arXiv:2611.32168v1 [q-bio.QM] 24 Nov 2026

将蛋白质结构分词器作为语言模型生成目标进行基准测试

技术报告 II：面向 LLM 的蛋白质结构分词器

刘勇[1,2], 石展鹏[3,4], 林沛聪[2], 钟振毅[1], 党一舟[5], 吴文迪[1], 李浩灵[6], 顾佳宇[2], 安多[2], 郑双佳[1,2]

[1]上海交通大学，上海，中国 [2]临港实验室，上海，中国
[3]复旦大学，上海，中国 [4]上海创新研究院，上海，中国
[5]东北大学，沈阳，中国 [6]南京大学，南京，中国

摘要

离散结构分词器是让语言模型获取蛋白质三维结构的标准方式，但人们通常依据其被读取的准确性，即其重建保真度来选择它们；而决定生成效果的属性，则是模型能够多好地写入这些结构分词器，即其可写性。我们直接测量这一属性：固定基础模型、数据、配方和计算资源，仅替换分词器，并训练相同的 9B 模型逐残基生成结构令牌；共测试八种分词器、四种计算预算和三个评估集（96 个评估单元）。在 20亿个 token 时，在完全相同的条件下，最终可写性的 TM评分 从 0.662 到 0.378 不等；八个分组中有五个达到 TM评分 0.5 这一通常认为折叠结构正确的标准，另有三个未达到。码本大小和重建上限都无法解释某个分组最终所处的位置；表现最好和最差的分组都使用了 8192 个码字。真正能够解释这种差异的是编码空间对局部结构的保持能力。在 DPLM-2 中，一个比特的码字错误仅造成 0.019 的 TM评分 损失，而随机码字会造成 0.347 的损失；训练后模型自身的错误同样具有局部性，其损失仅为随机汉明距离预期的 0.51。FoldToken4 在相同码本大小下未通过这两项测试，其可写性低 0.13 TM评分。三项控制实验限定了应如何解读这一基准测试。采用相同配方重新运行后，噪声下限为 $\sigma = 0.0067$ TM评分，这将表面上的六路排序压缩为四个等级。无参考几何面板与 TM评分 几乎正交（皮尔逊相关系数 -0.58）；按 TM评分 最差的分组反而生成了物理有效性最高的蛋白质骨架。此外，两项预注册干预均呈零效应，原因可由机制测量揭示：它们所针对的量在 15亿个 token 时已经饱和，而 TM评分 仍在上升。全部 131 个评估单元（包括原始生成令牌流）均已发布。

✉ **通信作者** liuyong9975@sjtu.edu.cn

主页 https://calpha-ai.com**GitHub**
https://github.com/yliust/Calpha**HuggingFace**
https://huggingface.co/liuyong9975/Calpha

1 引言

语言模型通过将蛋白质三维结构离散化为词元，并用这些词元扩展模型的词汇表，从而获得对蛋白质三维结构的访问权限。结构标记器会将每个残基映射为 K 个码字中的一个，因此，主链便成为模型可以与文本和序列一同读取的字符串。目前已有多种此类结构标记器可直接使用 [1–6], ，并且越来越多的蛋白质语言模型建立在其基础之上 [7–10]。本文研究的是这样一种设定：模型必须自回归地输出 该字符串，这使结构标记器从感知模块转变为生成目标。在其他模态中，同样的转变已被证明具有决定性作用：在保持语言模型不变的情况下替换分词器，是缩小图像生成中与扩散模型差距的关键 [11], ；而在音频领域，决定质量上限的是编解码器，而非序列模型 [12, 13]。

这一转变改变了分词器所需关注的属性，而这一属性至今尚未得到测量。分词器的评估和宣传依据都是重建保真度：对已知主链进行编码，再将其解码，并报告有多少结构得以保留。这个数值描述的是编解码器。它说明了从编码空间中读取结构的能力，却无法说明以预测下一个词元为目标训练的模型能否生成其中有效的字符串。我们将后一种属性称为可写性。以往的比较并未测量这一属性。结构分词器的系统性基准测试评估了重建和冻结表征探测 [4], ，两者都属于读取侧评估；最接近的机制性研究发现，编码器的局部性与解码器的重建能力之间存在权衡 [14], ，但该研究只评估了分词器本身，并未让模型尝试在编码空间中生成内容。已发表的、能够生成结构令牌的系统，在分词器、基础模型、数据和计算资源等方面同时存在差异，因此分词器本身的贡献从未被单独分离出来。

我们将其单独分离出来。我们在同一个90亿参数基础模型中训练八种分词器，统一使用同一混合数据集、配方、样本顺序和计算预算；在四种预算下导出模型，并在三个留出集上进行评分。随后通过三项控制来界定这些数值的含义。每项评分都以一个下限和一个上限为界：下限是将均匀随机令牌输入该分词器自身的解码器和评分器，上限则是对真实值编码并解码一次。设定这一范围很有必要，因为各解码器的结构先验差异足以使随机令牌的TM评分落在0.10至0.19之间。我们通过端到端重复一个组别来测量运行间的噪声下限，而非直接假定它。指标面板中有一半完全不参照参考结构，因为TM-score [15], 、lDDT [16], 、GDT [17] 和RMSD同样无法识别不符合合理蛋白质结构的输出；正是这种盲点促使分子对接领域引入物理有效性检查 [18]。

由此得出三项发现。第一，在相同条件下，训练至20亿个 token 时，最终可写性从 TM评分 0.662（DPLM-2）到 0.378（VanillaVQ）不等，跨越了通常判定折叠结构正确的 TM评分 0.5 门槛：五个组别达到该门槛，三个未达到。分词器的选择依据都无法预测某个组别最终处于什么水平：表现最好和最差的分词器都使用了8192个码字，而且八种重建结果中有六种的 TM评分 高于0.95。第二，真正能够预测结果的是相近的码字是否表示相近的结构。DPLM-2 的码字是13比特字符串，将30%的位置各翻转一比特会使 TM评分 降低0.019，而将这些位置替换为随机码字则会使其降低0.347；训练后的模型自身所犯的错误与正确码字相差3.29比特，约为随机情况下预期距离的一半。FoldToken4 使用二进制码，码本大小相同，却两项测试均未通过，且可写性低0.13 TM评分。若在编码空间中稍有偏差就只会造成轻微错误，那么这个编码空间便具有可写性。第三，三项对照并非只是形式上的记录，而是会改变结论：实测噪声下限将六个表面上的排名压缩成四个等级；按每种分词器的噪声下限进行扣除后，几何排名发生逆转，因为一种解码器面对随机输入时能输出81%的有效键长，而另一种仅能输出3%。

2 相关工作

结构分词器是作为编解码器设计和评估的。离散蛋白质结构标记化已趋于采用对学习得到的几何编码器进行逐残基量化，差异主要在于量化器。ESM3 [1]和 AminoAseed [4] 使用学习得到的向量量化码本 [19]；DPLM-2 [2] 采用从图像标记化 [11]引入的无查找二进制量化；Gaujac 等人 [5] 使用有限标量量化 [20]；ProSST [9] 对冻结的几何特征运行{k}-means；FoldToken4 [3] 和 ProTokens [6] 旨在构建紧凑、类似氨基酸的词汇表；ProTok [21] 则让序列、结构和功能共用一个词汇表。Foldseek 的 3Di 字母表 [22] 早于这条研究路线，其设计目标是搜索而非解码。所有这些方法的核心指标都是重建效果，即往返转换的 TM评分、lDDT 或 RMSD；它衡量的是编解码器的性能，而非语言模型能用它做到什么。

唯一系统性比较针对读取侧，而最接近的机制性结果仅涉及分词器本身。 StructTo- kenBench [4] 从四个维度评估结构分词器：重建、码本利用率，以及对冻结表示在监督任务上的探测。报告指出，没有任何一种分词器在所有维度上都占优。每个维度都在探究能够从编码中

解读出什么。AIDO.St [14] 最接近我们的问题：它揭示了编码器的局部性与解码器的重建能力之间的权衡，但测量对象仅为分词器本身；没有语言模型被要求在编码空间中进行写作，下游证据关注的也是表示质量，而非生成。我们的贡献是测量此前未被考察的写入侧，并采用一种将基础模型、数据、配方和计算资源固定不变的设计。

生成结构令牌的系统在所有方面都各不相同。 许多模型会将结构令牌扩展到语言模型的词汇表中，并生成这些令牌：采用 MaskGIT 风格的迭代掩码解码的 ESM3 [1] 和 DPLM-2 [2]；通过扩展蛋白质或通用语言模型的词汇表实现这一点的 ProtTeX [8] 和 SaProt [7]；以及使用直接在其编码上训练的 GPT 的 Gaujac 等人 [5]。每种方法都将一种分词器与一个基础模型、一个混合数据集和一份计算预算搭配，因此分词器的贡献与其他所有设计选择相互混杂。我们的实验方案正是为了解耦这些因素而设计的，这也正是我们愿意承担训练八个其他方面完全相同的组别所需成本的原因。

专业折叠模型提供了绝对参考。 AlphaFold2 [24], RoseTTAFold [25] 以及 Al- phaFold3 [26] 无需离散化步骤，便可在连续空间中预测结构；而 ESMFold [27] 仅凭蛋白质语言模型的特征就能做到这一点。ESMFold 是第 7节的恰当参考：它使用与我们的组别作为条件输入相同的 ESM2 特征，并直接生成坐标，因此两者的差异揭示的是通过离散编码和自回归解码器处理结构所带来的成本，而非输入较弱所造成的差异。CASP [28, 29] 和 FoldBench [30] 提供了留出的目标。

自回归退化和无参考有效性在其他领域已是众所周知的问题。在自然语言中，贪心解码和低温度解码会陷入重复； [31, 32], 而教师强制训练从自身指标中隐藏的误差累积，自计划采样 [33, 34]以来就一直受到研究。第二个经验在结构生物学中的对应观点是：针对参考结构的评分无法识别一个在物理上并非有效分子的输出，这正是 PoseBusters [18]所依据的论点。我们的无参考面板将这两点同时应用于结构令牌：统计生成字符串的标记级退化，以及对解码后的主链进行几何检查（键长、碰撞、回转半径、DSSP组成 [35]）。

3 衡量可写性

定义。当模型必须自行生成结构令牌字符串时所达到的结构质量，称为可写性：输入氨基酸序列和任务指令后，模型通过贪心解码为每个残基生成一个词元；随后，该字符串经过相应分词器自身的解码器处理，并使用 TM-score [15, 36]将坐标与实验结构进行比较评分。这一测量必然包含解码器，因为解码器是实践者所采用接口的一部分。正因如此，接下来的两项控制措施是强制性的。

下限和上限使评分具有可比性。 由于每个解码器本身贡献的结构量不同，原始TM评分无法跨分词器进行比较。我们通过在相同目标上使用完全相同的处理流程进行两次测量，为每个模型评分划定区间：来自均匀随机令牌的下限 S_4表示仅凭解码器的先验所能达到的水平，而将真实值编码并解码一次所得的上限 S_1，则是在该编码空间中任何模型所能达到的最佳水平。我们报告兑现率

$$\mathrm{RR} \;=\; \frac{\mathrm{TM}_{\mathrm{model}} - S_4}{S_1 - S_4}, \tag{1}$$

，即模型实际捕获的可达到区间的比例。图 2 将这三个量放在一起展示，这也是我们贯穿全文的报告方式：单独的模型评分没有可解释性。

兑现率存在一种失效模式，报告兑现率时必须同时报告这一点。 当区间较窄时，归一化值会被夸大。我们集合中的两个分词器重构效果不佳（$S_1 = 0.48$ 和 0.42），但它们的兑现率却位居表格前列（70.9% 和 81.0%），而按原始TM排序则垫底（0.395 和 0.378）。在较低上限上的高兑现率，意味着模型几乎耗尽了一个本来就无法表达蛋白质结构的编码空间。因此，我们从不报告不带有 S_1 的数据的兑现率，并将这两个分词器视为反例，而非成功案例。

实测噪声下限决定了排序允许达到多细的粒度。 此类基准测试在不了解测量本身方差的情况下报告严格排序，这种报告做法在机器学习领域一直受到批评 [37–39]。我们直接对其进行估计：一个组别（ESM3）在配方、数据和样本顺序完全相同的条件下，从头到尾训练了两次。九组配对差异的均值为 +0.0028，标准差为 0.0067，最大绝对值为 0.0123 TM评分，其中四个为正、五个为负。我们采用

$$|\Delta\mathrm{TM}| < 0.013 \;\Rightarrow\; \text{indistinguishable}, \tag{2}$$

并将其应用于本文中的每次跨分词器比较。这是真实噪声的下界，因为它来自单个组别的一次重复，因此该规则在关键方向上是保守的。

指标面板中有一半不得查看参考。 TM、lDDT [16], GDT [17] 和 RMSD 都通过解码器将输出与目标进行比较，并且都无法识别一类失败情况：输出根本不是合理的蛋白质——这正是引入物理有效性检查以弥补的缺口，类似于分子对接中的情况 [18]。因此，我们仅根据生成结构计算第二个面板：在词元层面，统计不同词元的比例、相邻重复率、最长连续重复和熵，这些指标能够检测神经文本生成中常见的贪心退化 [31, 32]；在几何层面，统计连续 CA–CA距离处于 $[3.6, 4.0]$ Å 以内的比例、超过 $|i-j|>2$ 的冲突率，以及以 $L^{0.34}$ 归一化的回转半径，这些指标的取值范围均与 TM评分使用的下限和上限相同。由于原始词元序列已存档，该面板得以对数周前训练的检查点进行追溯计算，无需重新运行模型。

4 实验设置

受控变量只有分词器，其他一概不变。 每个组别都从同一个90亿参数基础模型开始，使用该分词器的 K 码字扩展词汇表，并在同一混合数据集上训练（结构预测、逆折叠、填充、二级结构、接触和通用文本）。该数据集由同一脚本构建，任务比例、数据泄漏过滤和样本顺序也完全相同。各组别使用完全相同的训练配方：在20亿个 token上进行一次余弦调度，对主体、嵌入和适配器使用相同的学习率，采用相同的噪声前缀增强，并在0.5、1、1.5和2B个 token处导出模型。训练在 8×H200上进行，每个组别约需10小时。唯一有意设置的差异是：结构令牌由哪种分词器生成，以及由哪个解码器为输出评分。

分词器。我们评估了八种分词器，其码本大小从512到8192不等，涵盖三类量化方法（表5）：ESM3 [1, 40]（学习型向量量化码本 [19], K=4096）、DPLM-2 [2, 41]（无查找二进制量化 [11], K=8192）、ProFiT [42]（K=8192）、FoldToken4 [3, 43] 的10级和12级（K=1024 和4096，二进制码）、AminoAseed [44]（K=512）、GCP-VQVAE [45]（K=4096）以及VanillaVQ [44]（K=512）。AminoAseed和VanillaVQ使用相同的网络，仅码本训练方式不同，因此可以单独考察这一因素：VanillaVQ存在码本利用不足的问题，而AminoAseed正是为解决这一问题而设计的 [4, 46]。这八种分词器的代码库提交记录和权重校验和见附录B。

评估集与泄漏控制。我们报告三个留出集上的结果：FoldBench334 [30]（333 条链，2023 年 1 月之后发布）、CASP16 [28, 29]（155 个结构域）和 CAMEO2025 [47]（200 个于 2025年 发布的单链 X射线结构）。泄漏控制在构建训练集时完成，而不是在评估时进行：通过时间切分，并针对评估目标进行序列一致性过滤 [48]，同时记录每个目标最接近的剩余同源序列，以便将结果划分为同源和非同源两组。所有单元格均报告贪心解码结果；本轮不涉及采样和 best-of-n，因此我们不对它们得出结论。

5 结果

5.1 可写性的 TM评分 范围为 0.38 至 0.66，码本大小和保真度均无法解释这一点

网格。图1 展示了全部 96 个单元格。在 20亿个 token 的计算预算下，FoldBench334 上的最终 TM评分介于 0.3781（VanillaVQ）和 0.6623（DPLM-2）之间；这些组别之间唯一的差异是分词器，而且三个评估集给出的排序一致。我们将这些数值作为绝对分数给出，刻意不将其表示为二者之间的比率。TM评分没有真正的零点：随机令牌经过同一批解码器后，得分已经达到 0.103 至 0.185（表1），因此这个范围的两端都建立在约 0.17 的基准之上，而两个原始 TM评分值的商会随假定的原点而变化。若改为对扣除基线后的分数求比，结果又会不同。第3节定义的兑现率才具有明确的端点，我们报告的归一化数值正是它；本文中出现的比率

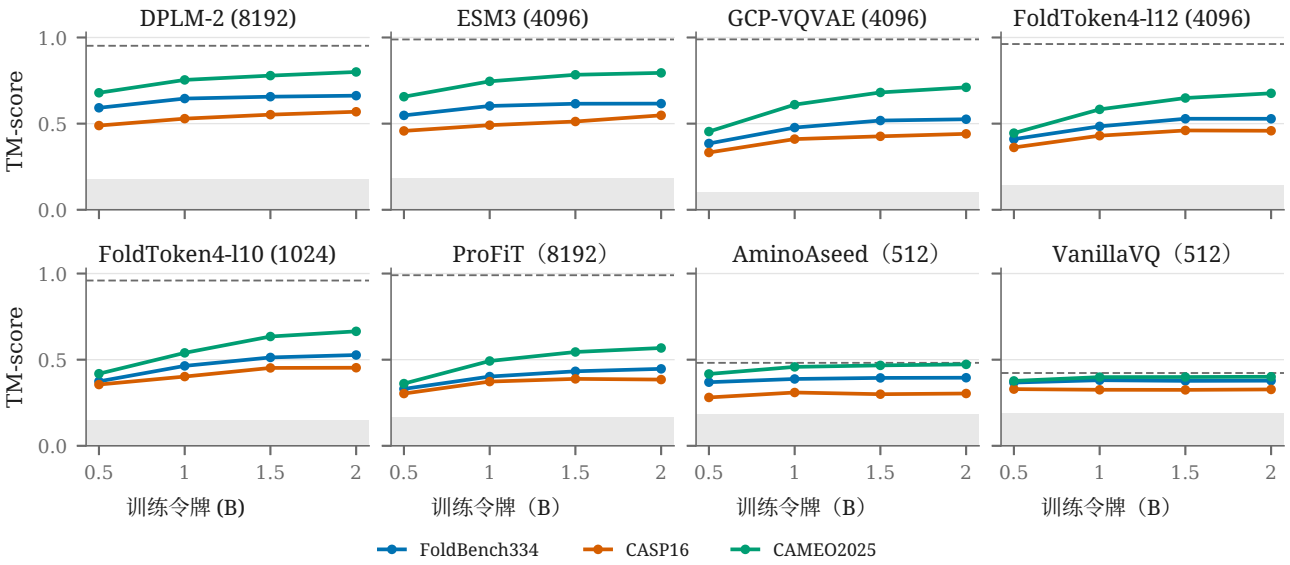


图1。 相同条件下的可写性。每个面板对应一个分词器（括号内为码本大小）；三条曲线分别对应三个评估集。阴影带表示该分词器的随机令牌下限 S_4，虚线表示其重建上限 S_1，两者均通过相同的解码器和评分器测得；各面板共用坐标轴。由于在色觉缺陷情况下无法可靠地区分八种分类色调，因此通过面板标题而非颜色来标识分词器。

表1。 FoldBench334 上 20亿个 token 时的最终可写性，使用每个分词器自身的下限和上限。RR 是兑现率。TM评分差异小于 0.013 实测噪声下限的分组归入同一等级，且不在组内排序。标记为 † 的两个分组的重建上限低于 0.5：它们较高的兑现率是狭窄区间造成的伪影，而非成功。

Tier	分词器	K	S_4 (下限) ↓	S_1 (上限) ↑	TM ↑	RR ↑
1	DPLM-2	8192	0.177	0.952	**0.6623**	62.6%
2	ESM3	4096	0.185	0.989	0.6162	53.7%
3	FoldToken4-l12	4096	0.144	0.962	0.5281	47.0%
	FoldToken4-l10	1024	0.151	0.959	0.5270	46.5%
	GCP-VQVAE	4096	0.103	0.989	0.5254	47.7%
4	ProFiT	8192	0.168	0.990	0.4467	33.9%
退化代码空间（作为反例报告，不参与排名）						
–	AminoAseed†	512	0.185	0.482	0.3952	70.9%
–	VanillaVQ†	512	0.189	0.423	0.3781	81.0%

都是 TM评分差值之间的比率，而差值确实有零点。除两个上限较低的分词器外，没有任何组别达到饱和；下文的每项结论都仅适用于这一计算预算。

码本大小并不能决定各组的排序。表中表现最好和最差的分词器使用的码字数都是 8192（DPLM-2 为 0.6623，ProFiT 为 0.4467）。FoldToken4 在 K=1024 和 K=4096 下的 TM评分相差不到 0.001。大小并非决定因素。

重建保真度也不能决定它们的排序。八种分词器中有六种的重建 TM评分高于 0.95，而这六种的可写性几乎覆盖了整个范围（0.4467 至 0.6623）。ProFiT 在这组中的上限最高（S_1 = 0.9902），兑现率却最低（33.9%）。一种分词器可以是出色的编解码器，却是不理想的载体；读取和写入是不同的属性。

5.2 测得的噪声下限将六个排名压缩为四个等级

三个分组之间无法区分。 表1 采用了第 3节中的噪声规则。FoldToken4-l12、FoldToken4-l10 和 GCP-VQVAE 之间的 TM评分差距为 0.001–0.003，比 0.013 的噪声下限低一个数量级；按排名顺序报告它们，无异于报告噪声。仅有的两个显著差距都很大：DPLM-2 优于 ESM3 的幅度为 0.046（3。5× 倍于噪声下限），ESM3 优于第 3 级的幅度为 0.088。

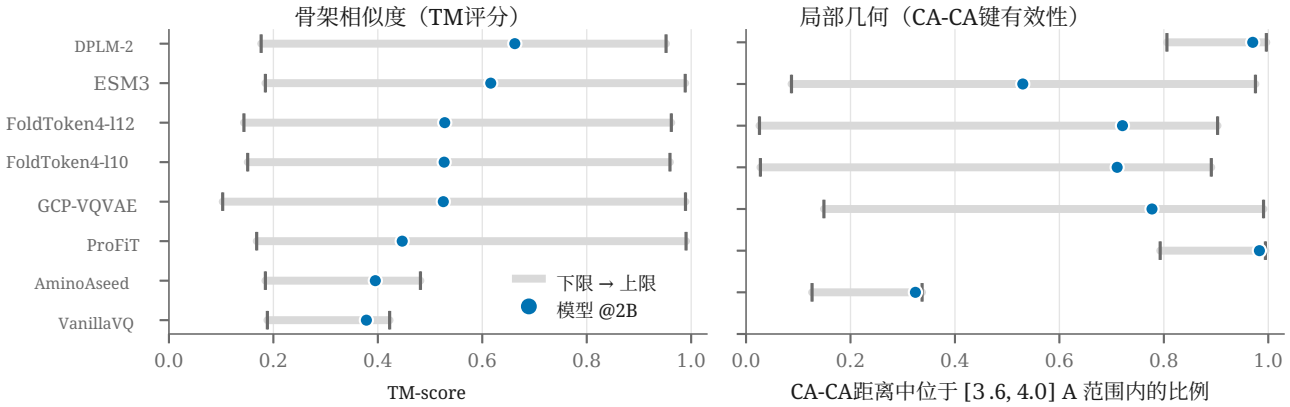


图2. 在 FoldBench334 上使用 20亿个 token 时，每项分数都应结合其各自的下限和上限来解读。左：骨架相似度。右：局部几何，即连续 CA–CA距离中处于 [3.6, 4.0] Å 以内的比例。条形两端分别表示该分词器的随机令牌下限和重建上限；圆点表示训练后的模型。两个面板中的分词器排序几乎完全相反。

一个差距在计算资源增加时保持稳定，另一个则正在缩小。 每个分组有四个计算点，因此我们可以判断差距究竟是分词器本身的特性，还是计算预算的结果。DPLM-2 相较于 ESM3 的优势基本不变（0.044、0.043、0.041、0.046）；而 ESM3 相较于 GCP-VQVAE 的优势则并非如此，从 0.163 降至 0.091。因此，我们对第一个差距不加限定地陈述，而对第二个差距仅在 ≤2B 词元时陈述。各分组在各自曲线上的位置也不尽相同（从 1.5B 到 2B，ESM3 提升 0.0007，FoldToken4-l12 −0.0002，而 ProFiT 和 FoldToken4-l10 提升 0.014），因此，单一计算预算下的快照在一定程度上反映了收敛速度。

5.3 解码器先验相差一个数量级，必须予以扣除 ed

在某些代码空间中，随机词元已经能生成有效的局部几何。 将均匀随机词元输入各个解码器后，DPLM-2 和 ProFiT 的 CA–CA键长有效性分别为 0.806 和 0.793，而 GCP-VQVAE 为 0.149，ESM3 为 0.087，FoldToken4 为 0.026–0.028。仅凭解码器本身、完全没有任何信息时，它们生成物理上合理骨架的能力就相差一个数量级。若不报告这一随机令牌下限，任何被引用的绝对几何指标都会把解码器的先验归功于模型。

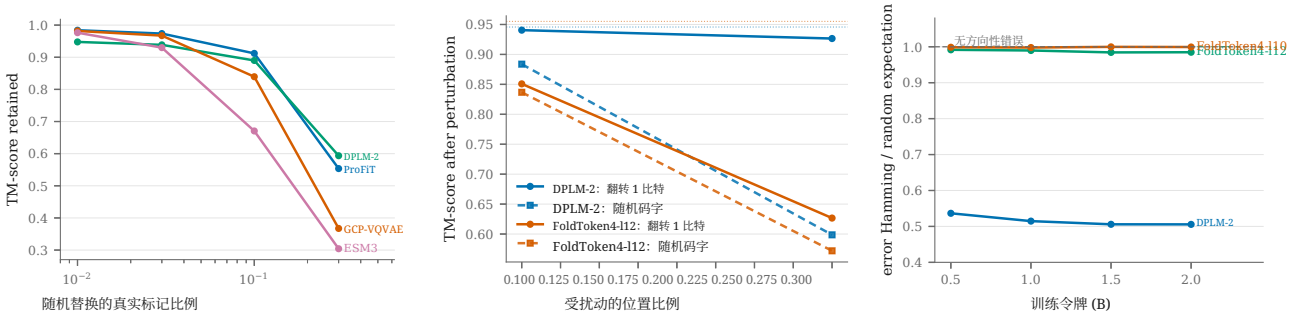
减去下限后，几何排名并不是TM排名。DPLM-2 的模型输出按原始键长计为 0.970，有效值接近其 0.996 的上限，但其下限为 0.806，因此捕获了其区间的 86.3%。ProFiT 按 TM评分计为最弱的组别，却达到 0.983 的原始值和其区间的 94.3%，为该组中最高。ESM3 按 TM评分排名第二，原始值为 0.530，占其区间的 49.9%；其归一化回转半径为 3.33，而实验结构为 2.79。图 2 展示了并排排列的两个面板。

两个面板几乎相互正交。在七个分组中，原始TM与经下限校正的键有效性之间的皮尔逊相关系数为 −0.58（斯皮尔曼相关系数为 −0.46）；与兑现率相比，+0.05。无参考面板代表第二个维度，而非冗余的检验。极端情况具体说明了这一点：ProFiT生成的东西看起来像蛋白质，却不是正确的蛋白质；而ESM3生成的东西虽然轮廓吻合，但其局部几何有一半无效，且链过度延伸。

退化分词器在这里也会失败，方式也完全相同。AminoAseed 达到了其键长有效区间的 93.9%，但其上限为 0.337，这意味着真实坐标经过自身编解码器处理一次后，并不能得到有效键长。这是第 3 节所述归一化陷阱在第二项指标上的重现，也是我们在所有地方都以绝对值报告上限的原因。

5.4 对损坏的容忍度 是表示的一种属性，而不是可写性的预测指标 ty

对随机词元损坏的鲁棒性无法预测可写性。随机替换 10% 的真实词元后，ProFiT 的 TM评分仍为 0.912，DPLM-2 为 0.890，GCP-VQVAE 为 0.839，ESM3 为 0.671（图3a）：对损坏最耐受的分词器可写性最低，而耐受性最低的分词器可写性排名第二。能够容忍任意错误，并不等于易于写入：关键在于错误是否落在结构上相近的码字上（第6节），以及远程接触是否能够保留（第7节）。我们报告这种不一致性，而不是选择与结论一致的指标。



(a) 抗损坏能力。 **(b)** 比特局部性。 **(c)** 训练过程中的误差几何。

图 3. (a) 将一部分真实标记替换为均匀随机标记；这种排序并非可写性排序。(b) 将一部分位置扰动一比特（实线）与扰动为随机码字（虚线）；点线表示各分词器的重建上限。(c) 模型自身词元错误的平均汉明距离，除以均匀随机错误码字的期望值；1.0 表示没有方向性。

6 什么使编码空间具备可写入性

假设。上一节排除了码本大小、重建保真度和损坏耐受性。剩下的候选因素是编码空间局部结构的一种性质：如果在模型输出空间中彼此接近的码字表示的是在笛卡尔空间中彼此接近的结构，那么一次近失误就会得到一个近似正确的结果；具备这一性质的分词器就是可写的。对于码字为二进制码的分词器而言，这一点可以直接检验，因为“在输出空间中接近”具有明确含义：汉明距离较小 [49]。这正是纠错输出码有意实现的性质 [50]；而在这里，它是量化器的副作用，因此在不同分词器之间差异很大。AIDO.St 以重建 [14] 为代价换取的编码器侧局部性，在读取端观察到的正是同一个量。我们的分词器中有两个符合条件：DPLM-2，其无查找量化器 [11] 使每个码字成为长度为 13 比特的字符串；以及 FoldToken4，其第 ℓ 级码为长度为 ℓ比特的字符串。

测试 1：相邻码字是否表示相邻结构？我们对真实主链进行编码，扰动部分位置，然后进行解码（图3b）。如果编码在局部保持结构，扰动一个比特的代价就应该很低；而扰动为随机码字则作为对照。在 30% 的位置上，对于 DPLM-2，一个比特的扰动代价为 0.019 TM（从 0.946 降至 0.927），而随机码字的代价为 0.347（降至 0.599），相差 18 倍。对于 FoldToken4-l12，两种扰动的代价几乎相同（10% 时分别为 0.851 和 0.837，30 % 时分别为 0.627 和 0.572）：尽管表现为比特，其比特并不编码结构相似性。

测试 2：模型自身的错误是否具有局部性？测试 1 衡量的是分词器的属性。真正重要的问题是，训练后的模型是否会利用这一属性。我们取出存档中的贪心生成结果，针对模型写错的每个词元，测量其写出的内容与正确内容之间的汉明距离，并与均匀随机错误码字的预期进行比较（图3c）。在 20亿个 token 时，DPLM-2 的模型平均在 13 个比特中错 3.29 个，比例为 0.506；FoldToken4-l12 为 0.985，FoldToken4-l10 为 0.999，与写入随机错误码字几乎没有区别。比特承载结构信息的分词器，其模型所犯的错误在结构上代价更低；同时，它也是可写入性高出 0.13 TM 的分词器。

这种数量在训练结束前很久就已饱和。由于四种计算预算下生成的结果都已归档，因此这呈现的是一段变化轨迹，而非单个数据点。DPLM-2 的比率在 0.5–2B 个词元期间依次降至 0.537、0.515、0.506、0.506，并在 1.5B 个词元时停止下降；与此同时，TM评分仍在上升（0.5919、0.6455、0.6564、0.6623）：无论 TM评分最后四分之一的提升由什么因素驱动，都不是模型错误局部性进一步增强所致。FoldToken4 的两个层级始终处于 0.98–1.00，经过 20亿个 token 的训练也毫无变化。

这些结果能说明什么，不能说明什么。局部性解释是基于八种分词器之间的相关性以及对机制本身的两项直接测量得出的，也是我们测试过的唯一经得起检验的候选解释。但它并未得到因果层面的确证：为检验这一解释而设计的干预（第8节）结果为零效应，正如上述饱和结果所解释的那样。我们将其表述为目前对排序最有力的解释，而非已证实的成因。

7 与专业折叠模型之间的差距

绝对数值，而非相对数值。前两节将各分词器相互比较，并与各自的上限比较。但都没有探究为何表现最好的组别仅达到 TM 0.662，而 ESMFold 在相同的

表2。 FoldBench334 上 20亿个 token 时，按序列间隔划分的接触捕获，以模型实现的下限至上限区间比例表示。各分组的下限（序列中最近邻）为 0.071 / 0.047 / 0.033，上限（解码真实标记）为 0.91–0.94 / 0.90–0.93 / 0.98–0.99。长距离区间的分组跨度几乎是短距离区间的三倍。

分词器	TM评分 @2B ↑	短 ↑	中 ↑	长 ↑
DPLM-2	0.6623	72.3%	71.3%	**63.2%**
ESM3	0.6162	67.6%	67.0%	58.7%
FoldToken4-l12	0.5281	67.9%	63.8%	44.6%
GCP-VQVAE	0.5254	56.9%	52.1%	35.4%
ProFiT	0.4467	59.5%	49.2%	**20.3%**
分布于各组		15.4 分	22.1 分	**42.8 分**
与 TM评分的相关性		+0.786	+0.906	**+0.976**

333 个目标上，使用相同的 ESM2-3B 特征却达到了 0.870。ESMFold 是合适的参考，因为它使用相同的特征并直接输出坐标，中间无需经过离散编码，因此这一差异反映的是令牌接口的代价，而非输入较弱造成的影响。配对差值为 +0.209（自助法95%置信区间 [51] [+0.194, +0.227]），且 ESMFold 在 96.3% 的目标上表现更好，因此这种差距是系统性的。本节将对此进行分解。表示并非限制因素：DPLM-2 使用真实标记重建时，TM评分达到 0.952，并恢复了 0.983 的长程接触，两者均高于 0.870。因此，损失发生在输入与输出字符串之间。

长程接触是分词器之间差异最大的地方，也是这个分词器的短板所在。 我们按序列间隔将接触分组，并通过相同的处理流程对每个区间进行评分，范围介于下限（序列中最近邻预测）和上限（解码真实标记）之间。DPLM-2 可捕获 72.3% 的可实现短程接触（$6 \le |i-j| < 12$）、 71.3% 的中程接触和 63.2% 的长程接触（$|i-j| \ge 24$）。单个组别内 9 个百分点的梯度看似微不足道；但跨组别来看，它却是主导因素（表 2）。长程兑现率介于 63.2% 至 20.3% 之间，跨度为 42.8 个百分点，而短程接触的跨度仅为 15.4 个百分点；长程兑现率与最终 TM评分的相关性为 +0.976，相比之下，短程接触的相关性为 +0.786。 **可写入的分词器与不可写入的分词器之间的区别，在于长程接触能否保留在生成字符串中**。 这也解释了第 5 节中出现的反转：ProFiT 生成的蛋白质骨架有效性最高，TM评分却最差，因为它写入的是局部看似合理、但全局拓扑结构错误的片段。

差距的一半来自模型未能迁移的远程接触。 输入中包含的远程信息已经多于模型所生成的信息：ESM2-3B 自身的无监督接触预测头 [52] 在这些目标上的 P@L/5 = 0.813 达到，而模型仅为 0.633。以每个目标的长程精度回归每个目标的 TM评分，DPLM-2 的斜率为 0.505，其他每个组别的斜率也基本相同（0.526、0.530、0.529； $r = 0.83$–0.87），因此接触精度与骨架相似度之间的兑换率是处理流程的属性，而不是某一个分词器的属性。将模型从 0.633 提升到输入的 0.813，相当于+0.091 TM，这相当于**0.209 差距的 44%**。按目标划分的结果也与此一致：在模型远程接触精度最高的三分之一目标上，它落后于 ESMFold 0.119；而在最差的三分之一目标上，则落后 0.325。

另一半则来自离散化本身的成本。将同一回归外推至该表示自身的上限（长程得分为 0.983），预测 TM 评分为 0.840，仍低于 ESMFold 的 0.870。即使模型能够迁移其编码空间所能表示的每一个长程接触，也无法弥合这一差距。同样，在模型长程接触表现最佳的三分之一目标上，它仍落后 0.119。因此，我们将这一差距大致分解为一半来自长程迁移，一半来自不可约离散化成本；但需注意，第二项是通过外推至观测范围之外得出的。

不存在一组稀疏的关键错误。如果少数灾难性词元起了主导作用，那么基于置信度的修复或局部重采样就能弥补大部分损失。我们从模型自身的贪心词元出发，将随机比例 x 替换为真实值后重新解码；恢复效果呈次线性。对于 DPLM-2， $x = 10\%$ 弥补了到上限距离的 5%， $x = 20\%$ 弥补了 12%， $x = 40\%$ 弥补了 28%；ESM3 的情况更为极端，分别为 3%、7% 和 18%。两条曲线都是凸的：修正词元的边际价值会随着已修正词元数量的增加而上升，因此必须大批量纠正错误，这与拓扑结构由协同作用决定的预期一致。我们的预注册阈值规定，若 $x = 10\%$ 弥补了一半差距，就会将这些错误判定为稀疏；但它实际只弥补了二十分之一。

表 3. 在相同的 333 个目标和相同的 ESM2-3B 特征上，最具可写入性的组别（DPLM-2，20亿个 token）与 ESMFold 之间 0.209 TM评分差距的分解。各项相互重叠，并非严格可加。

Term	幅度	占比	证据
未迁移的长程接触	≈ 0.091	44%	输入 0.813，对比模型 0.633；斜率 0.505–0.530
离散化成本	≈ 0.07	33%	上限 0.983 预测 TM评分 0.840 < 0.870
解码／选择（可实现）	0.0095	5%	中心样本与贪心解码，配对比较
（预言机上界）	(0.0515)	（25%）	与参考进行 32 次生成取最优

改进解码的作用有限，我们测量了其实际幅度。如果分布中存在远优于贪心解码所选结果的结构，那么改进解码将是成本最低的补救办法。我们对每个目标在 T=1.0下采样了 32 条链，并对它们全部评分，报告预言机最大值（按 TM评分选出的最佳结果，是利用答案得出的上界）、中心样本（与其他样本最相似的样本，不使用参考，因此实际可用，也是自洽性选择 [53]在结构上的对应方法），以及 32 个随机词元字符串中的最大值；按照极值统计，这样才能得到匹配的抽样结果。预言机最大值比贪心解码高出 +0.0515 TM评分（配对 95% 置信区间为 $[+0.0448, +0.0589]$，在 84.4% 的目标上更优），证实分布中确实存在更好的结构。中心样本仅弥补了 +0.0095 （$[+0.0036, +0.0157]$，占 53.2% 的目标）：**无需参考即可获得预言机收益的 18%**。随机词元对照组的得分为 0.2208，远低于前两者，因此这种效果并非多次采样造成的伪影。因此，实际中解码和选择约能带来 0.01 TM评分 的提升，约占差距的 5%；代价是推理成本增加至 32 倍，而理论上限虽为 25%，却无法达到。

分解结果总结。 表3 汇总了各项。长程迁移占主导，其次是不可约离散化成本，而解码所占比重较小。这三项并非严格可加，因为改善长程迁移会改变另外两项的剩余空间，因此这些百分比表示的是量级，而非精确划分。实际而言，仅靠更好的解码器或更好的训练目标都无法单独弥合这一差距：前者只能带来百分之一个 TM 评分点的提升，而后者会遇到约为 TM 0.84 的离散化下限。

7.1 处理流程彻底失败之处：超过 500 个残基的链

长度效应并不像通常假设的那样。 人们通常认为，长链的表现会单调下降，但对于可写入性最高的组别而言，事实并非如此。DPLM-2 的每个目标的 TM 与链长的相关性仅为 -0.054，其表现最佳的区间是 300–500 个残基（TM 0.726），高于 100–200 个残基（0.638）和 200–300 个残基（0.674）。对于少于 500 个残基的链，不存在值得报告的长度惩罚。

在 500 个残基处存在一道断崖，而且这是一次远程失效。 超过 500 个残基后，TM评分降至 0.549，而无需参考的分解说明了原因：远程接触捕获率从 300–500 区间的 66.0% 暴跌至 40.1%，而短程捕获率下降幅度要小得多，从 70.7% 降至 59.0%（图 4，右）。该区间包含 26 个目标（其中 25 个也有接触面板，捕获率数据即据此计算），因此点估计较为宽松；但该区间与 ESMFold 之间的差距（0.327，配对 95% 区间 $[+0.254, +0.390]$）与其他任何区间都没有重叠，而其他区间中最宽的也只达到 +0.248。这与第 7 节中作为运行区间出现的机制相同：长链包含比例更高的长程接触，而长程接触正是该处理流程无法迁移的部分。

随着链长增加，与专用模型之间的差距会扩大，因为专用模型的表现有所提升。 模型自身的表现随链长增加而下降得并不明显，但它与 ESMFold 的差距却持续扩大：从残基数少于 100 时的 0.151，增至残基数达到或超过 500 时的 0.327。这是因为 ESMFold的表现会 随着链长增加而提升（TM评分与 +0.172 链长正相关，而 DPLM-2 则为–0.054 ）。较长的链往往携带更多进化信号，结构域结构也更规整；专业折叠模型能够将这些特点转化为准确率，而词元流水线却不能。只报告链长混合基准测试上的平均值会掩盖这一点：两个系统在不同情况下落后的程度并不相同。

除了远程接触之外，长度本身也会产生影响。 链长与远程捕获呈负相关（–0.183），因此长度效应有一部分是接触效应。但并非全部如此：将每个目标的差距对远程捕获进行回归后，残差仍与链长相关，相关值为 +0.195。某种随链长增长的因素——无论是自回归步骤的数量、累积的承诺，还是注意力跨度——都会独立于捕获到哪些接触而降低准确率。较长生成过程中的误差累积

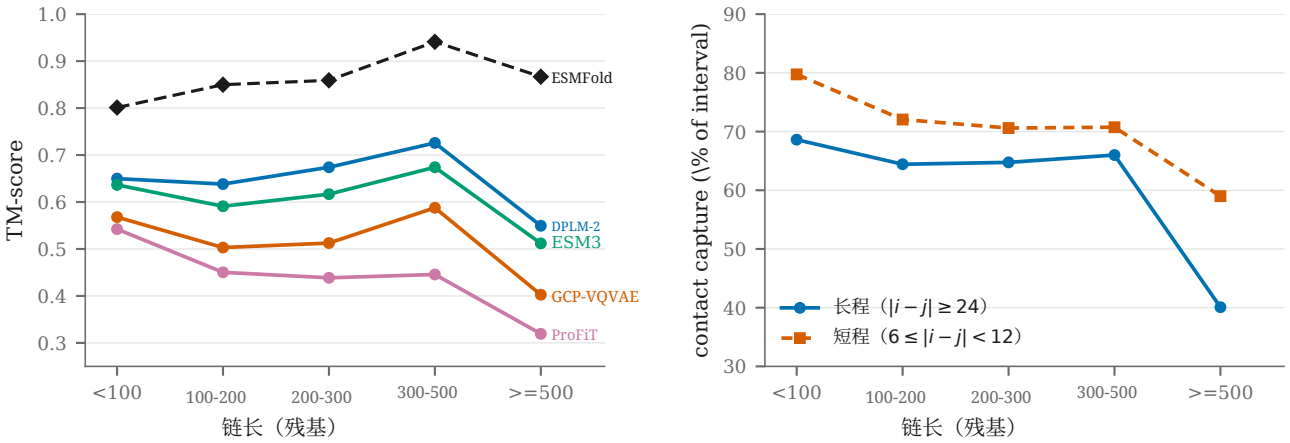


图4。 链长。左：四个分组及在相同目标上的 ESMFold 按长度区间统计的 TM评分。模型分组在 300–500 个残基处达到峰值，超过 500 个残基后下降；ESMFold 则随长度增加而上升，因此差距单调扩大。右：DPLM-2 按区间统计的接触捕获，以兑现下限至上限区间的比例表示。超过 500 个残基后，长距离区间崩溃，而短距离区间则缓慢退化。

正如暴露偏差分析所预测的那样 [33, 34]。我们尚未单独识别出这一项，也不声称对其进行了分解。

长度敏感性本身就是可写性信号。 TM评分与链长之间的相关性为 -0.054 （DPLM-2），-0.072 （ESM3），-0.102 （GCP-VQVAE），-0.167 以及 -0.187 （FoldToken4-l10 和 l12），而 -0.264（ProFiT）与可写性本身的排序一致（表1）。编码空间对长程结构的迁移能力较差的分词器，其质量也会随链长增加而衰减，这正是根据第 6 节所能预测的行为。

范围。 生成上限为 1024 个结构令牌。334 条评估链中恰有一条超过该长度，因此被排除；所以，500 个残基处的断崖位于生成范围之内，并非上限造成的伪影。本小节中的所有数字均针对 FoldBench334 上 2B 词元计算预算下的贪心解码；我们未测试断崖是否会随计算资源变化而移动，因为这需要在每个检查点重新评估长度分桶。

8 两个预注册的负面结果

8.1 直接监督比特不会带来任何改变

干预。 如果可写性源于模型落在比特相邻的码字上，那么训练模型预测比特 而不是码字索引应该会有所帮助。我们在运行实验前预注册了两种实现及其决策阈值。第一种（D2H）保留索引交叉熵，但使用 $\epsilon = 0.2$ 在汉明距离较近的码字上平滑目标。第二种（D2B）将索引目标替换为对比特边缘分布逐比特计算的二元交叉熵，这正是纠错输出码所规定的目标 [50]。两者均使用 DPLM-2 词元，除此之外与 DPLM-2 组别完全相同。

两者均为零效应。 D2H 训练至 20亿个 token，并在全部十二个单元上进行评估：与基线相比的所有差异均在 ± 0.016 TM评分 之内，预注册的主要比较（2B 时的 FoldBench334）结果为 -0.0001，配对自助法区间为 [51] $[-0.0078, +0.0075]$，胜率为 50.8%。D2B 在训练至 1.39B 个 token 后停止；在其 0.5B 检查点上，三个评估集的结果分别为 $+0.002$、$+0.014$ 和 -0.002。没有任何单元的变化超出噪声范围。

预注册机制检查正是使零效应结果具有可解释性的关键。同一计划要求，除非干预能够明确改变误差几何，否则必须舍弃零效应结果。对于 D2H，干预并未改变：误差汉明比为 0.507，而基线为 0.506；错误率也基本不变，分别为 92.3% 和 92.4%。因此，该次运行什么也没有检验，我们将其报告为无效结果，而不是将其视为反对比特监督的证据；$\epsilon = 0.2$ 这只不过是过于温和的标签平滑，梯度无法触及这些比特。D2B 确实改变了机制，在 0.5B 时将该比率从 0.537 推至 0.526，因此其零效应结果具有信息价值。其解释正是第 6节的饱和结果：在普通索引监督下，这些干预所针对的量在 1.5B 个词元时便不再改善，因此几乎没有剩余内容可供学习。

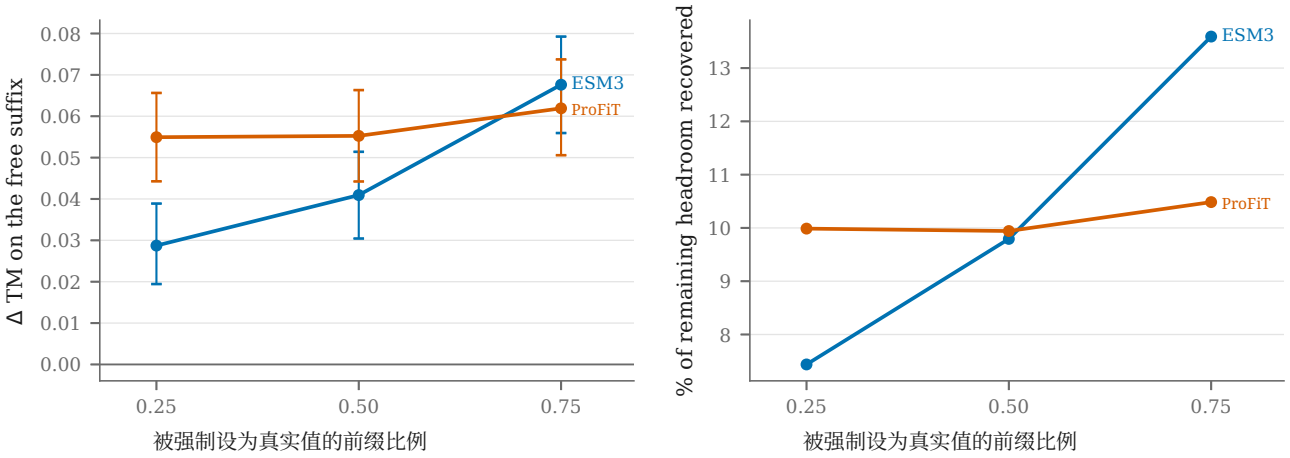


图5. 真实前缀补救，仅根据模型自行生成的后缀进行评分。左：绝对 Δ TM评分及配对自助法95%区间。右：将相同效果表示为各分词器达到其重建上限所剩余空间的比例，从而消除了ProFiT因起点较低而获得的优势。

一个价值超过干预本身的副产品。D2B 在 0.5B 个词元时的令牌错误率更高，为 95.1%，而基线为 93.9%，且TM评分相同。令牌准确率与结构质量彼此脱钩：模型写出哪个码字远不如它与正确码字相差多远重要。任何将 tokenaccuracy 作为结构质量代理指标的结构分词器评估，测量的都是错误的对象。

8.2 强制使用正确前缀无法挽救最差的分词器

正在检验的假设。ProFiT 的前缀一致性接近于零（$S_5 = 0.0011$ 相比之下，ESM3 为 0.5927；该数值是在主轴规范化输入上，以半长度前缀进行测量，以匹配训练条件）：对截断链重新编码会改变几乎每个词元。其可写性较差的一个自然解释是，自回归生成必须确定那些无法从前缀中推断出的早期词元。如果这就是其中的机制，那么将前缀替换为真实值应该能够挽回大部分损失。

测量必须排除这份馈赠。我们将前 $q \in \{0.25, 0.5, 0.75\}$ 个词元强制设为真实值，并且只对模型自行生成的残基进行评分，仅在该片段上进行叠合，并将其与普通贪心解码下的同一片段进行比较，同时采用配对自助法。对整条链进行评分会把复制的前缀计入模型表现：在 $q = 0.25$ 时，这会将 ESM3 的效果从 +0.029 夸大到 +0.078，且夸大程度会随着 q增加。

ProFiT 没有响应；ESM3 有响应。图 5 给出了结果。在三个设置（+0.055、 +0.055、 +0.062 TM；相对于其上限剩余空间的 10.0%、9.9%、10.5%）下，ProFiT 的增益都持平，而 ESM3 单调上升（+0.029、 +0.041、 +0.068；7.4%、9.8%、13.6%），并在 $q = 0.75$超过了 ProFiT。给予 ProFiT 四分之三的答案，其效果并不比给予它四分之一更好。六个区间均不包含零，因此该效应确实存在，但幅度很小；失败的是这一假设，而不是测量。

我们撤回对前缀一致性的因果解读。低 S_5 是与低可写性共同出现的症状，而不是其成因；ProFiT 的瓶颈在于表示本身（第 6节），而不是自回归解码所迫使模型作出的承诺。对链上错误发生位置的独立测量早已未能支持这一承诺假设，而本次干预也得出了相同结论。具体而言，这些数据并不支持将双向解码或扩散解码 [23, 54] 作为补救措施推荐；若依据因果解读，本应得出这一建议。

9 通用能力得以保留，而训练损失无法告诉你这一点

防护栏。 扩展词汇表并继续预训练，可能会因灾难性遗忘而直接损害基础模型 [55, 56], 这是领域自适应预训练的已知风险 [57], , 从而为任何可写性差异提供另一种解释。我们在预注册的 MMLU-Pro 350 题分层子集上评估全部七个 20亿令牌检查点 [58, 59], , 使用上游项目的题目列表和评分器，不作任何改动，以确保数值能够与早期轮次进行比较。基础模型的评分为 0.7743；扩展词汇表但不进行训练后，逐题结果仍为 0.7743。

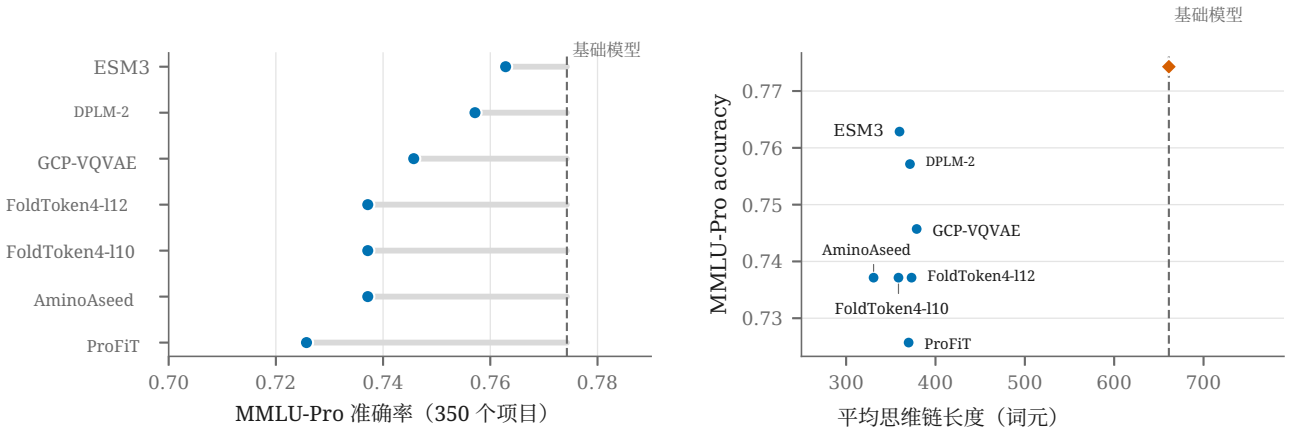


图 6. 20亿个 token 时的通用能力护栏。左：相对于基础模型（虚线）的 MMLU-Pro 准确率。右：准确率与平均思维链长度的关系；菱形代表基础模型。无论使用哪种分词器，每个组别的推理长度都减少了约一半。

没有任何组别发生崩溃。 准确率范围从 0.7629（ESM3，−1.14pp）到 0.7257（ProFiT，−4.86pp），解析失败率介于 0.00% 和 0.86% 之间，而基础模型为 0.29%（表 6）；没有任何结果类似于格式崩溃，因此另一种解释被排除。

模型的能力并未减弱，只是表达更简洁了。 平均思维链 [60] 长度从基础模型的 661 个词元降至各组别的 331–379 个词元，降幅为 43–50%，且与分词器没有系统性关联（图6）。由于七个组别的表现都相同，这是训练方案的属性，而非结构令牌的属性。仅报告准确率会完全掩盖这一点，原因与仅报告 TM评分会掩盖第5节中的失败相同。

混合内文本损失无法反映这一点。七个分组的最终通用文本训练损失分别为 2.1663、2.1673、2.1682、2.1691、2.1692、2.1694 和 2.1695，差异范围为 0.003，而它们的 MMLU-Pro 准确率相差 3.7 个百分点。在混合数据集文本上的训练损失衡量的是拟合程度，而非保留能力；若将其用作安全护栏，它会得出七个分组都同样安全的结论。

我们并未声称可写臂越多，保留能力就越强。原始 TM 与准确率下降之间存在皮尔逊相关系数+0.825 （置换检验 $p = 0.026$），但两两比较均无显著差异：差异最显著的一对是 ESM3 与 ProFiT，相差 3.7 个百分点，其精确 McNemar检验结果为 [61] $p = 0.079$；七个分组中有三个得分相同（258/350）。包含 350 个项目的集合，其二项标准误约为 2.3 个百分点。我们将这种关联作为观察结果报告，而非研究发现；要得出定论，需要使用完整的 12,032 项基准测试。

10 局限性与我们不作出的主张

我们明确说明每项主张的边界，因为第 3 节中的对照实验正是为了避免对本基准测试作出过度解读。

我们不声称存在最好的分词器。我们声称，在此基础模型、此配方和 20 亿词元的计算预算下，DPLM-2 是八种分词器中最具可写入性的。除两个退化组别外，没有任何组别达到饱和。
我们不声称 ESM3 优于 GCP-VQVAE。随着计算资源增加，这一差距从 0.163 缩小到 0.091；超过我们的计算预算后，差距很可能会消失；在此处，只有计算量匹配比较才有意义 [62, 63]。报告结果仅适用于 ≤2B 个词元。

我们不对第三等级进行排序，因为其中三个成员之间的差异小于噪声下限；而且**我们不会仅凭高兑现率就将其视为成功**：对于 AminoAseed 和 VanillaVQ，这反映的是一个狭窄的下限至上限区间。

我们并不声称比特局部性具有因果性。它是证据最充分的相关因素，并有两项独立的机制测量支持；能够确立因果关系的干预结果为零效应（第 8节），我们报告了这一零效应及其解释，而没有将其省略。

我们并不声称可写臂保留了更强的通用能力，因为组间关联在两两比较中并不显著，而且**我们不会将结论推广到单一配置之外**：一个基础模型、一种规模、一个混合数据集、一套配方，且仅使用贪心解码；没有关于采样或 best-of-n 的结果。

我们并不声称存在普遍适用的长度惩罚。在 500 个残基以下，最具可写入性的组别没有值得报告的长度效应（$r = -0.054$）；这一失效仅出现在长度为 500 个残基及以上的链中，而此时仅有 26 个目标可用（第 7.1 节）。我们也未能分离出在控制远程捕获后仍然存在的长度项，也未测试 500 个残基处的断崖式变化是否会随计算资源而移动。

噪声下限是一个下界，来自一个组别的一次重复实验（$n = 9$ 个配对单元）；每个组别进行三至五次重复实验、每次耗时30–50 GPU小时，才是合理的估计。**无参考面板并不完整**：二级结构组成需要完整的主链，因此对于仅CA解码器，我们报告键有效性、冲突率和回转半径，但不报告DSSP组成。**差距分解的第二项是外推结果**，超出了观测到的远程接触精度范围，且其中三项彼此重叠。

11 结论

结构分词器的选择取决于其可读性，同时也将其视为必须被写入的对象。在其他条件全部固定的情况下，我们直接测量后一属性，发现它在八个分词器之间的TM评分从0.662到0.378不等；这一差距反映了读者会认为属于正确折叠结构的主链与不会被如此认为的主链之间的距离。该属性既不能由码本大小预测，也不能由重建保真度预测，而是取决于编码空间是否在局部保持结构，从而使模型不可避免的错误在结构上代价较小。

在实践中，这意味着需要进行三项改变。为生成式接口选择分词器时，不要只衡量完美往返能够保留什么，还要衡量一个码字错误会带来多大代价。通过生成式解码器报告结果时，应在同一组目标上，用随机输入下限和真实值上限为每个数值划定范围，并以绝对值给出上限，以免将狭窄区间内的归一化分数误认为成功。此外，绝不能将令牌准确率作为结构质量的代理指标：我们展示了一个变体，它提高了令牌错误率，却保持结构质量不变。

我们对与专业折叠模型之间差距的分解，也界定了剩余改进手段的价值：更好的解码在实践中只能带来约 0.01 的 TM评分提升，将输入中已有的远程接触传递出来则可带来 0.09 的提升，而 TM评分约为 0.84 的离散化下限是两者共同受制的因素。因此，要缩小剩余差距，就需要更好的量化编码空间 [11, 14, 20] 或连续精修阶段 [26], ，而不是另一种解码启发式方法。

参考

[1] Hayes, T., Rao, R., Akin, H. 等. 用语言模型模拟5亿年的进化。《科学》，387(6736): 850–858，2025年。[2] Wang, X., Zheng, Z., Ye, F., Xue, D., Huang, S., 和 Gu, Q. DPLM-2：多模态扩散蛋白质语言模型。载于《国际学习表征会议》，2025年。[3] Gao, Z., Tan, C., Wang, J., Huang, Y., Wu, L., 和 Li, S. Z. FoldToken：通过向量量化及其他方法学习蛋白质语言。载于《AAAI人工智能会议论文集》，第39卷，第219–227页，2025年。doi: 10.1609/ aaai.v39i1.31998。[4] Yuan, X., Wang, Z., Collins, M., 和 Rangwala, H. 蛋白质结构分词：基准测试与新配方。载于第42届《国际机器学习会议论文集》，PMLR第267卷，2025年。arXiv:2503.00089。[5] Gaujac, B., Donà, J., Liu, L., Sarkissian, T., Beguir, K., 和 Pierrot, T. 学习蛋白质结构的语言。《arXiv预印本 arXiv:2405.15840》，2024年。[6] Cao, X., Feng, Z., Zhang, Z., 和 Gao, Y. Q. Protokens：一种通过机器学习获得的语言，用于紧凑且信息丰富地编码蛋白质三维结构。《化学科学》，2025年。doi: 10.1039/D5SC02055G。[7] Su, J., Han, C., Zhou, Y., Shan, J., Zhou, X., 和 Yuan, F. SaProt：采用结构感知词汇表的蛋白质语言建模。载于《国际学习表征会议》（ICLR），2024年。 https://openreview.net/forum?id=6MRm3G4NiU。[8] Ma, Z., Fan, C., Wang, Z. 等. ProtTeX：借助大语言模型进行蛋白质的结构上下文推理与编辑。《arXiv预印本 arXiv:2503.08179》，2025年。[9] Li, M., Tan, Y., Ma, X., Zhong, B., Yu, H., Zhou, Z., Ouyang, W., Zhou, B., Tan, P., 和 Hong, L. ProSST：采用量化结构与解耦注意力的蛋白质语言建模。载于《神经信息处理系统进展》（NeurIPS），第37卷，2024年。[10] Heinzinger, M., Weissenow, K., Sanchez, J. G., Henkel, A., Mirdita, M., Steinegger, M., 和 Rost, B. 用于蛋白质序列和结构的双语语言模型。《NAR基因组学与生物信息学》，6(4):lqae150，2024年。doi: 10.1093/nargab/ lqae150。

[11] Yu, L., Lezama, J., Gundavarapu, N. B., Versari, L., Sohn, K., Minnen, D., Cheng, Y., Gupta, A., Gu, X., Hauptmann, A. G., Gong, B., Yang, M.-H., Essa, I., Ross, D. A., 和 Jiang, L. 语言模型胜过扩散模型：分词器是视觉生成的关键。载于《国际学习表征会议》（ICLR），2024年。[12] Zeghidour, N., Luebs, A., Omran, A., Skoglund, J., 和 Tagliasacchi, M. SoundStream：端到端神经音频编解码器。《IEEE/ACM 音频、语音与语言处理汇刊》，30:495–507，2022年。[13] Défossez, A., Copet, J., Synnaeve, G., 和 Adi, Y. 高保真神经音频压缩。《机器学习研究汇刊》（TMLR），2023年。[14] Zhang, J., Yang, B. M.-P., Gong, J., Fan, X., Ellington, T., Xing, E. P., 和 Tan, L. S. 平衡蛋白质结构分词器中的局部性与重建。《bioRxiv预印本》2024.12.02.626366，2024年。[15] Zhang, Y. 和 Skolnick, J. 用于自动评估蛋白质结构模板质量的评分函数。《蛋白质：结构、功能与生物信息学》，57(4):702–710，2004年。[16] Mariani, V., Biasini, M., Barbato, A., 和 Schwede, T. lddt：一种利用距离差异检验比较蛋白质结构和模型的局部无叠合评分。《生物信息学》，29(21):2722–2728，2013年。[17] Zemla, A. LGA：一种用于发现蛋白质结构三维相似性的方法。《核酸研究》，31(13):3370–3374，2003年。[18] Buttenschoen, M., Morris, G. M., 和 Deane, C. M. PoseBusters：基于人工智能的对接方法无法生成物理有效构象，也无法泛化到新序列。《化学科学》，15(9):3130–3139，2024年。[19] van den Oord, A., Vinyals, O., 和 Kavukcuoglu, K. 神经离散表征学习。载于《神经信息处理系统进展》，第30卷。Curran Associates, Inc.，2017年。 https://proceedings.neurips.cc/paper_files/paper/2017/hash/7a98af17e63a0ac09ce2e96d03992fbc-Abstract.html。[20] Mentzer, F., Minnen, D., Agustsson, E., 和 Tschannen, M. 有限标量量化：让 VQ-VAE 更简单。载于《国际学习表征会议》（ICLR），2024年。[21] Hardy, R., Zhang, X., 和 Rajpurkar, P. ProTok：通过模块化多码本量化实现统一的蛋白质分词。arXiv预印本 arXiv:2610.18481，2026年。[22] van Kempen, M., Kim, S. S., Tumescheit, C., Mirdita, M., Lee, J., Gilchrist, C. L. M., Söding, J., 和 Steinegger, M. 使用 Foldseek 快速、准确地搜索蛋白质结构。《自然·生物技术》，42(2):243–246，2024年。doi: 10.1038/ s41587-023-01773-0。[23] Chang, H., Zhang, H., Jiang, L., Liu, C., 和 Freeman, W. T. MaskGIT：掩码生成式图像变换器。载于《IEEE/CVF 计算机视觉与模式识别会议》（CVPR），2022年。[24] Jumper, J., Evans, R., Pritzel, A. 等. 使用 AlphaFold 高精度预测蛋白质结构。《自然》，596: 583–589，2021年。[25] Baek, M., DiMaio, F., Anishchenko, I. 等. 使用三轨神经网络准确预测蛋白质结构与相互作用。《科学》，373(6557):871–876，2021年。doi: 10.1126/science.abj8754。[26] Abramson, J., Adler, J., Dunger, J. 等. 使用 AlphaFold 3 准确预测生物分子相互作用的结构。《自然》，630:493–500，2024年。[27] Lin, Z., Akin, H., Rao, R. 等. 使用语言模型进行进化尺度的原子级蛋白质结构预测。《科学》，379(6637):1123–1130，2023年。[28] Kryshtafovych, A., Schwede, T., Topf, M., Fidelis, K., 和 Moult, J. 深度学习在计算结构生物学中的进展与瓶颈：CASP 第十六轮。《蛋白质：结构、功能与生物信息学》，94(1):5–14，2026年。[29] Yuan, Q., Tian, C., 和 Zhang, Y. CASP16 蛋白质单体结构预测评估。《蛋白质：结构、功能与生物信息学》，94(1):86–105，2026年。[30] Xu, S., Zhang, O., Wang, Y. 等. 使用 FoldBench 对全原子生物分子结构预测进行基准测试。《自然·通讯》，2025年。doi: 10.1101/2025.05.22.655600。[31] Holtzman, A., Buys, J., Du, L., Forbes, M., 和 Choi, Y. 神经文本退化的奇特案例。载于《国际学习表征会议》（ICLR），2020年。[32] Welleck, S., Kulikov, I., Roller, S., Dinan, E., Cho, K., 和 Weston, J. 使用非似然训练进行神经文本生成。载于《国际学习表征会议》（ICLR），2020年。[33] Bengio, S., Vinyals, O., Jaitly, N., 和 Shazeer, N. 用于循环神经网络序列预测的计划采样。载于《神经信息处理系统进展》（NeurIPS），2015年。[34] Ranzato, M., Chopra, S., Auli, M., 和 Zaremba, W. 使用循环神经网络进行序列级训练。载于《国际学习表征会议》（ICLR），2016年。[35] Kabsch, W. 和 Sander, C. 蛋白质二级结构词典：氢键与几何特征的模式识别。《生物聚合物》，22(12):2577–2637，1983年。[36] Zhang, Y. 和 Skolnick, J. TM-align：一种基于 TM-score 的蛋白质结构比对算法。《核酸研究》，33(7):2302–2309，2005年。doi: 10.1093/nar/gki524。

[37] Reimers, N. 和 Gurevych, I. 报告分数分布会带来不同：LSTM网络在序列标注中的性能研究。载于自然语言处理实证方法会议（EMNLP），2017年。[38] Henderson, P., Islam, R., Bachman, P., Pineau, J., Precup, D. 和 Meger, D. 值得关注的深度强化学习。载于AAAI人工智能会议，2018年。[39] Dodge, J., Gururangan, S., Card, D., Schwartz, R. 和 Smith, N. A. 展示你的工作：改进实验结果的报告方式。载于自然语言处理实证方法会议（EMNLP），2019年。[40] Esm3结构分词器。 https://github.com/evolutionaryscale/esm。于2026年访问；确切的提交版本和权重校验和记录在发布的档案中。[41] Dplm-2结构分词器。 https://github.com/bytedance/dplm.git，提交版本 8a2e15e534。于2026年访问；确切的提交版本和权重校验和记录在发布的档案中。[42] ProFiT。 https://github.com/QDKStorm/ProFiT.git，提交版本 be4ff7cb6b。于2026年访问；确切的提交版本和权重校验和记录在发布的档案中。[43] FoldToken4。 https://github.com/A4Bio/FoldToken_open.git，提交版本 fed13c4d11。于2026年访问；确切的提交版本和权重校验和记录在发布的档案中。[44] Structtokenbench（AminoAseed、VanillaVQ）。 https://github.com/KatarinaYuan/StructTokenBench.git，提交版本2681b240f6。于2026年访问；确切的提交版本和权重校验和记录在发布的档案中。[45] GCP-VQVAE。 https://github.com/mahdip72/vq_encoder_decoder.git，提交版本 68c4c284fe。于2026年访问；确切的提交版本和权重校验和记录在发布的档案中。[46] Yu, J., Li, X., Koh, J. Y., Zhang, H., Pang, R., Qin, J., Ku, A., Xu, Y., Baldridge, J. 和 Wu, Y. 采用改进VQGAN的向量量化图像建模。载于国际学习表征会议（ICLR），2022年。[47] Robin, X., Haas, J., Gumienny, R., Smolinski, A., Tauriello, G. 和 Schwede, T. 连续自动模型评估（CAMEO）——关于结构预测方法全自动评估未来的展望。《蛋白质：结构、功能与生物信息学》，89(12):1977–1986，2021年。doi: 10.1002/prot.26213。[48] Steinegger, M. 和 Söding, J. MMseqs2支持灵敏的蛋白质序列搜索，用于分析海量数据集。《自然·生物技术》，35(11):1026–1028，2017年。doi: 10.1038/nbt.3988。[49] Hamming, R. W. 检错码与纠错码。《贝尔系统技术期刊》，29(2):147–160，1950年。[50] Dietterich, T. G. 和 Bakiri, G. 通过纠错输出码解决多类学习问题。《人工智能研究期刊》，2:263–286，1995年。[51] Efron, B. 自助法：重新审视刀切法。《统计年刊》，7(1):1–26，1979年。[52] Rao, R., Meier, J., Sercu, T., Ovchinnikov, S. 和 Rives, A. Transformer蛋白质语言模型是无监督结构学习器。载于国际学习表征会议（ICLR），2021年。[53] Wang, X., Wei, J., Schuurmans, D., Le, Q., Chi, E., Narang, S., Chowdhery, A. 和 Zhou, D. 自洽性提升语言模型中的思维链推理。载于国际学习表征会议（ICLR），2023年。[54] Wang, X., Zheng, Z., Ye, F., Xue, D., Huang, S. 和 Gu, Q. 扩散语言模型是多功能的蛋白质学习器。载于第41届国际机器学习会议，《机器学习研究论文集》第235卷，第52309–52333页。PMLR，2024年。 https://proceedings.mlr.press/v235/wang24ct.html。[55] Kirkpatrick, J., Pascanu, R., Rabinowitz, N. 等。克服神经网络中的灾难性遗忘。《美国国家科学院院刊》，114(13):3521–3526，2017年。doi: 10.1073/pnas.1611835114。[56] Luo, Y., Yang, Z., Meng, F., Li, Y., Zhou, J. 和 Zhang, Y. 持续微调期间大语言模型灾难性遗忘的实证研究。《IEEE音频、语音与语言处理汇刊》，33:3776–3786，2025年。doi: 10.1109/TASLPRO.2025.3606231。[57] Gururangan, S., Marasović, A., Swayamdipta, S., Lo, K., Beltagy, I., Downey, D. 和 Smith, N. A. 不要停止预训练：使语言模型适应领域和任务。载于计算语言学协会年会（ACL），2020年。[58] Wang, Y., Ma, X., Zhang, G. 等。MMLU-Pro：更稳健、更具挑战性的多任务语言理解基准。载于神经信息处理系统进展（数据集与基准专题），第37卷，第95266–95290页。Curran Associates, Inc.，2024年。[59] MMLU-Pro。MMLU-Pro：更稳健、更具挑战性的多任务语言理解基准。 https://huggingface.co/datasets/TIGER-Lab/MMLU-Pro。在一个预注册的350题分层子集上进行评估，使用上游题目列表和评分器。[60] Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E. H., Le, Q. V. 和 Zhou, D. 思维链提示能够引导大语言模型进行推理。载于神经信息处理系统进展，2022年。[61] McNemar, Q. 关于相关比例或百分比之差的抽样误差的注记。《Psychometrika》，12(2):153–157，1947年。

[62] Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J. 和 Amodei, D. 神经语言模型的缩放定律。*arXiv*预印本 *arXiv:2001.08361*，2020年。[63] Hoffmann, J., Borgeaud, S., Mensch, A. 等。训练计算最优的大语言模型。载于神经信息处理系统进展（*NeurIPS*），2022年。

附录

A 测量事件及其代价

第 3 节中的每项控制措施，都源于缺少它曾导致错误结果。我们记录这四起事件，是因为由此形成的规则是本研究中可迁移的部分，也因为这些问题都是在较晚阶段才被发现，值得一述。

报告了低于下限的能力分数。 在早前一轮中，我们将 0.168 的贪心 TM评分报告为能力结果。事后，我们在相同样本量下，针对相同目标测得该分词器的随机令牌下限为 0.185。模型的输出比经同一解码器处理的随机输入还差。报告中已经记载了两项足以发现这一问题的事实：局部几何由解码器而非模型保证；贪心解码在约 25% 的位置重复输出同一令牌。但我们都没有追究到底。采纳的规则： 下限是结果的一部分，而不是后续补充；发现异常后，应追查到底，而非仅仅记录下来。

评估这份礼物。 前缀补救脚本的标头声称它只对未经强制的后缀进行评分；但其实现调用了通用评分器，而该评分器会对整条链进行评分。由于强制前缀是复制的真实值，报告中的效应在 $q = 0.25$ 处被夸大了 2.7 倍，在更大的 q处则夸大得更多。脚本自身的日志行写着“整条链 TM”；但标头和日志从未相互核对。已采纳的规则：当注释与日志不一致时，以日志为规范；现在，该脚本结束时会自动运行仅后缀评分。

没有文件支撑的数字。 第 6 节所依据的误差几何测量最初只存在于终端输出中。后来，这些测量被重新计算并以 JSON 形式归档；此时，将其从一个检查点扩展到全部四个检查点无需额外成本，并产生了解释无效干预的饱和结果。已采纳的规则： 报告中的每个数字都必须对应一个已归档文件，分析脚本应写入结果，而不是将结果打印出来。

一次过于温和、无法检验任何东西的干预。 第一个比特监督变体使用了 $\epsilon = 0.2$，其选择是为了与现有的标签平滑惯例保持一致。它使误差汉明比从 0.506 变为 0.507。预注册机制检查发现了这一点，并判定该次运行无效；如果没有这项检查，一个看似干净的零效应本可能会被作为比特监督没有帮助的证据发表。 采纳的规则：保守的超参数并不意味着安全，每项干预都必须附带一项预注册检查，以确认它确实产生了作用。

自我纠正。 随着数据不断积累，研究期间修订了四项结论：码本大小的作用（在最终确定为“没有能够突破噪声下限的效应”之前，修订了四次）、两种分词器的相对表现（三次）、对编码瓶颈的诊断（错误判断了究竟是哪种资源达到饱和），以及一项只在三组评估数据中的一组上成立、后来被撤回的结果。 采纳的规则：报告记录每项判断是在何时、依据哪些数据作出的。

B 可复现性

发布内容。 全部 131 个评估单元均已归档，每个单元都包含汇总指标、逐目标得分、无参考面板以及原始生成令牌流。最后一项让档案真正发挥作用：第 5 节的几何面板、第 6 节的误差几何轨迹，以及第 8 节的仅后缀补救得分，都是事后基于归档的生成结果计算得出的；这些结果来自数周前训练的检查点，计算时无需加载任何模型。此外，档案还包括 11 条逐步记录的训练曲线、涵盖所有下限、上限、稳定性和容差测量的跨分词器评分卡、各组别混合数据集的数据构建统计，以及 MMLU-Pro 的逐项预测。

第三方组件。 每个分词器都固定对应一个代码库提交，其权重则固定对应一个 SHA-256 校验和；二者均记录在归档的来源目录中。模型检查点（每个组别导出四个，每个约 17 GB）不予分发；训练混合数据集可根据记录的构建参数重新生成，而检查点无法重建，这一点已在归档中注明。

已知缺口。 ESM3 在发布网格中的四个计算点来自基于本文处理流程进行的一次重新运行，而非原始运行；原始运行没有 1.5B 导出版本，其训练检查点也已被轮换移除。原始运行的三个点被单独保留，仅用于噪声估计。VanillaVQ 的 1.5B 和 2B 点来自恢复其 1B 检查点，而非连续运行。二者均不影响本文所作的比较，并已在归档中标记。

C 完整结果表

表4。 完整网格：所有分词器、计算预算和评估集的贪心解码 TM-score（共96个单元格）。同一列内小于0.013 TM的比较结果低于实测的运行间噪声下限（第3节）。

分词器	计算预算	FoldBench334 ↑	CASP16 ↑	CAMEO2025 ↑
DPLM-2	0.5B	0.5919	0.4888	0.6791
	1B	0.6455	0.5289	0.7538
	1.5B	0.6564	0.5521	0.7787
	2B	0.6623	0.5689	0.7999
ESM3	0.5B	0.5478	0.4577	0.6562
	1B	0.6025	0.4910	0.7455
	1.5B	0.6155	0.5123	0.7836
	2B	0.6162	0.5482	0.7946
GCP-VQVAE	0.5B	0.3847	0.3324	0.4540
	1B	0.4771	0.4102	0.6102
	1.5B	0.5179	0.4263	0.6810
	2B	0.5254	0.4407	0.7106
FoldToken4-l12	0.5B	0.4101	0.3617	0.4449
	1B	0.4841	0.4299	0.5827
	1.5B	0.5283	0.4604	0.6487
	2B	0.5281	0.4586	0.6761
FoldToken4-l10	0.5B	0.3740	0.3556	0.4185
	1B	0.4639	0.4019	0.5395
	1.5B	0.5129	0.4521	0.6344
	2B	0.5270	0.4532	0.6645
ProFiT	0.5B	0.3303	0.3033	0.3611
	1B	0.4020	0.3728	0.4923
	1.5B	0.4325	0.3885	0.5446
	2B	0.4467	0.3845	0.5678
AminoAseed	0.5B	0.3692	0.2807	0.4170
	1B	0.3880	0.3094	0.4581
	1.5B	0.3941	0.2996	0.4668
	2B	0.3952	0.3036	0.4727
VanillaVQ	0.5B	0.3672	0.3291	0.3767
	1B	0.3808	0.3252	0.3987
	1.5B	0.3769	0.3246	0.3992
	2B	0.3781	0.3272	0.4008

表 5. 在 FoldBench334 上测得的分词器评分卡，未训练任何模型。 S_1 表示通过该分词器自身的解码器得到的重建上限，S_4 表示均匀随机下限。键有效性列表示在上限和下限情况下，连续 CA–CA距离位于 $[3.6, 4.0]$ Å 范围内的比例。 S_3 表示随机替换 10% 真实标记后保留的 TM评分。 S_5 表示前缀一致性。

		TM		CA–CA有效性			
分词器	K	$S_1 \uparrow$	$S_4 \downarrow$	上限	下限	S_3@10% ↑	$S_5 \uparrow$
DPLM-2	8192	0.9519	0.1766	0.996	0.806	0.890	0.2100
ESM3	4096	0.9886	0.1847	0.975	0.087	0.671	0.5927
GCP-VQVAE	4096	0.9891	0.1029	0.991	0.149	0.839	0.0490
FoldToken4-l12	4096	0.9620	0.1436	0.903	0.026	0.853	0.0205
FoldToken4-l10	1024	0.9594	0.1508	0.891	0.028	0.849	0.0284
ProFiT	8192	0.9902	0.1680	0.995	0.793	0.912	0.0011
AminoAseed	512	0.4816	0.1847	0.337	0.126	0.445	0.5375
VanillaVQ	512	0.4226	0.1885	0.246	0.083	0.407	0.6545

表 6. 通用能力护栏：在 20 亿令牌检查点上，使用官方 5-shot 思维链协议，对预注册的 350 项分层子集进行 MMLU-Pro 评测。成对差异均无显著性（精确 McNemar 检验）；参见第 9 节。

Arm	TM评分 @2B	准确率 ↑	Δ 与基准相比	解析失败 ↓	平均思维链令牌数
基础模型	–	0.7743	–	0.29%	661
DPLM-2	0.6623	0.7571	-1.71个百分点	0.86%	371
ESM3	0.6162	0.7629	-1.14个百分点	0.29%	360
GCP-VQVAE	0.5254	0.7457	-2.86个百分点	0.29%	379
FoldToken4-l12	0.5281	0.7371	-3.71pp	0.57%	373
FoldToken4-l10	0.5270	0.7371	-3.71pp	0.00%	359
ProFiT	0.4467	0.7257	-4.86 个百分点	0.00%	370
AminoAseed	0.3952	0.7371	-3.71 个百分点	0.29%	331