou, F., Huang, F., Hu, F., Ge, H., Wei, H., Lin, H., Tang, J., Yang, J., Tu, J., Zhang, J., Yang, J., Yang, J., Zhou, J., Zhou, J., Lin, J., Dang, K., Bao, K., Yang, K., Yu, L., Deng, L., Li, M., Xue, M., Li, M., Zhang, P., Wang, P., Zhu, Q., Men, R., Gao, R., Liu, S., Luo, S., Li, T., Tang, T., Yin, W., Ren, X., Wang, X., Zhang, X., Ren, X., Fan, Y., Su, Y., Zhang, Y.,Zhang,Y., Wan, Y., Liu, Y., Wang, Z., Cui, Z., Zhang, Z., Zhou, Z., 和 Qiu, Z. Qwen3 技术报告，2025。URL https://arxiv.org/abs/2505.09388。

Zellers, R., Bisk, Y., Farhadi, A.和Choi, Y. Hellaswag：机器真的能补全你的句子吗？载于第*57*届计算语言学协会年会（*ACL*）论文集，2019年。URLhttps://aclanthology.org/P19-1472。

Zhao, B., Cui, Q., Song, R., Qiu, Y.和Liang, J. 解耦知识蒸馏。载于*IEEE/CVF*计算机视觉与模式识别会议（*CVPR*）论文集，第11953–11962页，2022年。URLhttps://arxiv.org/abs/2203.08679。

Zhong, Q., Ding, L., Shen, L., Liu, J., Du, B., and Tao, D. 重新审视自回归语言模型的知识蒸馏。 *arXiv* 预印本 *arXiv:2402.11890*，2024。网址 https://arxiv.org/abs/2402.11890。

Zhou, J., Lu, T., Mishra, S., Brahma, S., Basu, S., Luan, Y., Zhou, D., and Hou, L. 面向大型语言模型的指令遵循评估，2023年。链接 https://arxiv.org/abs/2311.07911。

A. 证明：位置 上的随机采样选择是加权知识的无偏估计量 D

在本节中，我们证明，采用位置随机采样选择方法的知识蒸馏在期望上与加权知识蒸馏相匹配。需要注意的是，可以轻松地将这种选择方法转换为在期望上与完整知识蒸馏相匹配，但我们有意不这样做，因为我们的目标是匹配一种根据词元的熵对其加权的加权知识蒸馏。

考虑一个包含 N个词元位置的序列，并以 $t \in \{1, \ldots, N\}$为索引。令 L_t表示位置 t处的每令牌蒸馏损失，令 $w(t)$为分配给该词元的非负重要性权重。我们从以下分布中独立同分布地采样 K 个词元索引 t_k：

$$q(t) = \frac{w(t)}{\sum_{j=1}^{N} w(j)}.$$

这里， $q(t)$表示词元位置上的采样分布， $\mathcal{L}$ 是经验损失估计量， $\mathbb{E}[\cdot]$ 表示对采样过程取期望。

使用这一记号，我们有

$$\begin{aligned}\mathbb{E}[L_{t_k}] &= \sum_{t=1}^{N} q(t) L_t = \sum_{t=1}^{N} \frac{w(t)}{\sum_{j=1}^{N} w(j)} L_t \\ &= \mathcal{L}_{\text{weighted}}.\end{aligned}$$

因此，采样 token t的概率与其对加权知识蒸馏目标的贡献成正比。

$$\begin{aligned}\mathbb{E}[\widehat{\mathcal{L}}] &= \mathbb{E}\left[\frac{1}{K}\sum_{k=1}^{K} L_{t_k}\right] = \frac{1}{K}\sum_{k=1}^{K} \mathbb{E}[L_{t_k}] \\ &= \frac{1}{K} \cdot K \cdot \mathcal{L}_{\text{weighted}} = \mathcal{L}_{\text{weighted}}.\end{aligned}$$

也可以将 c_t 视为 token t 被采样的次数：

$$c_t = \sum_{k=1}^{K} \mathbf{1}_{t_k = t}, \quad \widehat{\mathcal{L}} = \frac{1}{K}\sum_{t} c_t L_t,$$
$$\mathbb{E}[c_t] = K q(t).$$

So,

$$\begin{aligned}\mathbb{E}[\widehat{\mathcal{L}}] = \frac{1}{K}\sum_{t} \mathbb{E}[c_t] L_t &= \frac{1}{K}\sum_{t} K q(t) L_t \\ &= \sum_{t} q(t) L_t.\end{aligned}$$

因此， $\mathcal{L}$是**加权知识蒸馏目标**的无偏估计量：

$$\mathcal{L}_{\text{weighted}} := \sum_{t=1}^{N} q(t)\, L_t = \frac{\sum_t w(t)\, L_t}{\sum_j w(j)}.$$

重要性校正的位置随机采样。 为完整起见，我们指出，通过重要性校正，也可以使位置随机采样成为完整的无加权知识蒸馏目标的无偏估计量。具体而言，若将每个采样损失按其采样概率的倒数重新加权，$\widehat{\mathcal{L}}_{\text{IC}} = \frac{1}{KN}\sum_{k=1}^{K} \frac{L_{t_k}}{q\ t_k()}$，那么 $\mathbb{E}[\widehat{\mathcal{L}}_{\text{IC}}] = \frac{1}{N}\sum_{t=1}^{N} L_t$, 便可在期望意义上恢复完整知识蒸馏。我们将这种变体称为重要性校正的位置随机采样 ，并在实验中单独评估了它。

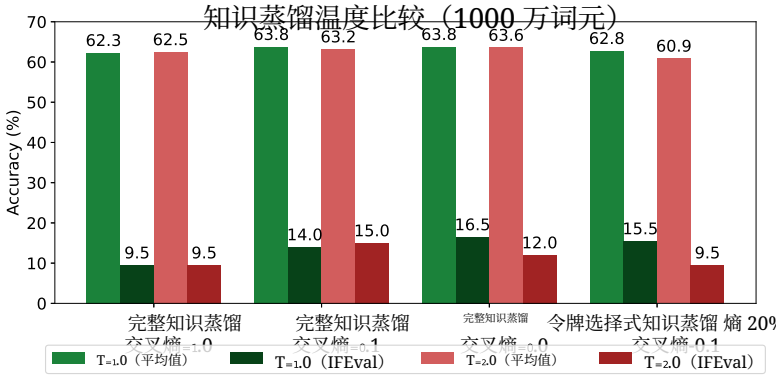

图*4.* **完整知识蒸馏的温度消融**。 我们比较 T=2.0 与 T=1.0 ，并报告五个基准测试（ArcEasy、GSM8K、HellaSwag、PIQA和 LAMBADA OpenAI）上的平均准确率。

B. 超参数

表8 和9 列出了所有运行共享的超参数选择，以及不同蒸馏变体之间有所差异的设置。

组件	值
教师模型	Qwen/Qwen3-8B（在线，未量化）
学生模型	Qwen/Qwen3-1.7B
数据集	FineWeb-Edu 流（8000 万词元）
序列长度	512 个词元（最大_序列_长度=512）
训练轮次	遍历流式子集 1 次
小批次	批次大小 2×8梯度累积步数（有效批次大小 16）
优化器	bitsandbytes 8位 Adam（学习率 1×10^{-5}）
知识蒸馏温度	1.0
交叉熵混合权重	$\alpha_{\text{CE}} = 1 - \lambda = 0.0$ （纯 KL 散度损失）
离线缓存	启用，缓存类别为 $U = 64$
随机种子	1337、1338、1339（GSM8K 设置中则为 1340、1341、1342）

表*8*。 所有实验的共享超参数。

变体	附加设置
完整知识蒸馏	蒸馏所有词元（$k = 100\%$）。
SE-KD（学生熵排名前 k）	$k = 20\%$；按序列长度归一化选择
课程学习	选择课程学习步骤= 4000。
随机令牌选择	$k = 20\%$；均匀随机令牌选择，并按长度归一化。
Pos-RS-KD	$k = 20\%$；学生熵评分；POS_RS_MATCH_FULL_KD= 1 用于校正后的变体。

表 *9.* 主表中报告的各蒸馏变体的具体设置。

C. 补充结果

本节报告辅助实验，以说明本文中所采用的超参数设置。我们比较了完整知识蒸馏基线的温度设置和交叉熵混合权重。在这些消融实验中，温度 T=1.0大多优于更高的温度，而交叉熵分量带来的益处微乎其微；此外，加入交叉熵会妨碍我们采用部分基于选择的效率优化（例如，将携带梯度的逻辑值限制在选定位置）。因此，在所有主要实验中，我们采用 T=1.0 ，并设定 λ=1 在公式 1 中使用（纯 KL 散度）。

D. 位置随机采样表现欠佳

图 6 展示了确定性 Top-k%位置选择与相同预算下的位置随机采样（Pos RS-KD）之间的差异。尽管 Pos RS-KD 根据不确定性-

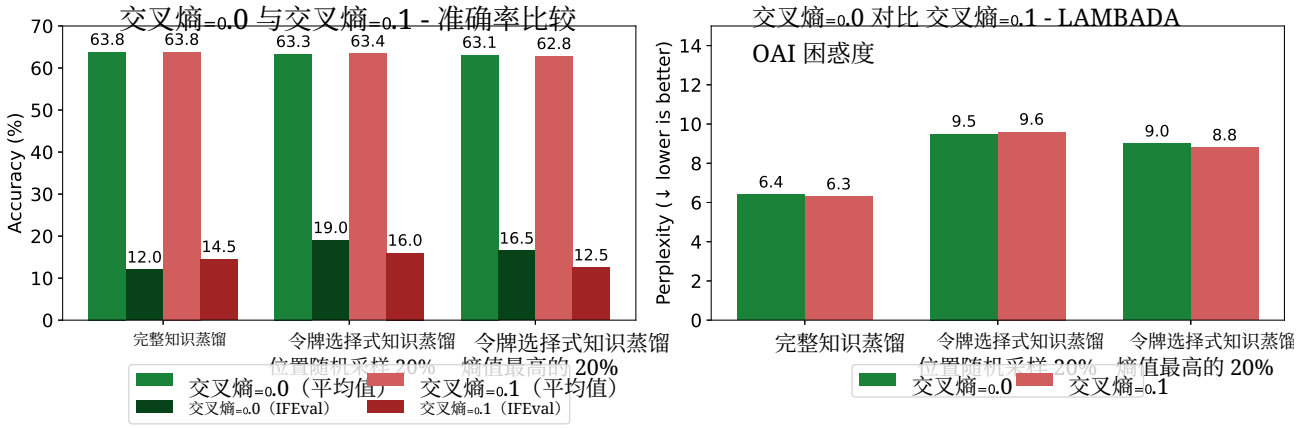

图 *5.* **完整知识蒸馏的交叉熵混合消融实验。** 我们比较 $\alpha_{\mathrm{CE}} = 1 - \lambda$=0.1 与 α_{CE}=0.0，并报告相同的平均准确率指标。本研究采用较小规模的 1000 万令牌运行，作为合理性检查，而非完全收敛的比较。

导出的权重引入随机性，但在通用设置下，其准确率和校准表现均不如 Top-k （表 3）。

一种可能的解释是序列内部的熵质量集中：如果逐序列熵分布高度尖峰化，那么按熵比例采样可能会将很大一部分预算分配给一小组极端熵位置（通常靠近序列开头）。这会降低其他信息丰富的位置的覆盖范围，而 Top-k 会确定性地纳入这些位置，并且可能增加不同更新之间的方差。

未来有几种简单的缓解措施可能会改善 Pos RS-KD：(i) 对采样分布进行温度平滑 （例如，从 $\propto H(q_t)^{1/T}$ 中采样，并使用 $T > 1$），以平缓过度尖峰化的序列；(ii) 采用轻量级启发式方法，例如排除前几个符合条件的位置，或裁剪极端熵值；以及 (iii) 将按熵比例采样与一个小型的确定性“覆盖范围”组件相结合（例如，预留部分预算用于 Top-k% 并对剩余部分进行采样）。我们将这些变体的系统性研究留待未来工作。

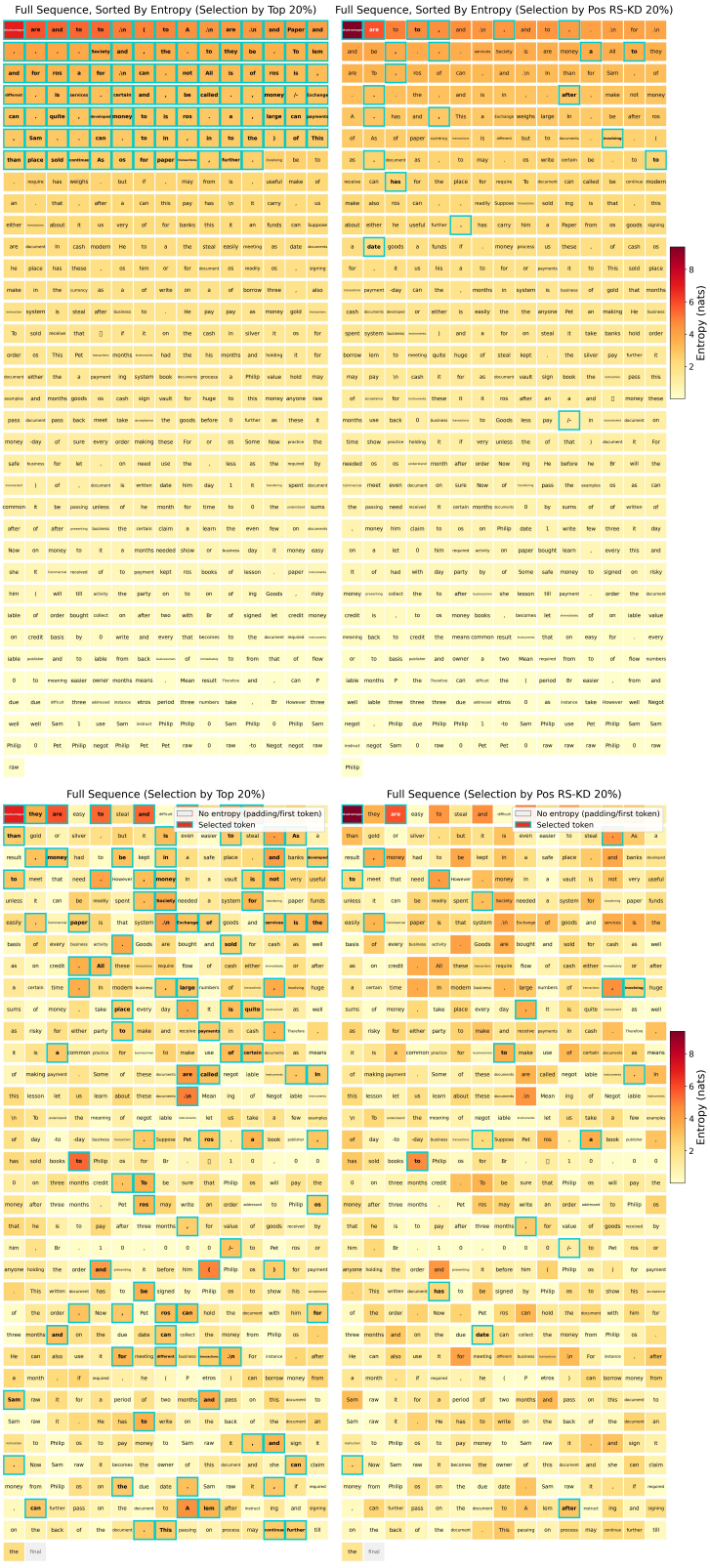

图 *6.* **Top-k 与固定预算下的位置随机采样对比**（k=20%）。词元按学生熵着色；青色轮廓标出选定位置。**上排：**同一序列按熵排序，突出每种策略如何在熵分布上分配其预算。**下排：**原始词元顺序（填充显示为“无熵”），展示选择如何沿序列分布。**左：**确定性 Top-k；**右：**熵比例位置 RS-KD（Pos RS-KD / Pos RS-KD*）。

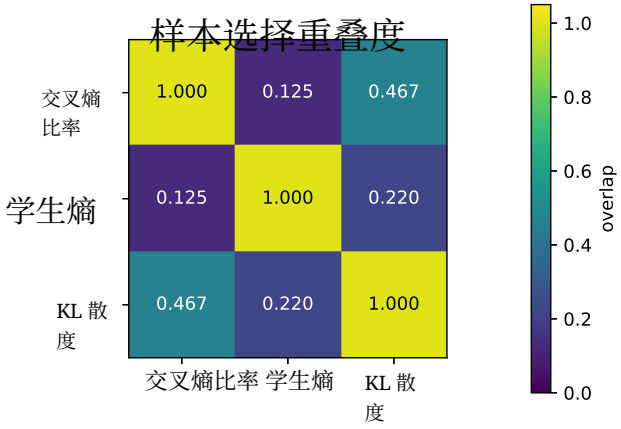

图 7. **不同指标间的样本选择重叠情况。** 根据学生熵、交叉熵比率和 KL 散度选出的样本之间的两两重叠情况。基于教师的指标之间相互重叠度更高，而与学生熵的重叠度较低，这表明选择稳定性存在差异。

E. 离线缓存权衡

SE-KD$_{3X}$ 的缓存占用还可以进一步降低：仅存储固定子集的选定位置对应的教师逻辑值，使存储量随位置预算 k% 缩放。然而，这会引入根本性的权衡：位置级缓存可最大限度地节省存储，但会破坏适应性；样本级缓存则能保留课程效应，但代价是缓存更大。因此，我们避免使用位置级缓存，因为默认的学生熵选择器会隐式形成课程学习——高熵位置会在训练过程中不断变化，而冻结预先计算的掩码会消除这种适应性，并可能降低蒸馏质量。

相比之下，我们假设样本级选择在学生模型的学习动态下更稳定，因此适合采用单次预筛选，以减少教师查询和缓存大小。这一假设与Xu 等人 (2023年) 的研究一致：他们发现，对学生模型而言不确定的样本对教师来说也很难，这表明样本难度主要由数据本身决定。不过，我们并未声称已确凿地证实这一点。如图 7所示，探索性的重叠分析显示，基于教师的指标（KL 散度和交叉熵比率）之间存在显著一致性（46.7%），但学生熵与任一指标之间的重叠度要低得多（分别为 22.0% 和 12.5%），这凸显了更系统地研究样本选择稳定性的必要性。

另一种方案是在蒸馏过程中在线构建缓存，记录不断演化的学生模型所选择的位置或样本。虽然这保留了课程效应，并且可能在不同学生模型之间迁移，但它牺牲了离线缓存的一项关键优势——即在内存中仅保留一个模型的情况下进行蒸馏的能力——并且不太适合多轮次训练。未来的工作可以将在线课程（例如 GLS）下选择的样本与蒸馏前处理得到的样本进行比较，以更好地刻画选择稳定性。

F. 验证集划分表

本附录报告开发过程中用于模型/指标选择和消融实验的验证集划分结果。论文正文中的所有比较均基于留出测试集划分；验证集划分不用于最终评估。我们使用三个随机种子（1337、1338、1339）评估每项验证实验，并报告平均结果。验证基准套件根据 ArcEasy、GSM8K、HellaSwag 和 PIQA 在各验证集划分上的平均准确率构建。

表*10*1337、1338、133920**80**。 **在验证集上采用 Top- % 选择的位置重要性指标**，在三个固定随机种子（ ）下取平均，使用 FineWeb-Edu 的 M 个词元进行训练。根据这些结果，我们选择学生熵作为位置重要性指标（表 2）。它取得了最高的验证准确率，并且在表现最佳的指标中独有的优势是无需教师端信息，因此能够在教师模型上使用选择性语言模型头，避免对未选定位置进行 logit 计算。

方法	准确率 ↑
Qwen3 1.7B	62.2
Qwen3 8B	***75.1***
AT-KD	65.2
完整知识蒸馏	65.6
随机 20%	65.5
位置选择策略：前*20%*	
学生熵	**66.0**
教师熵	65.4
学生交叉熵	65.8
教师交叉熵	65.3
KL	**66.0**
反向 KL	**66.0**
交叉熵比率	65.9
交叉熵比率 +学生熵	65.8
学生熵 +KL散度	65.7

表*11*。 **验证集上基于学生熵的位置选择策略**，对三个随机种子取平均。我们在主要实验中选择了前20%（表 3）：尽管 GLS 和课程学习的验证准确率略高，但差异很小（0.1–0.2 个百分点），而前20%更简单，无需额外超参数（GLS 的队列大小、课程学习的调度计划）。

方法	准确率 ↑
Qwen3 1.7B	62.2
Qwen3 8B	***75.1***
AT-KD	65.2
完整知识蒸馏	65.6
随机 20%	65.5
位置选择策略：前*20%*	
前20%	66.0
课程学习 20%	66.1
GLS 30K 20%	**66.2**
Pos RS-KD 20%	64.9
Pos RS-KD* 20%	65.5

G. 标准差

我们报告三个固定随机种子下的标准差，以量化其他训练设置完全相同的情况下运行间方差。

表 *12*1337, 1338,13392 20. 表格的标准差 (位置重要性指标（Top- % 选择），在三个固定随机种子（ ）上计算得出。

方法	准确率 ↑	IFEval ↑	困惑度 ↓	期望校准误差 ↓
完整知识蒸馏	0.20	0.56	0.18	0.07
随机位置 20%	0.15	1.25	0.21	0.59
AT-KD	0.04	0.78	0.06	0.04
位置选择策略：前*20%*				
学生熵（SE-KD）	0.14	0.81	0.26	0.12
教师熵	0.68	0.24	1.28	0.52
教师交叉熵	0.30	0.67	0.38	0.79
学生交叉熵	0.16	0.36	0.35	0.20
KL	0.16	1.19	0.09	0.11
反向 KL	0.10	0.45	0.09	0.04
交叉熵比率	0.02	0.12	0.34	0.06
交叉熵比率 + 学生熵	0.06	0.44	0.04	0.07
学生熵 + KL散度	0.13	0.13	0.48	0.49

表 *13*。 表 3 （基于学生熵的位置选择策略），在三个固定随机种子（1337、1338、1339）上计算得出。

方法	准确率 ↑	IFEval ↑	困惑度 ↓	期望校准误差 ↓
完整知识蒸馏	0.20	0.56	0.18	0.07
随机位置 20%	0.15	1.25	0.21	0.59
AT-KD	0.04	0.78	0.06	0.04
位置重要性指标：学生熵，$k = 20\%$				
前20% (SE-KD)	0.14	0.81	0.26	0.12
前20% GLS 30K	0.23	0.33	0.55	0.14
课程学习 20%	0.09	0.41	0.21	0.08
Pos RS-KD* 20%	0.39	0.36	0.44	0.10
Pos RS-KD 20%	0.08	0.99	0.26	0.15

表*14*。 表格的标准差4（通用蒸馏；测试集划分，80M 个词元），根据三个固定随机种子计算。

方法	准确率（%） ↑	IFEval（%） ↑	困惑度 ↓	期望校准误差（%） ↓
完整知识蒸馏	0.20	0.56	0.18	0.07
随机位置 20%	0.15	1.25	0.21	0.59
随机样本 20%	0.27	0.06	0.71	0.03
SE-KD	0.13	0.48	0.26	0.08
RS-KD	0.06	5.33	0.06	0.01
TopSmp 20%	0.58	0.48	0.64	0.05
RS-KD +TopSmp 20%	0.11	0.21	0.00	0.04
SE-KD +TopSmp 20%	0.07	0.77	0.27	0.11
SE-KD_{3X}	0.15	1.61	0.12	0.16

表*15*1340, 1341, 13425 8. 表 （任务特定 GSM K 蒸馏；测试集），基于三个固定随机种子（ ）计算的标准差。

	方法	GSM8K 准确率	Acc.	PPL
Off Policy Distill.	完整知识蒸馏	0.90	0.10	0.06
	随机 20%	0.10	0.12	0.15
	SE-KD	0.12	0.06	0.10
	Pos-RS-KD 20%	0.80	0.17	0.23
	Pos RS-KD* 20%	1.22	0.36	0.31
	TopSmp 20%	0.12	0.00	0.06
	SE-KD +TopSmp 20%	0.06	0.06	0.06
	SE-KD_{3X}	0.31	0.15	0.00
On Policy Distill.	完整知识蒸馏	0.21	0.00	0.06
	随机 20%	0.72	0.06	0.12
	SE-KD	0.15	0.00	0.00
	Pos-RS-KD 20%	2.25	0.00	0.58
	Pos-RS-KD* 20%	0.58	0.06	0.21
	TopSmp 20%	0.49	0.06	0.12
	SE-KD +TopSmp 20%	1.00	0.31	0.66

表16. **选择性语言模型头配置的内存与速度对比**。实验使用 Qwen3-8B（教师）→ Qwen3-1.7B（学生模型），在 NVIDIA GeForce RTX 3090 图形处理器上运行，批次大小为 B=2，序列长度为 T=512，以及 λ=1。默认流程 对应于标准知识蒸馏实现。分块流式流程 重构计算以匹配选择性实现（例如，流式熵计算和位置索引），同时选择所有位置（k=100%），从而隔离代码重组的开销。在 k=20% 时，仅前20%的位置（按学生熵）参与知识蒸馏损失。加速比相对于默认流程基线进行报告。

	配置	k	学生峰值（GB）	教师峰值（GB）	运行时间	加速比
1M tokens	完整知识蒸馏 （默认流程）	100%	15.88	17.30	16.3 分钟	1.00×
	完整知识蒸馏 （分块流程）	100%	14.15	17.58	14.9 分钟	1.09×
	教师选择性语言模型头（分块流程）	100%	14.15	17.29	14.9分钟	1.09×
	无选择性语言模型头（默认流程）	20%	13.59	17.30	13.6分钟	1.20×
	无选择性语言模型头（分块流程）	20%	**11.42**	15.97	12.6 分钟	1.29×
	教师选择性语言模型头（分块流程）	20%	**11.42**	**15.68**	12.6 分钟	1.29×
	学生选择性语言模型头 （分块流程）	20%	**11.42**	15.97	12.2 分钟	1.34×
	教师 +学生选择性语言模型头 （分块流程）	20%	**11.42**	**15.68**	**12.0 分钟**	1.36×
5M tokens	完整知识蒸馏 （默认流程）	100%	15.88	17.30	79.2 分钟	1.00×
	完整知识蒸馏 （分块流程）	100%	14.15	17.58	72.0 分钟	1.10×
	无选择性语言模型头（默认流程）	20%	13.59	17.30	69.7 分钟	1.14×
	无选择性语言模型头（分块流程）	20%	**11.42**	15.97	65.5 分钟	1.21×
	教师选择性语言模型头（分块流程）	20%	**11.42**	**15.68**	62.5 分钟	1.27×
	学生选择性语言模型头 （分块流程）	20%	**11.42**	15.97	61.8 分钟	1.28×
	教师 +学生选择性语言模型头 （分块流程）	20%	**11.42**	**15.68**	**60.1 分钟**	1.32×
80M tokens	完整知识蒸馏 （默认流程）	100%	15.88	17.30	21小时10分钟	1.00×
	完整知识蒸馏 （分块流程）	100%	14.15	17.58	19小时37分钟	1.08×
	教师选择性语言模型头（分块流程）	20%	**11.42**	**15.68**	17小时04分钟	1.24×
	学生选择性语言模型头 *(分块流程)*	20%	**11.42**	15.97	16小时57分钟	1.25×
	教师 +学生选择性语言模型头 *(分块流程)*	20%	**11.42**	**15.68**	**15小时47分钟**	1.34×

H. 内存效率　　选择性语言模型头与分块流式熵计算的效率

表 16 详细展示了不同知识蒸馏实现的内存使用量和训练速度消融结果。具体而言，我们比较了：

1. **默认知识蒸馏实现**：标准知识蒸馏，对教师模型和学生模型均计算完整的 $[B, L, V]$ 逻辑值。
2. **分块流式实现**：采用分块流式熵计算和选择性代码路径，但仍在所有位置计算逻辑值。这样可以独立于位置稀疏化，单独考察分块流式处理的效果。

3. **选择性语言模型头变体**：根据学生熵选定一部分位置，并在这些位置上计算知识蒸馏损失；教师端和/或学生端的选择性语言模型头将 logit 计算和梯度传播限制在选定位置上。

表16呈现了几个观察结果。首先，即使在 k=100%的情况下，分块流式流程也能将学生峰值内存从 15.88 GB 降至 14.15 GB（降低 11%），并通过避免在反向传播期间物化完整的学生端逻辑值张量，实现 9% 的加速比。其次，将 k 从 100% 降至 20% 可带来显著的额外节省：即使不使用选择性语言模型头，学生峰值内存也会降至 13.59 GB（默认流程）或 11.42 GB（分块流式流程），加速比分别为 1。20× 和 1。29×。第三，在 k=20% 时加入选择性语言模型头，可将教师峰值内存从 15.97 GB 进一步降至 15.68 GB，同时保持学生端内存占用不变；教师端和 + 学生端选择性配置相结合，可实现最佳运行时间（12.0 分钟，加速比为 1。36× ）。

图 8 随时间可视化了这些影响。在 k=100%（左图）时，内存峰值源于每个训练步骤中对完整$[B, L, V]$ logit 张量的瞬时分配。降至 k=20% 且不使用选择性语言模型头（中图）时，峰值内存已经降低，因为参与知识蒸馏损失的位置更少，尽管完整逻辑值仍会被物化。在 k=20% 使用选择性语言模型头（右图）时，这些峰值被完全消除，因为逻辑值仅在选定的 ∼20% 的位置上计算。

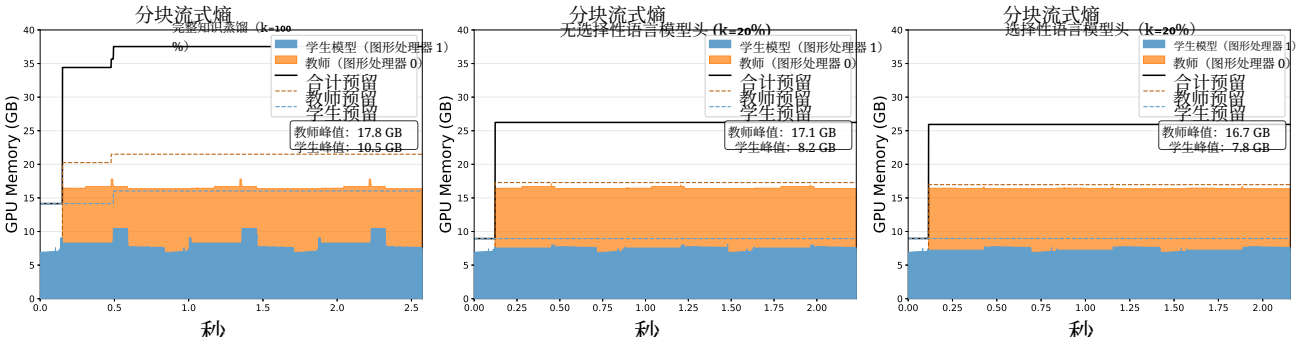

图 *8.* **不同选择性语言模型头配置下的 GPU 内存配置。** 使用分块流式熵计算时，PyTorch 分析器记录的多个训练步骤中的内存轨迹。**左：** k=100% 的完整知识蒸馏，其中分配完整的 $[B, L, V]$ logit 张量会引发周期性的内存峰值。**中：** k=20% 未使用选择性语言模型头，参与知识蒸馏的位置更少，但仍会实例化完整的逻辑值。**右：** k=20% 使用选择性语言模型头，仅在选定位置计算逻辑值，从而消除瞬时峰值并进一步降低峰值内存。教师和学生模型运行在独立 GPU 上；堆叠区域表示已分配的内存，虚线表示预留的内存。