arXiv:2611.32801v1 [cs.LG] 24 Nov 2026

语言模型中集成不确定性的缩放定律

Siamak Ravanbakhsh、David Nitchi、Khang Ngo麦吉尔大学和
Mila {siamak.ravanbakhsh,david.nitchi,khang.ngo}@mila.quebec

摘要

大语言模型（LLM）的许多应用场景——从人工智能安全到探索和人机交互——都可以从认知不确定性估计中受益。我们认为，现有的贝叶斯深度学习解决方案通常局限于固定的模型系列，因而无法捕捉实际大语言模型规模下完整的认知不确定性。这是因为，模型系列有限模型容量所导致的误差，对于先验支持的模型之间的分歧而言是不可见的。我们将这一误差与众所周知的、随模型规模变化的大语言模型损失幂律缩放联系起来。随后我们证明，第二个幂律刻画了随着模型规模增大，集成内部不断缩小的分歧。相较于上述容量定律，这一分歧衰减得更慢，这表明集成能够捕捉越来越大比例的认知不确定性；与单个更大模型相比，集成作为扩展方向也变得更具竞争力。最后，第三个幂律刻画了这样一种集成多样性损失：集成成员在分支前共享的训练比例越大，集成多样性损失就越大。因此，除了容量误差之外，该幂律还揭示了 LoRA 集成，以及用于近似单一训练解附近不确定性的基于优化或二阶方法所面临的另一项限制。

从总体上看，这些发现表明，多样化的集成可能是一个很有前景的方向，既有助于扩展规模，也有助于在更大规模上进行不确定性量化。

1 引言

认知不确定性反映了模型在知识上的欠缺，而更多数据或更高的模型容量可以减少这种欠缺。贝叶斯推断通过对合理模型建立分布来表示这种不确定性。这种不确定性可以指导主动学习中信息性标签的选择（Houlsby 等，2011）、实验设计（Rainforth 等，2024）以及自主材料发现（Kusne 等，2020）。认识到知识空白还可以在分布偏移下指导安全探索，并帮助决定何时寻求人工监督（Amodei 等，2016）；而对人类偏好的不确定性则可以促使智能体将决策交由人工干预（Hadfield-Menell 等，2017；Russell，2019）。

许多贝叶斯深度学习（BDL）方法，包括变分推断、蒙特卡罗 dropout、拉普拉斯近似、基于优化的方法以及集成的诸多变体（

Graves，2011年；Gal 和 Ghahramani，2016年；Daxberger 等人，2021年；Maddox等，2019年；Lakshminarayanan 等人，2017年），在固定架构内近似权重的分布。因此，当最优模型位于该族之外时，其报告的不确定性会变得相当不可靠。族内可用的最佳预测器与“真实”预测器之间的损失差异，就是近似误差或容量损失。这种损失对于贝叶斯深度学习方法而言是不可见的，因为它位于先验的支持集之外。我们将不可见残差与贝叶斯深度学习方法可以访问的部分（即可见部分）区分开来（图 1）。经验语言模型损失

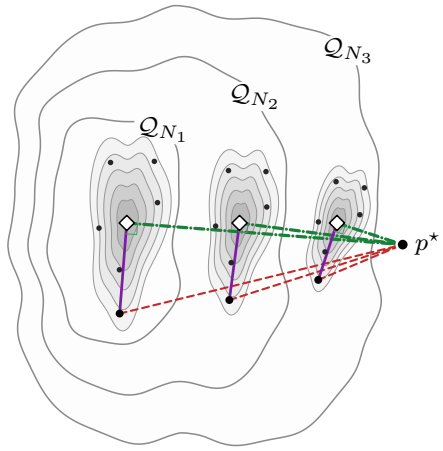


图 1 可见和不可见的认知不确定性。 示意图展示了表达能力不断增强的模型家族逐渐逼近真实的下一词元预测器 $p^\star$。阴影轮廓表示贝叶斯后验分布；圆点表示集成成员，菱形表示它们的池化预测（这些预测可能不属于同一族）。**红色虚线边** 表示高于不可约损失的总认知损失，该损失会随着模型规模增大，按照一种众所周知的幂律下降。**紫色实线边**表示通过集成恢复的可见部分；我们发现，该部分会随着模型规模增大而相当缓慢地按幂律下降。**绿色点划线边** 表示不可见部分，其中包括各成员共享的误差。它们不同的衰减率意味着，随着规模增大，紫色部分与绿色部分的比率会提高，集成能够捕获更大比例的认知不确定性。我们研究了这一现象及其影响。边长的平方表示加性损失分解。

定律包含一个依赖于大小的项（Kaplan 等人，2020年；Hoffmann 等人，2022年；Bahri 等人，2024年），可以宽泛地将其视为这种容量损失或不可见部分。[1]

现在来看误差或不确定性的可见部分，即使存在模型错设，贝叶斯后验也能够集中于支持度最高的预测器附近，后验方差会随着 D^{-1} 对于 D独立观测值而缩小（Kleijn 和 van der Vaart，2012年）。然而，这一集中性结果需要满足一些假设，而由于对称性和冗余参数化，这些假设在神经网络中并不成立（Watanabe，2013年）。在一项相关工作中，Rosso 等人（2025年）研究了跨数据和模型规模的不确定性缩放，并报告称，在视觉领域存在幂律趋势，而在语言模型微调中则表现出较弱的数据依赖性。我们认为，他们的结果不应被解读为 认知不确定性的缩放定律，而应被解读为可见部分的缩放定律。本文使用深度集成研究大语言模型（LLM）中不确定性的可见和不可见部分的行为，并确定其若干影响。具体而言，我们研究：

1. 集成分歧关于数据、模型和上下文规模的缩放定律。我们通过实证报告了一个浅幂律，涵盖各种开放权重模型以及我们自己的字符级模型族中的集成离散度（可见部分）：分歧按如下方式缩放：$N^{-\gamma}$ 其中 $\gamma \approx 0.08$–0.20，其中 N 表示模型参数数量；见图2b和图5。图3显示，这种集成离散度相比认知损失的不可见部分下降得更慢。我们比较了采用不同集成方式时的这种关系，包括几何平均与算术平均池化、训练词元预算和上下文长度；见图5和图13。超过中等训练预算后，可见部分变化不大（图5a）。超过某个中等长度后，对更长上下文的依赖也很弱（图5b；附录 D）。

2. 大规模集成的成本效益。可见部分和不可见部分的独立幂律预测，随着模型规模增大，可见比例也会提高（图3）。这种不确定性部分源于对不可约损失的不确定性。该预测与我们使用参考模型与成员模型之间的KL散度估计可约损失时发现的类似趋势相互印证（图6），表明在大规模条件下，集成能够更有效地捕获不断增加的认知不确定性。该缩放定律的另一个推论是，除了不确定性估计之外，在参数总数和数据集规模均固定的情况下，就损失而言，随着规模增大，集成将逐渐具备与采用更大模型相当的竞争力；见图7。

3. 通过后期分支进行集成的缩放定律。创建集成的一种经济高效的方法是后期分支：从同一个部分训练模型出发，使用不同的训练数据顺序继续训练多个模型，

[1]随模型规模变化的幂律可能归因于可见部分加不可见部分（即总可约损失），也可能仅归因于不可见部分。在我们的实验规模下，这两种对损失曲线进行建模的方式基本上无法区分；参见附录E.3和图17。

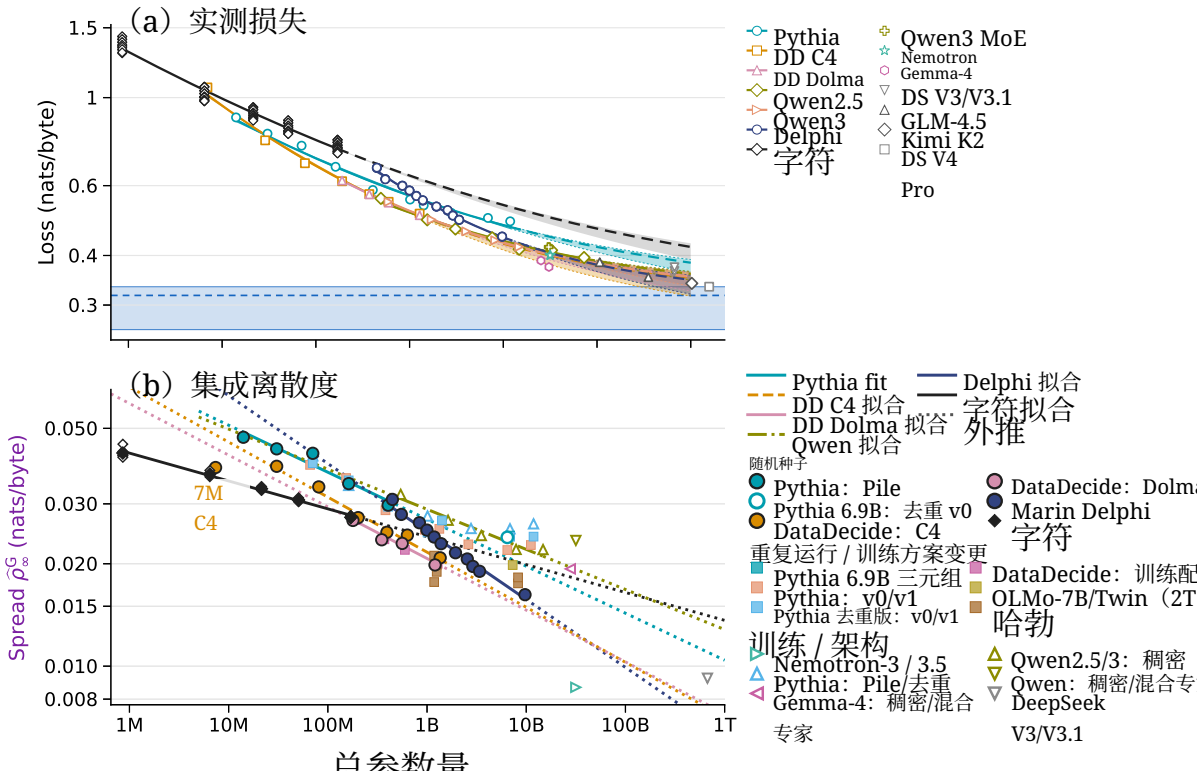


图2 模型损失和集成离散度随大小的变化。（a）在相同的近期文本片段上，对77个公开模型和98个字符级检查点的测量损失。拟合采用 $L(N) = E + A(N/10^9)^{-\eta}$，位于 $E = 0.31727$纳特/字节。彩色带对范围 $[0.26019, 0.33373]$（蓝色）内的损失底线重新拟合这些定律。（b）72组公开模型比较和35种字符模型设置的校正几何离散度，拟合为 $\widehat{\rho}_\infty^G(N) = c(N/10^9)^{-\gamma}$。点线表示外推结果。拟合指数（$\eta$，$\gamma$）分别为：Pythia（0.191，0.138）、DataDecide C4（0.253，0.162）、DataDecide Dolma（0.246，0.152）、Qwen2.5（0.222，0.118）、Qwen3（0.247，0.118）、Delphi（0.313，0.200）以及字符模型（0.164，0.082）。离散度拟合独立于 E。

这些模型都从同一个部分训练模型开始。问题在于，这种方法是否能带来与历史完全独立的“正规”集成相似的收益。我们的结果表明，后期分支会显著减小随机种子离散度，而且其变化遵循一个简单的幂律，可以预测；见图8。这意味着，在训练后期围绕单个随机种子生成不同变体的集成技术，其多样性会显著降低；其中包括 LoRA 集成（Wang等，2023年），以及使用训练轨迹的最后阶段来近似后验分布的方法（Maddox等，2019年）。

2 关系　损失、认知不确定性与集成之间

本节首先确立损失与认知不确定性之间的关系，然后根据集成是否能够获取认知部分的损失，将其划分为两个组成部分。这促使我们使用几何离散度来衡量集成中的分歧。本节的推导并不局限于种子集成，对更多样化的集成技术同样成立。

2.1 将认知不确定性视为可约损失

对于随机变量使用大写字母，设 X 为上下文，Y 为有限词汇表 $\mathcal{V}$ 中的下一个词元，其真实条件分布为 p_x，模型预测为 q_x。平均交叉熵损失可分为与模型无关的不可约部分和可约减部分，后者仅在 $q = p$ 时才会消失：

$$L(q) = \mathbb{E}_{X,Y}[-\log q_X(Y)] = \underbrace{\mathbb{E}_X H(p_X)}_{\text{irreducible loss } E=H(Y|X)} + \underbrace{\mathbb{E}_X \,\mathrm{KL}(p_X \| q_X)}_{\text{reducible loss } \mathcal{E}(q)}. \tag{1}$$

对于完全校准模型，总损失等于平均预测熵 $\mathbb{E}_X H(q_X)$。要理解这一点，请注意，完全校准意味着，给定预测 q_X 时的真实词元分布就是 q_X 本身，即 $\Pr(Y = \cdot \mid q_X) = q_X$——也就是说，每当模型为某个词元分配例如 70% 的概率时，在模型做出该确切预测的上下文中，该词元实际有 70% 的概率出现在下一位置。以 q_X 为条件并取期望，再利用 $\Pr(Y = \cdot \mid q_X) = q_X$，可得

$$L(q) = \mathbb{E}_X \left[-\sum_y \Pr(Y = y \mid q_X) \log q_X(y) \right] \overset{\text{calibration}}{=} \mathbb{E}_X H(q_X). \tag{2}$$

校准模型仍可能遗漏信息；例如，真实的一元预测无论上下文如何都始终预测词元边际频率，因此是完全校准的。接下来，我们将这种遗漏的信息识别为以条件互信息形式表示的认知不确定性。首先注意，对于完全校准模型，根据校准的定义，我们有

$$H(Y \mid q_X) = \mathbb{E}_X \, H\big(\Pr(Y = \cdot \mid q_X)\big) \overset{\text{calibration}}{=} \mathbb{E}_X H(q_X).$$

将其代入条件互信息的定义，可得

$$I(Y; X \mid q_X) = H(Y \mid q_X) - H(Y \mid X, q_X) = \mathbb{E}_X H(q_X) - H(Y \mid X).$$

其中我们使用了 $H(Y \mid X,\ q_X) = H(Y \mid X)$，因为 q_X 是 X 的函数。重新整理可得，将平均预测熵分解为偶然不确定性和认知不确定性（Bröcker，2009）：[2]

$$\underbrace{\mathbb{E}_X H(q_X)}_{\substack{\text{平均预测熵} \\ \text{总损失 } L(q),\ \text{在校准时}}} = \underbrace{H(Y \mid X)}_{\substack{\text{偶然不确定性} \\ \text{不可约损失 } E}} + \underbrace{I\ (Y;\ X \mid q_X)}_{\substack{\text{认知不确定性} \\ \text{可约损失 } \mathcal{E}\ (q)\ ,\ \text{在经过校准时}}} \tag{3}$$

这为我们理解大语言模型中的损失提供了清晰的解释；对于经过校准的模型，可约损失属于认知不确定性，因为它涉及上下文与下一个词元之间、但我们的模型未能捕捉到的信息；另见（Lahlou 等人，2023年；Kotelevskii 等人，2025年）。

至于在实际设置中校准假设是否合理，实证研究报告称，预训练语言模型在多种设置下都具有良好的校准效果（Desai 和 Durrett，2020年；Kadavath 等人，2022年；OpenAI，2023年；Spiess 等人，2025年；Nakkiran 等人，2026年）。我们的重新校准测试发现，对于字符级模型，留出对数损失的增益低于估计可约损失的 1%；在 45 个 Pythia、DataDecide、Qwen 和 Delphi 检查点中，该增益低于原始对数损失的 0。34%（附录 A）。

2.2 集成中认知不确定性的可见与不可见部分

几何平均离散度和算术平均离散度。 设 $S \geq 2$ 表示集成成员的数量 q_1，……，q_S，并设 $\bar{L}_S = S^{-1} \sum_s L(q_s)$ 表示它们的平均损失。我们考虑两种在集成成员上进行池化的标准方式，即几何平均池化和算术平均池化：

$$q^{\mathrm{A}}_{S,x}(y) = \frac{1}{S} \sum_s q_{s,x}(y), \qquad q^{\mathrm{G}}_{S,x}(y) = \frac{1}{Z_x} \exp\{S^{-1} \sum_s \log q_{s,x}(y)\}. \tag{4}$$

[2]未进行校准时，可约损失还包括一个校准误差：$\mathbb{E}_X \,\mathrm{KL}(c(q_X) \| q_X)$，其中 $c(q_X) = \Pr(Y = \cdot \mid q_X)$。

此处 $Z_x = \sum_{y \in \mathcal{V}}$指数函数$\{S^{-1} \sum_s$对数函数$q_{s,\ x}(y)\}$用于对几何池进行归一化。算术池化对概率取平均，而几何池化对逻辑值或对数概率取平均并重新归一化。对于任一选项$P \in \{\mathrm{A},\ \mathrm{G}\}$，离散度是成员平均损失与集成损失之差：$\rho_S^P = \bar{L}_S - L(q_S^P)$。成员的认知损失（平均而言）可以分解为离散度加上集成的可约损失：

$$\underbrace{\bar{L}_S - E}_{\text{单个成员的认知损失}} = \underbrace{\bar{L}_S - L(q_S^P)}_{\text{离散度 } \rho_S^P \text{（可见）}} + \underbrace{L(q_S^P) - E.}_{\text{集成的认知损失（不可见）}} \tag{5}$$

这个等式将可见部分与不可见部分分离开来。我们使用$f_S^P = \rho_S^P / \left(\bar{L}_S - E\right) \in [0, 1]$表示成员的认知不确定性中可通过集成看到的“比例”。尽管在实践中更常使用算术池化，但几何池化在概念上更接近于回归中采用高斯似然的均值池化；在几何池化中，离散度等于该池与其成员之间的平均散度[3]（Wood 等，2023）：

$$\rho_S^{\mathrm{G}} = S^{-1} \sum_s \mathbb{E}_X \, \mathrm{KL}(q_{S,X}^{\mathrm{G}} \| q_{s,X}) \tag{6}$$

这种几何离散度在高斯集成中起到均值方差的作用：当所有成员达成一致时，其值为零；同时，它等于集成所挽回的损失；对于特定上下文而言，其值与观测到的下一个词元无关，仅取决于预测结果。相比之下，对于算术池化，损失降低量取决于观测到的下一个词元。

基于分歧的贝叶斯主动学习（BALD）。BALD 提供了集成中第三种、也是最常见的一项分歧度量（Houlsby 等，2011）：

$$\underbrace{\mathbb{E}_X H(q_{S,X}^{\mathrm{A}})}_{\text{pooled predictive entropy}} = \underbrace{\frac{1}{S} \sum_s \mathbb{E}_X H(q_{s,X})}_{\text{mean member entropy}} + \underbrace{\frac{1}{S} \sum_s \mathbb{E}_X \, \mathrm{KL}(q_{s,X} \| q_{S,X}^{\mathrm{A}})}_{\text{ensemble disagreement } J_S \text{ (BALD)}}. \tag{7}$$

将 BALD 与离散度联系起来的一种方式是注意到：相对于平均成员，算术池化通过离散度降低对数损失，并通过 BALD 提高熵。虽然在某些设置中，BALD 被解释为认知不确定性，但在上述公式中，成员平均熵包含了所有成员共有的认知不确定性（如容量损失中的认知不确定性）；不同于公式（3）中的认知项，BALD 并不度量真正的认知不确定性。鉴于公式（5）中离散度与认知损失之间的直接联系，我们主要关注这一集成多样性度量，尤其是几何离散度，因为它形式更简单，并且与回归集成中的方差存在联系。不过，也可以从公式（7）和（6）中的 KL 形式看出 BALD 与几何集成之间的相似性；对于相近的预测，BALD 与几何离散度在领先的二次阶上相一致（Nielsen 和 Nock，2014），并且我们的评估显示，这三项度量表现出相似的定性行为。

将不同的集成规模置于同一尺度上。在实验中，我们希望对量进行归一化，以校正集成规模的影响。S 个独立同分布样本的均值周围的变异除以 S，其期望为总体方差的 $(S-1)/S$ 倍。对于相近的预测，离散度和 BALD 都具有主要的二次方差项（Oehlert，1992；Nielsen 和 Nock，2014）。因此，我们将这些度量乘以$S/(S-1)$，并使用下标 ∞ 表示近似的总体估计。例如，$\widehat{\rho}_\infty^{\mathrm{G}} = {}^{2}_{1}\rho_2^{\mathrm{G}}$表示两个集成成员。请注意，我们的集成通常并不具有独立同分布的成员，因此上述论证在实践中并不完全成立。

[3]对于任一汇聚规则，只需代入 L 项的定义，我们即可看到，离散度就是相对于真实条件分布的散度降低量：$\rho_S^P = S^{-1} \sum_s \mathbb{E}_X \, \mathrm{KL}(p_X \| q_{s,\ X}) - \mathbb{E}_X \, \mathrm{KL}(p_X \| q_{S,\ X}^P)$。

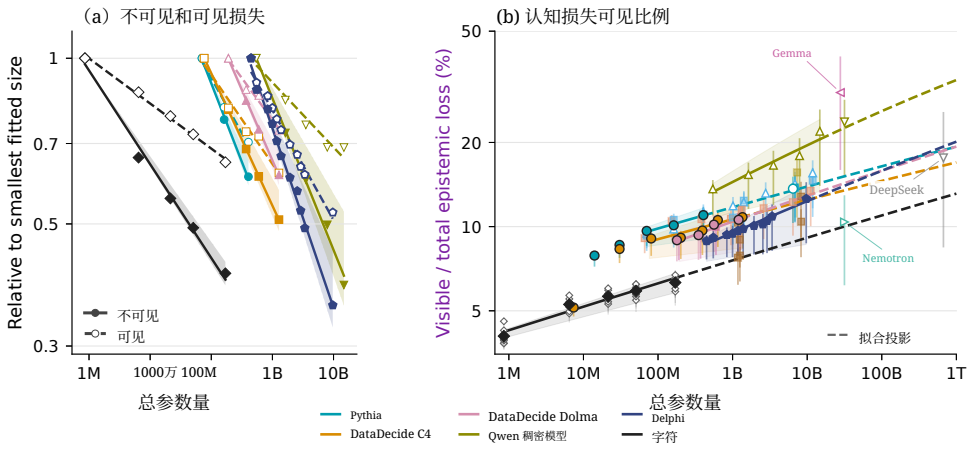


图3 在更大的规模下，更大比例的认知不确定性是可见的。(a) 实线和实心标记表示不可见损失；虚线和空心标记表示可见损失。颜色用于标识各族。Pythia 的离散度和不可见损失指数（γ、α）为 (0.200, 0.285)，DataDecide C4 为 (0.162, 0.240)，DataDecide Dolma 为 (0.152, 0.256)，稠密 Qwen 为 (0.118, 0.276)，Delphi 为 (0.200, 0.326)，字符模型为 (0.082, 0.170)。每个组成部分均以其族中最小的拟合规模进行归一化。(b) 在近期文本上的 72 个公开比较和 35 个字符设置中的可见比例。不确定性带反映了下限范围。

3 经验观测及其含义

3.1 实验中使用的集成

我们使用现有的开放权重模型和一个新的字符级族，并根据集成成员之间的关联紧密程度对其进行分组。模型的详细信息和明确列表见附录 B.1 和表 9。

独立随机种子。 这些模型之间的差异仅在于随机种子；我们的实验具体使用了 Pythia、DataDecide 和 Delphi 族，其中 DataDecide 包含两个因数据集选项不同而有所区别的族，即 C4 和 Dolma。我们的字符级变换器也使用独立随机种子。对于这一族，我们改变模型规模、训练预算和共享训练历史；其训练详情请参见附录 B.7。

发布重跑和小幅配方变更。 这些模型固定架构，但允许对训练配方或语料库进行小幅变更，也允许跨发布版本进行重复运行。

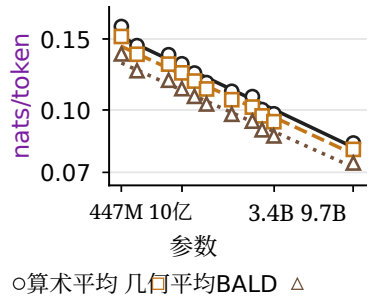


图 4 Delphi 种子集成中的集成分歧度量以相似的方式缩放。

更大规模的训练或架构变更。这些模型配对的模型大小相近且词元词汇表对齐，但允许在训练数据、训练流程或架构方面进行更大幅度的变更。其中一些模型还共享早期训练阶段。我们广泛使用的这一组集成包括大小相近的 Qwen 2.5 和 Qwen 3 模型二元组。该组还包括稠密模型/MoE 对比。

3.2 可见认知不确定性随模型、数据和上下文的扩展

在界定不可约损失时，我们使用近期文本（2026年8月 Common Crawl）进行 72 次公开模型比较，以减少预训练重叠。附录I给出了 C4 验证集（2019年 Common Crawl）上的相应结果。

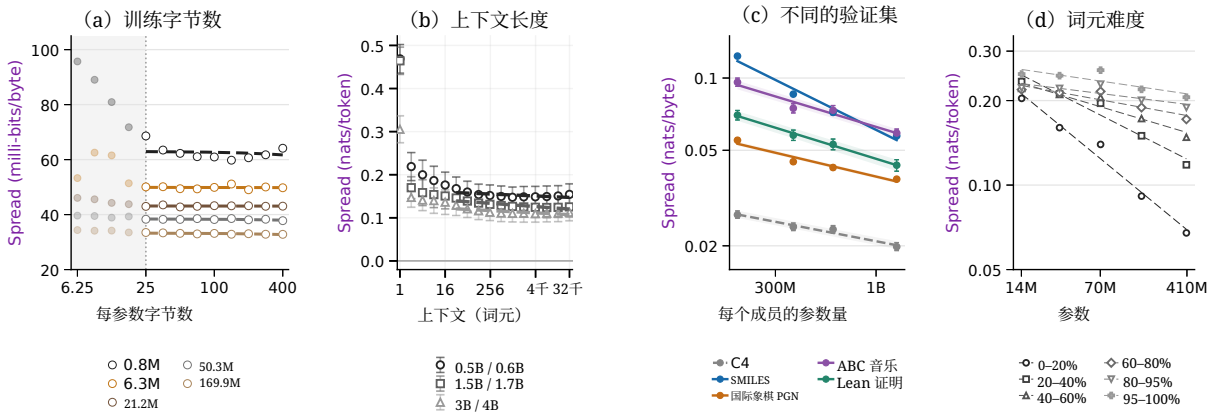


图 5 训练预算、上下文、评估领域和难度下的离散度。 （a）随着数据集大小增加，从每参数 25 字节开始，字符级模型的离散度变化很小。（b）对于三个 Qwen2.5/Qwen3 二元组，当上下文长度超过 16 个词元后，其离散度变化相对较小。（c）DataDecide-Dolma 的四种规模在 C4（灰色）和四个专门领域／分布外领域上的校正几何离散度；后者的离散度更高，但仍保持幂律关系。（d）Pythia 中按词元难度（熵）分箱后的离散度。对于难度更高的词元，离散度的收敛速度更慢。

随模型规模变化。 在近期文本上，对 $\hat{\rho}^{\mathrm{G}}_{\infty} \propto N^{-\gamma}$ 进行拟合可得 $\gamma = 0.118$–0.200，覆盖拟合得到的公开模型家族（图 2b）。在共享下限处，公开模型的损失指数为 $\eta = 0.191$–0.313（面板 a）。附录 B.5 和表 11、表 16 给出了统一的拟合方案和系列系数。我们的字符级模型在留出的 FineWeb-Edu 文本上也显示出同样平缓的规模趋势；按总参数量计算时，指数为 $\gamma = 0.118$（附录 B.7.2；图12）。图3a显示，在近期文本上，不可见分量在各模型家族中的下降幅度比离散度更为明显。几何离散度、算术离散度和 BALD 也表现出类似的平缓规模趋势（图 4；附录 C.2）。

随数据扩展。 由于对于多随机种子的开放权重模型系列，我们没有冷却阶段之后的可靠检查点，并且训练期间学习率会因（例如余弦）调度而变化，因此无法在这些模型家族中可靠地测量离散度随词元数量的变化。然而，对于我们的字符级系列，我们可以做到这一点，并且看到，在某个初始阶段之后，直到每参数 25 字节，离散度的变化都很小（图5a）。详细信息见附录 B.7。

随上下文长度扩展。 上下文长度是语言模型损失可以近似按照幂律下降的另一个维度（Kaplan 等人，2020年；Xiong 等人，2024年）。由于我们的字符级模型上下文长度相对较短，因此我们考察 Pythia、DataDecide、Delphi 和 Qwen 模型家族。图 5b 展示了三个 Qwen2.5/Qwen3 基础模型对在固定目标上的几何离散度，随着上下文增长至 32,768 个词元时的变化。在初始阶段显著收缩后，离散度对额外上下文的依赖很弱；详情以及 BALD 和算术离散度的类似结果，另见附录 D；Pythia、DataDecide 和 Delphi 集成也呈现出类似趋势。

随数据集变化。 图 5c 表明，将评估领域转向包括 ABC 音乐、化学 SMILES、国际象棋对局和 Lean 证明在内的专业领域数据集，可能会改变离散度定律的偏移量和斜率。在这些领域中，集成离散度仍然会随模型规模近似遵循幂律。附录 B.5 描述了评估集，并在图 11 中给出了其他模型家族的结果。

基于词元难度的集成离散度变化。直观而言，集成离散度量化了集成内部的分歧，因此应取决于下一个词元的难度。我们根据某个模型对词元进行难度分箱（预测熵），以评估这一假设。我们使用固定的参考模型，从而确保在每个模型规模下，每个难度分箱都包含相同的词元。

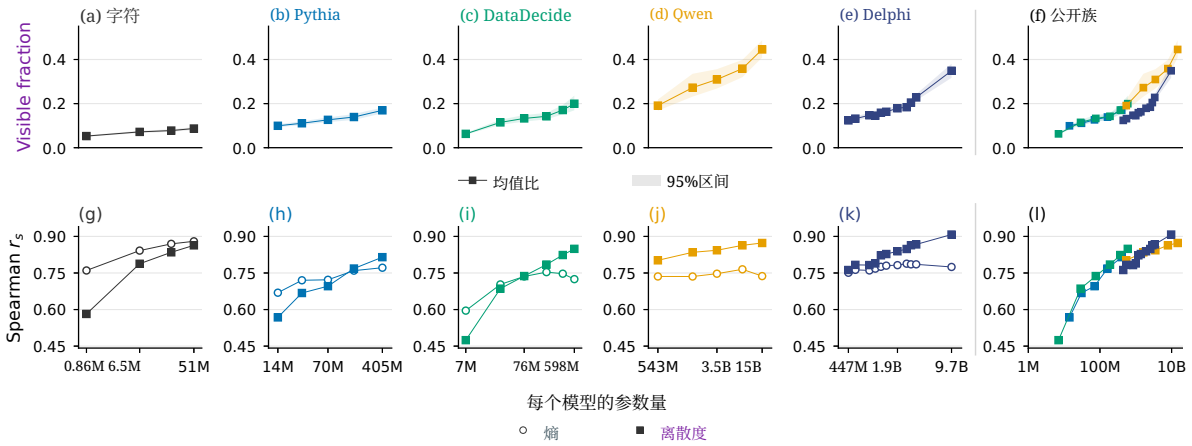


图6 所有族中的上下文依赖的不确定性。 (a–e) 使用基于参考分布的估计来计算每个上下文的认知不确定性所得的图3(b)对应图。平均校正后的离散度 除以平均参考模型与成员模型之间的KL散度。带状区域表示对留出窗口进行重采样得到的 95% 区间。(g–k) 单模型熵和集成离散度 与参考模型与成员模型之间的KL散度估计的可约损失之间的斯皮尔曼相关系数。(f,l) 在相同的40段近期文本上，四个公开族的可见比例和离散度相关性。字符模型使用留出的 FineWeb-Edu 文本。在每个族内，所有规模使用相同的测试上下文、词汇表和参考分布。

图5d展示了Pythia种子集成的趋势：我们观察到幂律趋势，其中衰减指数取决于难度分箱。困难位置在集成离散度中所占的比例不断增加，而它们随着模型规模增长而缓慢得到解决，这一直观上可以解释集成离散度的温和缩放；附录C详述了相关内容。

3.3 规模越大，可见的认知不确定性占比越高

给定可见、不可见和总认知损失这三个项，以及两个幂律，一个尚未解决的问题是：究竟哪些项确实具有幂律缩放。总认知损失的直接幂律关系，$\bar{L}_S - E \simeq KN^{-\eta}$，结合离散度 $\rho_\infty^P \simeq c_P N^{-\gamma_P}$，预测出的可见比例是二者之比$f_\infty^P \simeq (c_P/K)\, N^{\eta-\gamma_P}$。根据我们的实证结果（$\eta > \gamma_P$），由于前者的斜率更陡，这种外推最终会超过100%。另一种方法是，分别为公式（5）中的可见 和 不可见 项设定幂律关系，得到

$$\rho_\infty^P(N) = c_P N^{-\gamma_P}, \qquad L(q_\infty^P) - E = AN^{-\alpha}, \qquad f_\infty^P(N) = \frac{\rho_\infty^P(N)}{L(q_\infty^P) - E + \rho_\infty^P(N)}. \tag{8}$$

在这里，该比例保持在100%以下，但其总和通常并不服从幂律。问题变成了：这两种替代方案中，哪一种能更好地解释数据？在我们目前观察到的尺度范围内，对多个集成系列进行比较后发现，这两种方法的拟合质量相近，没有哪一种始终占优（附录E.3，尤其是图17）。要确定正确的模型，需要更大的集成系列；在这种情况下，两种模型对可见比例的预测会明显分化，因为第二个模型会达到饱和。图3b展示了分项拟合及其对不可约损失的敏感性。在共同的中心下限处，所有六个拟合序列中的不可见分量都比离散度衰减得更快，因此估计的可见比例会随着大小增加而上升（表17）。这些结果表明，在更大的尺度下，集成涵盖了总认知不确定性的更大比例，而集成在估计不确定性方面也变得更加有效。然而，取决于下限估计值，可见比例上升的斜率可能很小。在不同设置下，（Rosso等，2025；Kirsch，2025）也得到了类似的定性结果：他们观察到集成分歧会随着尺度增加而减小；但是，他们没有考虑模型容量及其更陡峭的缩放

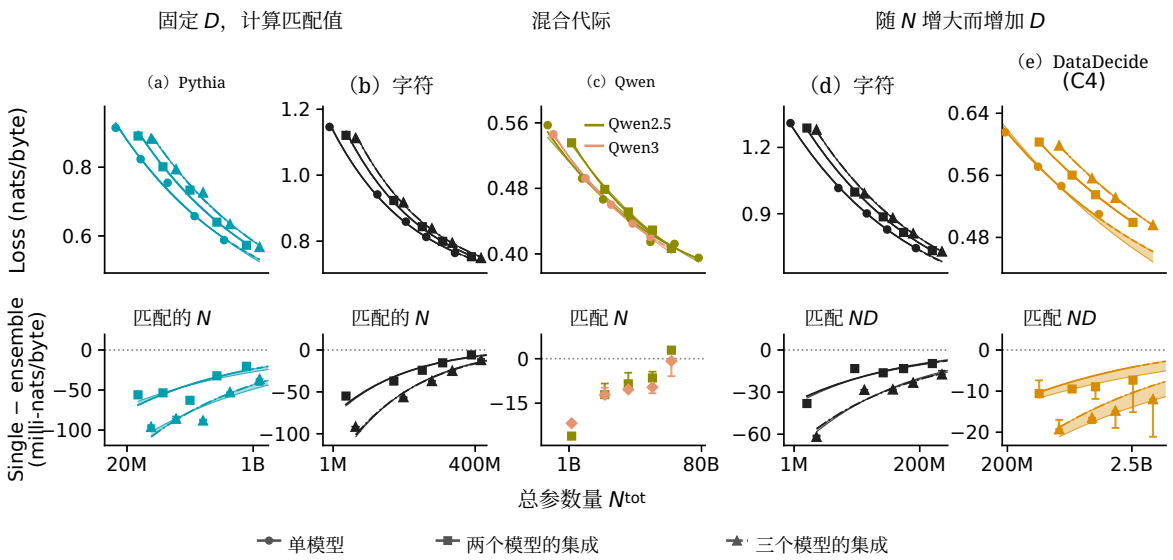


图7 五种设置下的单模型、二元组和三元组。 上：单模型和集成的损失与总参数量的关系。下：拟合的单模型损失减去实测集成损失（标记），以及根据分别拟合的损失定律和扩散定律得出的预测（曲线）。阴影和垂直区间表示对语料库特定下限范围的敏感性，并且在每个下限处重新拟合损失定律。(a、b) Pythia 和字符级模型与训练数据及总参数量相匹配。每个字符级模型使用 8.493B 字节。(c) Qwen 与参数量相匹配，但混合了训练预算不同的代际。在 (d、e) 中，每个间隔都将集成与总训练计算量成本 ND相同的单模型进行比较。字符模型使用 FineWeb-Edu；公开模型使用近期文本。由于 Delphi 的已发布检查点既不满足固定数据、匹配参数量的约束，也不满足匹配训练计算量的约束，因此将其排除。对于分布外数据，我们观察到交叉现象；参见附录图19

定律在这一上下文中的作用。通过这样做，我们表明，尽管集成的离散度在缩小，但利用集成恢复出的认知不确定性比例仍会随着尺度增加而上升。附录E介绍了详细信息、局限性及其他结果。

跨上下文的可见比例。我们使用固定参考分布的前向KL散度（公式（1））来估计各个上下文中的可约损失。参考分布分别为：字符模型使用BLT，Pythia使用MPT-30B，DataDecide使用OLMo-7B，Qwen使用Qwen2.5-72B，Delphi使用Delphi 25B。

估计的可见比例在五个模型系列中总体上都随规模增大而上升（图6a–f）。其幅度取决于所选的参考分布。相比之下，图3b衡量的是高于共同下限的损失。

图6g–k比较了熵和离散度作为评分指标，用于根据参考分布估计的可约损失对上下文进行排序。附录F.4给出了通用估计器。在较大规模的公开模型中，离散度在预测基于参考分布的认知不确定性方面也比熵或采用Adam训练的隐藏状态线性探针更有效，后者遵循Ahdritz等（2024）的方法；参见图18和附录F.3.1。

3.4 集成：更大模型的替代方案

可约损失中越来越大的比例由集成离散度解释，这一事实表明，在更大模型规模下，集成可以提供一个新的扩展维度。在这里，我们在匹配的总参数量或训练计算量下比较不同的集成规模，并将单模型视为规模为一的集成。对于计算最优训练，在 Chinchilla损失定律中假设指数 $\alpha = \beta$ 大致相等$L(N,\ D)$，并在固定训练预算下使损失最小化，可得 $D = \kappa N$，其中$\kappa = (b/a)^{1/\alpha}$（Hoffmann 等，2022年）。由于训练计算量同时与模型规模和数据规模成正比

ND，因此在计算最优训练中，它随 N^2 增长。由此，我们可以通过两种方式，将一个由 S 个模型组成的集成（每个模型具有 N 个参数，并使用 D 个词元进行训练）与一个更大的单模型进行比较。**（1）匹配总参数量和计算量**。在此设置下，我们使用 SN 参数，在相同的 D 词元上训练一个单模型。对于稠密模型而言，这在主导阶上也匹配了集成的训练计算量和推理计算量。

（2）在计算最优性下匹配计算量。在此设置下，假设集成成员和更大模型均为计算最优，并且具有相同的总训练计算量。这意味着更大模型具有 $\sqrt{S}N$ 参数，并使用 $\sqrt{S}D$ 词元进行训练。该模型使用的数据多于每个成员，但参数少于总参数量。

要了解何时集成更为可取，请回顾其成员平均损失的降低量恰好等于离散度（公式（5））。因此，当这一降低量超过更大模型所带来的改进时，集成的表现就优于更大的单模型：

$$\rho_S^P > \bar{L}_S - L(q_{\text{big}}). \tag{9}$$

我们用 $L(N、\ D)$表示成员平均损失，用 $\rho_S^P = (1 - S^{-1})c_P N^{-\gamma_P}$ 表示离散度，并忽略其在初始阶段之后对训练数据的微弱依赖（图5a）。以下论断依据这两条缩放定律，将比较扩展到任意两个固定的集成规模，其中 $S = 1$ 对应于单模型。

论断（在超过某个规模后，在预算相同的情况下，成员更多的集成表现更优）。假设两个规模为 $1 \leq S_1 < S_2$ 的集成遵循相同的 Chinchilla 定律和扩散定律，其中$a,\ b,\ c_P > 0$ 以及 $\alpha = \beta > \gamma_P > 0$。定义 $r = S_2/S_1$ 和 $u_i = 1 - S_i^{-1}$。令由 S_2 个成员组成的集成中的每个成员均为计算最优，并具有 N 个参数和 D 个训练词元。将其与由 S_1 个成员组成的集成进行比较；在情景（1）中，该集成的成员使用（$rN,\ D$），以匹配训练数据和总参数量；在情景（2）中，成员使用（$\sqrt{r}\,N,\ \sqrt{r}\,D$），以在成员为计算最优的情况下匹配总训练计算量。在情景 1 中，当 $N > N_1^\star(S_1,\ S_2)$ 时，S_2 个成员的集成具有更低的损失；在情景 2 中，当 $N > N_2^\star(S_1,\ S_2)$ 时具有更低的损失，其中

$$N_1^\star(S_1, S_2) = \left[\frac{a(1 - r^{-\alpha})}{c_P(u_2 - u_1 r^{-\gamma_P})}\right]^{1/(\alpha-\gamma_P)}, N_2^\star(S_1, S_2) = \left[\frac{2a(1 - r^{-\alpha/2})}{c_P(u_2 - u_1 r^{-\gamma_P/2})}\right]^{1/(\alpha-\gamma_P)}. \tag{10}$$

两个阈值都是有限值，并且 $N_2^\star > N_1^\star$。设置 $S_1 = 1$ 可恢复单模型比较；设置$S_1 = 2,\ S_2 = 3$ 可得到三元组–二元组交叉点。证明见附录G.1。

因此，对于任意两个固定的集成规模，缩放定律都会预测一个规模阈值：超过该阈值后，只要这些定律在该规模下仍然成立，将相同预算分配给更多集成成员就能获得更低的损失。图7支持模型规模增大时竞争力随之增强这一预测。我们发现，集成与更大单模型之间的损失差距总体上会缩小，无论是在第一个情景 (a,b) 中固定训练数据，还是在第二个情景 (d,e) 中匹配总训练成本 ND。尽管存在这一趋势，大多数集成的损失仍高于更大模型。底行曲线由分别拟合的两条定律得出：集成离散度幂律带来的增益，以及基于 Chinchilla 的更大单模型所带来的改进。我们对二元组和三元组拟合出一条共享的离散度定律，然后减去更大模型预测的改进量，并在固定 D 或总训练成本 ND 下匹配参数。Qwen 未显示下方曲线，因为其成员来自不同代际。附录G.4展示了专用数据上的类似结果；对于某些模型而言，这些数据并非分布内数据（领域包括音乐、数学证明、化学 SMILES 和国际象棋），在这些数据上，离散度可能高出数倍，并且在不同模型家族中都能观察到交叉点

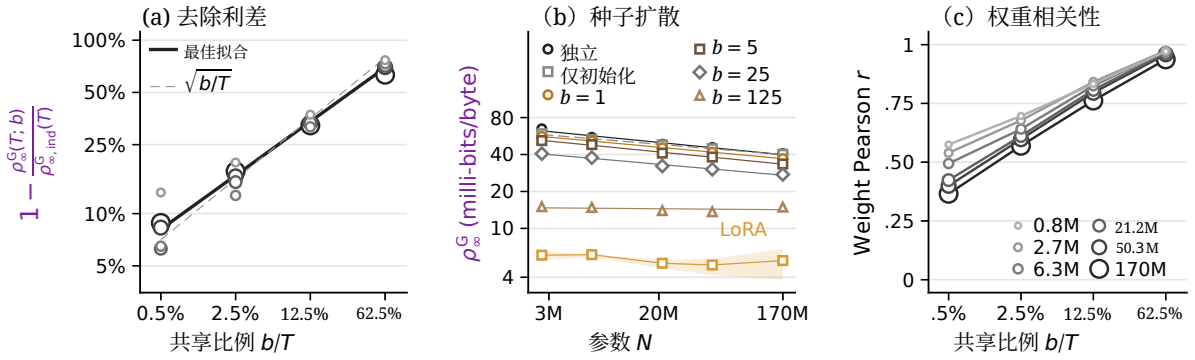


图8 共享训练会降低集成离散度。 （a）独立随机种子间的分散度损失相对于训练共享比例的变化，b/T，其中 b 是分支点，T 是最终训练预算。离散度损失近似遵循平方根定律。（b）独立随机种子、后期分支和 LoRA 集成的几何离散度与模型规模的关系。图（a、b）使用近期文本。（c）在 $T = 200$每个解析参数对应的字节数下，父模型与分支模型完整最终参数向量之间的皮尔逊相关系数。大小图例适用于（a、c）；（c）包含全部六种宽度。附录 H.3 给出了计算细节以及相应的离散度比较。

。附录 G 介绍了详细信息、交叉点的数值范围，以及更多关于在训练检查点中观察到交叉点的结果。

这些发现与 Kim 等人（2026）的研究相互补充；他们表明，在广泛的数据重用和经过调优的正则化条件下，集成的表现可以超过更大型的单个语言模型；另请参见（Lobacheva 等人，2020）。我们的分析根据损失与集成离散度的相对缩放关系，推导出了一个交叉点判据。该判据基于我们的两条缩放定律，但不要求特定的正则化方案；从理念上说，其关于集成的含义也与 Abe 等人（2022）的观点相悖，后者认为更大型的单模型可以达到与集成相当的性能。

3.5 采用后期分支的集成离散度与 LoRA 集成

后期分支通过在较晚的检查点分离集成成员的训练来降低集成成本。因此，集成成员将具有相同的初始化和早期更新。我们首先研究采用全参数延续训练的这类集成，然后研究 LoRA 集成，即在权重的低秩适配器上继续训练。

离散度遵循近似的平方根定律。使用 FineWeb-Edu 字节模型，我们在 $b \in \{1, 5, 25, 125\}$ 每参数字节数处分支，并在最终预算 $T = 200$处进行评估。来自同一父模型的不同延续从父模型的参数开始，但以不同顺序看到相同的剩余数据；训练期间会更新所有参数。本节中的独立参照同时改变初始化和数据顺序。[4] 对于不同的分支点，我们会测量相对于完全独立运行的参考分布，离散度保留了多少——即离散度的保留份额为 $R(N,\ t,\ b) = \rho^{\mathrm{G}}_{\infty,\ \mathrm{branch}}(N,\ t,\ b)\,/\,\rho^{\mathrm{G}}_{\infty,\ \mathrm{ind}}(N,\ t)$.

其中，$\rho_{\infty,\ \mathrm{branch}}$ 和 $\rho_{\infty,\ \mathrm{ind}}$ 分别表示后期分支和参考集成的离散度。图8所示的拟合规律表明，损失掉（而非保留）的离散度占比呈现明显的幂律缩放：

$$1 - R(N, T, b) \simeq 0.857 \left(\frac{b}{T}\right)^{0.445}. \tag{11}$$

[4]我们看到，与仅在初始化方面存在差异的集成相比，改变数据顺序会使离散度略有增加；在五种拟合规模中，平均仅增加约 4%。

例如，预计在训练进行到一半时进行分支，会损失约 63% 的独立随机种子离散度；若外推至仅剩 10% 训练时间时才进行分支，则损失约为 82%。这些结果表明，后期分支会大幅降低集成多样性，凸显了共享训练历史以高效构建集成所需付出的代价（Wasay 等人，2020年）。这一幂律也可用于通过后期分支，以较低成本预测真实随机种子的平均离散度，从而在决定开展大规模多随机种子实验之前进行估算。图 8c 显示，若能访问模型，可通过权重相关性来估算这种多样性；随着分支时间推迟，在不同模型规模下，权重相关性始终更高（采用对数尺度）；另见 Fort 等人（2019）。关于离散度与后期分支之间联系的这些观测结果，表明贝叶斯深度学习方法可能存在局限：这类方法通过训练轨迹的最后阶段近似后验分布（Maddox等，2019年），或更普遍地依赖于对损失“盆地”的建模。详情及更多结果见附录 H。

多样性惩罚项。 人们可以尝试使用多样性惩罚项来增加离散度，例如负相关学习（Liu 和 Yao，1999年）及其几何池化扩展（Webb 等人，2021年）所采用的方法。在附录 H.4 中，我们报告了这样一项实验：结果显示，多样性惩罚项确实在每种模型规模下都增加了离散度，但代价是更高的成员损失，这在很大程度上抵消了总集成损失的额外增益。这意味着，尽管集成现在捕获了更多不确定性，但这种不确定性的增加主要归因于单模型性能下降，因此该惩罚项失去了意义。

LoRA 集成。 LoRA 集成（Wang等，2023年）和贝叶斯适配器（Yang 等人，2024年；Lin 等人，2026年；Rahmati 等人，2025年）同样共享一个预训练骨干网络，因此我们根据后期检查点上独立训练的适配器来测量离散度。在这里，我们比较了后期分支中的 LoRA 与独立随机种子的离散度，发现，在 $b = 125$处分支的 LoRA 二元组相较于来自同一检查点的完整参数分支，进一步降低了 2 到 4 倍（图 8b ）。更高的秩能够改善单模型损失，但即使秩高达 $d/2$，保留的完整分支离散度仍不到一半（附录 H.5；图 22）。Mühlematter 等人（2026年）报告称，LoRA 集成在分类任务中具有良好的多样性和校准效果。然而，他们的共享骨干网络微调基线不同于我们的独立预训练参照。

4 讨论与局限性

在这里，我们讨论研究所得的理论和实证结果。可以说，其中许多片段在过去的文献中已经有所体现，但据我们所知，整体图景及其蕴含意义一直付之阙如。在理论方面，我们认为，可约损失是经过校准的大语言模型中认知不确定性的正确平均度量——这一校准条件十分重要，因为在公式 3 中，互信息项（具有明确的认知不确定性解释）与可约损失之间的等式依赖于该条件。此外，我们发现，相较于 BALD，离散度，尤其是 几何离散度，是衡量集成分歧的合适指标，因为它与认知损失存在直接联系。

在这一背景下，两个指数范围不同的明确缩放定律的存在具有重要意义。随着模型规模扩大，对于给定上下文而言，越来越大比例的真实认知不确定性会通过集成多样性变得可见，而集成即使仅为改善预测也会变得具有竞争力，而不取决于其是否用于不确定性估计。最后，第三个幂律表明，通过后期分支和 LoRA 进行低成本近似可以显著降低离散度。我们的发现促使我们训练更大规模的随机种子集成，以验证离散度的幂律缩放；同时开展更广泛的集成实验，以在更多样化的设置中检验后期分支幂律。

我们的理论主体以及实验中发现的趋势，不仅适用于种子集成，也适用于任意集成。对能力各异的模型进行集成时，变化之处在于，成员的平均损失将不再反映单个成员的表现，

而集成离散度相较于平均值带来的改进，也未必意味着优于表现最佳的集成成员。无论是将认知不确定性估计为可约损失，还是将集成与单模型缩放进行比较，同样的论证都适用。

我们的结果和预测最重要的局限在于对不可约损失的不确定性。不过请注意，这一下限估计仅影响容量定律（即 Chinchilla 缩放定律），而与我们所确定的另外两种缩放定律——扩散定律和后期分支幂律——无关。只要模型规模的 Chinchilla 缩放指数 η 高于我们估计的指数 $\gamma \approx 0.08$–0.20（扩散定律的指数），我们预测的趋势就仍然有效。附录中将进一步阐述这一局限以及其他局限。

未来一个有趣的理论研究方向，是探究与集成相关的两种缩放定律的来源：1）尽管我们将扩散的幂律缩放的缓慢斜率与困难词元的缓慢解析联系起来，但幂律产生的确切机制仍有待研究；2）同样，图8所示后期分支幂律的产生机制也尚未得到解释。

参考文献

Taiga Abe、Estefany Kelly Buchanan、Geoff Pleiss、Richard Zemel 和 John Cunningham。《深度集成有效，但有必要吗？》载于《神经信息处理系统进展》，第35卷，第33646–33660页。Curran Associates, Inc.，2022年。doi：10.52202/ 068431-2438。URL：https://proceedings.neurips.cc/paper_files/paper/2022/hash/da18c47118a2d09926346f33bebde9f4-Abstract-Conference.html。Gustaf Ahdritz、Tian Qin、Nikhil Vyas、Boaz Barak 和 Benjamin L. Edelman。《使用语言模型区分可知与不可知》，2024年。URL：https://proceedings.mlr.press/v235/ ahdritz24a.html。ICML。Dario Amodei、Chris Olah、Jacob Steinhardt、Paul Christiano、John Schulman 和 Dan Mané。《人工智能安全中的具体问题》，2016年。URL：https://arxiv.org/abs/1606.06565。arXiv：1606.06565。Yasaman Bahri、Ethan Dyer、Jared Kaplan、Jaehoon Lee 和 Utkarsh Sharma。《解释神经网络缩放定律》，2024年。URL：https://arxiv.org/abs/2102.06701。美国国家科学院院刊；预印本首次发布于2021年。Stella Biderman 等人。《Pythia：用于分析大语言模型训练与缩放的一套工具》，2023年。URL：https://arxiv.org/abs/2304.01373。ICML。Jochen Bröcker。《适当评分的可靠性、充分性与分解》。《英国皇家气象学会季刊》，135(643)：1512–1519，2009年。doi：10.1002/qj.456。URL：https://arxiv.org/abs/ 0806.0813。Erik Daxberger、Agustinus Kristiadi、Alexander Immer、Runa Eschenhagen、Matthias Bauer 和 Philipp Hennig。《拉普拉斯重制版——轻松实现贝叶斯深度学习》，2021年。URL：https://arxiv.org/abs/2106.14806。NeurIPS。DeepSeek-AI。《DeepSeek-V3 技术报告》，2024年。URL：https://arxiv.org/abs/2412.19437。arXiv：2412.19437。Shrey Desai 和 Greg Durrett。《预训练 Transformer 的校准》。载于《*2020*年自然语言处理实证方法会议论文集（*EMNLP*）》，第295–302页，2020年。doi：10.18653/v1/2020. emnlp-main.21。URL：https://aclanthology.org/2020.emnlp-main.21/。EvaByte。《EvaByte 模型卡》，2025年。URL：https://huggingface.co/EvaByte/EvaByte。基础模型发布。Stanislav Fort、Huiyi Hu 和 Balaji Lakshminarayanan。《深度集成：损失景观视角》，2019年。URL：https://arxiv.org/abs/1912.02757。arXiv 预印本。Eric Foxley。《诺丁汉音乐数据库》，2011年。URL：https://ifdo.ca/~seymour/nottingham/ nottingham.html。由 Seymour Shlien 编辑的 ABC 版本；最后更新于2011年10月2日。

Yarin Gal 和 Zoubin Ghahramani。《Dropout 作为贝叶斯近似：表示深度学习中的模型不确定性》，2016年。URL：https://proceedings.mlr.press/v48/gal16.html。ICML。Leo Gao 等人。《The Pile：用于语言建模的800GB多样化文本数据集》，2020年。URL：https://arxiv. org/abs/2101.00027。arXiv：2101.00027。Gemma Team。《Gemma 4 技术报告》，2026年。URL：https://arxiv.org/abs/2607.02770。arXiv：2607.02770。Alex Graves。《神经网络的实用变分推断》。载于《神经信息处理系统进展》，第24卷，第2348–2356页。Curran Associates, Inc.，2011年。URL：https://proceedings.neurips.cc/paper/2011/hash/7eb3c8be3d411e8ebfab08eba5f49632-Abstract.html。Dirk Groeneveld 等人。《OLMo：加速语言模型科学研究》，2024年。URL：https://arxiv.org/ abs/2402.00838。arXiv：2402.00838。Dylan Hadfield-Menell、Anca Dragan、Pieter Abbeel 和 Stuart Russell。《关机游戏》。载于《第二十六届国际人工智能联合会议论文集》，第220–227页，2017年。doi：10.24963/ijcai.2017/32。URL：https://www.ijcai.org/proceedings/2017/32。Will Held。《可外推至 300× 拟合范围之外的缩放定律》。Marin，2026年。URL：https://marin.community/ blog/2026/05/11/delphi/。Jordan Hoffmann 等人。《训练计算最优的大语言模型》，2022年。URL：https://arxiv.org/abs/2203.15556。NeurIPS。Neil Houlsby、Ferenc Huszár、Zoubin Ghahramani 和 Máté Lengyel。《用于分类和偏好学习的贝叶斯主动学习》，2011年。URL：https://arxiv.org/abs/1112.5745。arXiv：1112.5745。Saurav Kadavath 等人。《语言模型（大体上）知道自己知道什么》，2022年。URL：https://arxiv.org/ abs/2207.05221。arXiv 预印本。Jared Kaplan 等人。《神经语言模型的缩放定律》，2020年。URL：https://arxiv.org/abs/2001. 08361。arXiv 论文。Konwoo Kim、Suhas Kotha、Percy Liang 和 Tatsunori Hashimoto。《无限计算下的预训练》，2026年。URL：https://proceedings.iclr.cc/paper_files/paper/2026/hash/79682ef54677a2cb3dfb8383b372037f-Abstract-Conference.html。ICLR；预印本首次发布于2025年9月18日。Andreas Kirsch。《（隐式）集成的集成：大型模型中的认知不确定性坍缩》，2025年。URL：https://arxiv.org/abs/2409.02628v2。TMLR；预印本首次发布于2024年。B. J. K. Kleijn 和 A. W. van der Vaart。《模型错设下的 Bernstein–von Mises定理》。《*Electronic Journalof Statistics*》，6：354–381，2012年。doi：10.1214/12-EJS675。URL：https://doi.org/10.1214/12-EJS675。Nikita Kotelevskii、Vladimir Kondratyev、Martin Takáˇc、Eric Moulines 和 Maxim Panov。《从风险到不确定性：通过贝叶斯估计生成预测不确定性度量》。载于《国际学习表征会议》，2025年。URL：https://proceedings.iclr.cc/paper_files/paper/2025/hash/e6f32e64b9c27d153b46c94f0fe22b56-Abstract-Conference.html。A. Gilad Kusne、Heshan Yu、Changming Wu、Huairuo Zhang、Jason Hattrick-Simpers、Brian DeCost、Suchismita Sarker、Corey Oses、Cormac Toher、Stefano Curtarolo、Albert V. Davydov、Ritesh Agarwal、Leonid A. Bendersky、Mo Li、Apurva Mehta 和 Ichiro Takeuchi。《通过贝叶斯主动学习即时开展闭环材料发现》。《*NatureCommunications*》，11：5966，2020年。doi：10.1038/s41467-020-19597-w。URL：https://www.nature. com/articles/s41467-020-19597-w。Salem Lahlou、Moksh Jain、Hadi Nekoei、Victor Ion Butoi、Paul Bertin、Jarrid Rector-Brooks、Maksym Korablyov 和 Yoshua Bengio。《DEUP：直接认知不确定性预测》。《*Transactions onMachine LearningResearch*》，2023年。URL：https://openreview.net/forum?id=eGLdVRvvfQ。Balaji Lakshminarayanan、Alexander Pritzel 和 Charles Blundell。《使用深度集成进行简单且可扩展的预测不确定性估计》，2017年。URL：https://arxiv.org/abs/1612.01474。NeurIPS。Jeffrey Li、Alex Fang、Georgios Smyrnis、Maor Ivgi、Matt Jordan 等人。《DataComp-LM：寻找新一代语言模型训练集》，2024年。URL：https://arxiv.org/abs/2406.11794。arXiv：2406.11794。

Lichess。Lichess 开放数据库：2013年1月标准等级分对局，2013年。URL https://database.lichess.org/。PGN 数据集。Moule Lin、Shuhao Guan、Andrea Patane、David Gregg 和 Goetz Botterweck。Bayesian-LoRA：大语言模型的概率低秩适配。载于 国际机器学习会议，2026年。URL https://arxiv.org/abs/2601.21003。Yong Liu 和 Xin Yao。通过负相关进行集成学习。神经网络，12(10)：1399–1404，1999年。doi：10.1016/S0893-6080(99)00073-8。URL https://www.sciencedirect.com/science/article/pii/ S0893608099000738。Ekaterina Lobacheva、Nadezhda Chirkova、Maxim Kodryan 和 Dmitry Vetrov。深度集成中的幂律，2020年。URL https://arxiv.org/abs/2007.08483。NeurIPS。Wesley J. Maddox、Pavel Izmailov、Timur Garipov、Dmitry Vetrov 和 Andrew Gordon Wilson。深度学习中贝叶斯不确定性的简单基线，2019年。URL https://proceedings.neurips.cc/paper_files/paper/2019/hash/118921efba23fc329e6560b27861f0c2-Abstract.html。NeurIPS。Ian Magnusson、Nguyen Tai、Ben Bogin、David Heineman、Jena D. Hwang、Luca Soldaini、Akshita Bhagia、Jiacheng Liu、Dirk Groeneveld、Oyvind Tafjord、Noah A. Smith、Pang Wei Koh 和 Jesse Dodge。DataDecide：如何通过小规模实验预测最佳预训练数据，2025年。URL https://arxiv.org/abs/2504.11393。arXiv 论文。Marin Community。复现额外的 Delphi 随机种子。Marin 文档，2026年。URL https://marin.readthedocs.io/en/latest/reproducibility/delphi-seed-repeats/。Moonshot AI。Kimi-K2-Base。Hugging Face 模型发布，2025年。URL https://huggingface.co/ moonshotai/Kimi-K2-Base。检查点修订版 ce72df012259dcc55d945e890f815fe7ef69159c。Dominik J. Mühlematter、Michelle Halbheer、Alexander Becker、Dominik Narnhofer、Helge Aasen、Konrad Schindler 和 Mehmet Ozgur Turkoglu。LoRA 集成：用于自注意力网络的高效不确定性建模。机器学习研究汇刊，2026年。URL https://openreview.net/forum?id=yhXXmOMpSQ。Preetum Nakkiran、Arwen Bradley、Adam Golinski、Eugene Ndiaye、Michael Kirchhof 和 Sinead Williamson。在词元上训练，在概念上校准：大语言模型中语义校准的涌现。载于 国际学习表征会议，2026年。URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/39b228b1a83232b14a1c81b0ed5a7a05-Abstract-Conference.html。Frank Nielsen 和 Richard Nock。用于逼近 f-散度的卡方距离与高阶卡方距离。*IEEE* 信号处理快报，21(1)：10–13，2014年。doi：10.1109/LSP.2013.2288355。URL https://arxiv. org/abs/1309.3029。英伟达。NVIDIA Nemotron 3 Nano 30B-A3B Base BF16：模型卡。https://huggingface.co/nvidia/NVIDIA-Nemotron-3-Nano-30B-A3B-Base-BF16，2025年。英伟达。NVIDIA Nemotron 3.5 Lightning 30B-A3B Base BF16：模型卡。https://huggingface.co/nvidia/NVIDIA-Nemotron-3.5-Lightning-30B-A3B-Base-BF16，2026年。Gary W. Oehlert。关于德尔塔方法的注记。《美国统计学家》，46(1)：27–29，1992年。doi：10.1080/00031305. 1992.10475842。URL https://doi.org/10.1080/00031305.1992.10475842。OpenAI。GPT-4 技术报告。*arXiv* 预印本 *arXiv:2303.08774*，2023年。URL https://arxiv.org/abs/ 2303.08774。Artidoro Pagnoni、Ram Pasunuru、Pedro Rodriguez、John Nguyen、Benjamin Muller、Margaret Li、Chunting Zhou、Lili Yu、Jason Weston、Luke Zettlemoyer、Gargi Ghosh、Mike Lewis、Ari Holtzman 和 Srinivasan Iyer。字节潜在变换器：补丁比词元扩展得更好，2024年。URL https://arxiv.org/abs/2412.09871。arXiv:2412.09871。Guilherme Penedo 等人。FineWeb 数据集：在规模化条件下提炼网络，获取最优质的文本数据，2024年。URL https://arxiv.org/abs/2406.17557。arXiv:2406.17557。Qwen 团队。Qwen2.5 技术报告，2024年。URL https://arxiv.org/abs/2412.15115。arXiv:2412.15115。Jack W. Rae、Anna Potapenko、Siddhant M. Jayakumar 和 Timothy P. Lillicrap。用于长程序列建模的压缩型变换器，2019年。URL https://arxiv.org/abs/1911.05507。

Colin Raffel、Noam Shazeer、Adam Roberts、Katherine Lee、Sharan Narang、Michael Matena、Yanqi Zhou、Wei Li 和 Peter J. Liu。通过统一的文本到文本转换器探索迁移学习的极限。机器学习研究期刊，21(140)：1–67，2020年。URL https://www.jmlr.org/papers/v21/20-074.html。Amir Hossein Rahmati、Sanket Jantre、Weifeng Zhang、Yucheng Wang、Byung-Jun Yoon、Nathan Urban 和 Xiaoning Qian。C-LoRA：用于大语言模型不确定性估计的上下文低秩适配。载于 神经信息处理系统进展，第38卷，2025年。URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/615e48e9f5200d8f6d31ae08a348a497-Abstract-Conference.html。Tom Rainforth、Adam Foster、Desi R. Ivanova 和 Freddie Bickford Smith。现代贝叶斯实验设计。统计科学，39(1)：100–114，2024年。doi：10.1214/23-STS915。URL https://doi.org/10.1214/ 23-STS915。Mattia Rosso、Simone Rossi、Giulio Franzese、Markus Heinonen 和 Maurizio Filippone。深度学习不确定性的缩放定律，2025年。URL https://arxiv.org/abs/2506.09648v2。arXiv 预印本；v2 于2026年2月9日修订。Stuart Russell。《人类兼容：人工智能与控制问题》。Viking，2019年。URL https://www.penguinrandomhouse.com/books/566677/human-compatible-by-stuart-russell/。Luca Soldaini 等人。Dolma：用于语言模型预训练研究的三万亿词元开放语料库，2024年。URL https://arxiv.org/abs/2402.00159。arXiv:2402.00159。Claudio Spiess、David Gros、Kunal Suresh Pai、Michael Pradel、Md Rafiqul Islam Rabin、Amin Alipour、Susmit Jha、Prem Devanbu 和 Toufique Ahmed。代码语言模型的校准与正确性。载于 *2025 IEEE/ACM 第47届软件工程国际会议（ICSE）*，第540–552页，2025年。doi：10.1109/ICSE55347.2025. 00040。URL https://www.software-lab.org/publications/icse2025_calibration.pdf。Databricks AI研究团队。MPT-30B：提升开源基础模型的标准，2023年。URL https://www.databricks.com/blog/mpt-30b。模型发布，6月22日。mathlib社区。Lean 数学库。载于 第*9*届 *ACM SIGPLAN* 国际认证程序与证明会议论文集，CPP 2020。ACM，2020年。doi：10.1145/3372885.3373824。URL https://doi.org/10.1145/3372885.3373824。Oskar van der Wal、Pietro Lesci、Max Müller-Eberstein、Naomi Saphra、Hailey Schoelkopf、Willem Zuidema 和 Stella Biderman。PolyPythias：五十次语言模型预训练运行中的稳定性与离群值。载于 国际学习表征会议，2025年。URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/d611d06e3207330555fbc10810e70163-Abstract-Conference.html。Xi Wang、Laurence Aitchison 和 Maja Rudolph。用于大语言模型微调的 LoRA 集成，2023年。URL https://arxiv.org/abs/2310.00035。arXiv:2310.00035。Abdul Wasay、Brian Hentschel、Yuze Liao、Sanyuan Chen 和 Stratos Idreos。MotherNets：快速深度集成学习。载于机器学习与系统会议论文集，2020年。URL https://proceedings.mlsys. org/paper_files/paper/2020/file/4a420924d20bc025ebb37849169e6ebd-Paper.pdf。Sumio Watanabe。广泛适用的贝叶斯信息准则。机器学习研究期刊，14(27)：867–897，2013年。URL https://jmlr.org/papers/v14/watanabe13a.html。Andrew Webb、Charles Reynolds、Wenlin Chen、Henry Reeve、Dan Iliescu、Mikel Luján 和 Gavin Brown。是否进行集成：端到端训练何时失败？载于 数据库中的机器学习与知识发现（*ECML PKDD 2020*），第12459卷《计算机科学讲义》，第109–123页。Springer，2021年。doi：10.1007/978-3-030-67664-3_7。URL https://arxiv.org/abs/1902.04422。Danny Wood、Tingting Mu、Andrew M. Webb、Henry W. J. Reeve、Mikel Luján 和 Gavin Brown。集成学习中多样性的统一理论。机器学习研究期刊，24(359)：1–49，2023年。URL https://jmlr.org/papers/v24/23-0041.html。Zhenqin Wu、Bharath Ramsundar、Evan N. Feinberg、Joseph Gomes、Caleb Geniesse、Aneesh S. Pappu、Karl Leswing 和 Vijay Pande。MoleculeNet：分子机器学习基准。《化学科学》，9(2)：513–530，2018年。doi：10.1039/C7SC02664A。URL https://doi.org/10.1039/C7SC02664A。Wenhan Xiong、Jingyu Liu、Igor Molybog、Hejia Zhang、Prajjwal Bhargava、Rui Hou、Louis Martin、Rashi Rungta、Karthik Abinav Sankararaman、Barlas Oguz、Madian Khabsa、Han Fang、Yashar Mehdad、Sharan Narang、Kshitiz

Malik、Angela Fan、Shruti Bhosale、Sergey Edunov、Mike Lewis、Sinong Wang 和 Hao Ma。《基础模型的有效长上下文扩展》。载于*2024*年北美计算语言学协会分会年会论文集：人类语言技术（第*1*卷：长篇论文），第 4643–4663 页。计算语言学协会，2024年。doi：10.18653/v1/2024.naacl-long.260。URL：https://aclanthology.org/2024.naacl-long.260/。Adam X. Yang、Maxime Robeyns、Xi Wang 和 Laurence Aitchison。《面向大语言模型的贝叶斯低秩适配》，2024年。URL：https://arxiv.org/abs/2308.13111。ICLR。An Yang 等人。《Qwen3技术报告》，2025年。URL：https://arxiv.org/abs/2505.09388。arXiv：2505.09388。

附录内容

章节	涵盖内容	Page
A. 校准	字符模型、Pythia 的留出集再校准增益，DataDecide、Qwen 和 Delphi，以及置信度校准和这些测试的局限 .	19
B. 模型、评估与大小–数据缩放	实验中使用的模型家族和随机种子选择的详细信息；评估语料库和单位；离散度拟合与敏感性检查；字符模型训练和不可约损失参考值。	24
C. 池化、熵与固定难度组	几何离散度、算术离散度和 BALD；其规模律拟合，以及固定的基于熵的难度组内的比较。	32
D. 更改可用上下文	针对 Pythia 的上下文长度实验，其中目标固定，DataDecide 和 Qwen，包括评估流程、拟合趋势以及不同段落之间的变化。	34
E. 可见份额缩放与损失定律比较	分别讨论可见损失律和不可见损失律、可见比例对不可约损失估计的敏感性，以及与一个直接成员损失定律。	36
F. 上下文依赖的不确定性	基于参考的上下文级目标、熵和离散度评分，隐藏状态探针、预测评估和恢复比例跨上下文 .	42
G. 集成与更大型模型的比较	交叉点主张的证明、在参数匹配条件下的损失拟合和训练成本预算，以及实测 DataDecide 比较。	44
H. 共享训练与 LoRA 集成研究	全参数分支流程与拟合；多样性惩罚以及集成损失权衡（H.4）；LoRA 训练、秩、离散度，以及预测损失。	50
I. C4 验证结果	主要图表对应的 C4 结果，包括损失下限、上下文依赖的不确定性，以及匹配预算比较。	55

表 1 仅使用预测的重新校准映射。 所有映射均在拟合折上最小化下一字节交叉熵。多层感知机使用 GELU 隐藏激活函数。

Map	参数	拟合设置
温度	1	有界最小化，对数 $T \in [-2, 2]$
向量	512	L-BFGS，最多 150 次迭代； $w = 1$， $b = 0$ 初始
MLP	131,584	256 个隐藏单元；输入 0。1 最大值（log q， -30）
		AdamW，25 个训练轮次，批次大小为 8,192
		学习率 10^{-3}，余弦衰减；权重衰减 10^{-2}

A 校准

第 2 节将可约损失与校准条件下的认知不确定性联系起来。我们通过拟合仅基于预测的校正，并测量其对留出对数损失的降低，来检验校准误差。语言模型保持不变；映射不接收上下文或隐藏状态。

令 $c(q_X) = \Pr(Y = \cdot \mid q_X)$ 且 $C_{\text{cal}}(q) = \mathbb{E}\,\mathrm{KL}(c(q_X) \| q_X)$。仅使用预测的映射 a 满足

$$L(q) - L(a \circ q) = C_{\text{cal}}(q) - \mathbb{E}\,\mathrm{KL}(c(q_X) \| a(q_X)) \le C_{\text{cal}}(q).$$

因此，已取得的增益为校准误差提供了下界；较小的增益并不能为剩余误差提供上界。我们测试字符模型成员和几何池，以及单个公开检查点。字符模型的 256 字节词汇表使灵活的全分布校正更易于实现。

A.1 字符级模型

对于附录 B.7.1 中的架构，我们分别拟合每个成员及其有限几何池，不采用用于离散度的总体重缩放。奇数和偶数编号的留出的 FineWeb-Edu 序列分别定义拟合折和评分折；交换二者的角色，并对评分结果取平均。

温度缩放使用一个标量；向量缩放为每个对数概率分别提供缩放因子和偏移量；残差多层感知机（MLP）在归一化之前添加一个学习得到的修正。MLP 的零初始化输出层从恒等映射开始。表1给出了拟合设置，表2报告了每项评估设置，图9总结了大小和数据依赖性。

A.2 针对公开模型的常见重新校准测试

我们将相同的双参数映射应用于表3中的全部45个公开检查点：十个 Pythia、十一个 DataDecide、十三个 Qwen 和十一个 Delphi 模型。对于每种大小，我们使用发布的默认检查点或随机种子 0。令 $k = \arg\max_y q(y)$、 $p = q(k)$ 和 $g(p) = \sigma(a\,\mathrm{logit}(p) + b)$，其中逻辑函数为 σ。校正后的分布为

$$q'(k) = g(p), \qquad q'(y) = q(y)\frac{1 - g(p)}{1 - p} \quad (y \neq k).$$

我们使用交换后的奇偶源文本折，针对 $Y = k$ with $a \in [0.05, 20]$ and $b \in [-10, 10]$ 最小化二元交叉熵。恰好存储为 1 的概率保持不变。对于 $z = \mathbf{1}\{Y = k\}$，下一个词元损失的降低量为

$$\log \frac{q'(Y)}{q(Y)} = z \log \frac{g(p)}{p} + (1 - z) \log \frac{1 - g(p)}{1 - p}.$$

因此，保存最高概率词元和正确性指标就足以评估校正效果。所有模型都在相同的 96 个近期文本来源（附录 B.3）上使用 32位浮点数推理：Pythia 和 DataDecide 产生了 92,160 个预测，涵盖 432,878

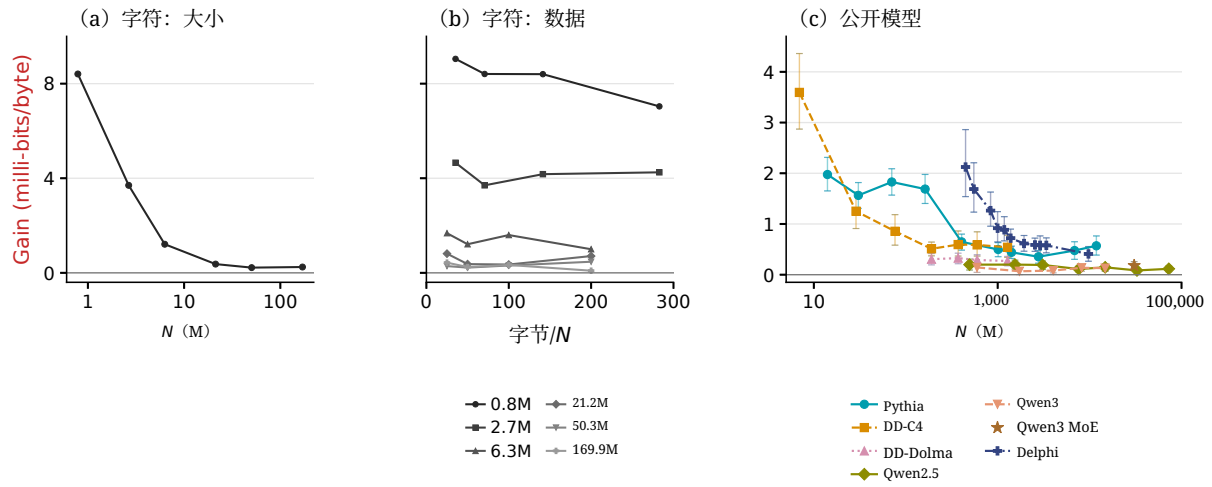


图9 跨模型和数据规模的重新校准。 以毫比特/字节计的留出多层感知机增益。（a）第二个保留预算下的字符模型几何池（每个解析参数对应 50–71 字节）。（b）实际 D/N；图例给出解析规模。（c）表4中的单个公开检查点；DD 表示 DataDecide，且 N 统计总参数量，包括所有混合专家。面板（c）使用更小的纵轴范围；柱状图表示条件95% 来源自助法区间，且拟合映射固定。所有增益均对交换后的拟合/评分折取平均。公开拟合和区间的局限性详见附录A.3。

字节；Qwen 产生了 89,485 个预测，涵盖 429,055 字节；Delphi 产生了 87,683 个预测，涵盖 431,488 字节。这些计数因分词和评分前缀截断而有所不同。该映射保留了非最高概率的相对概率，但无法检测其校准误差。

A.3 公开模型的全分布重新校准

我们对与上述相同的检查点拟合温度、向量和残差多层感知机映射，并对相同的目标进行评分。表4报告了所有带符号的增益；图9c按规模绘制了多层感知机结果。

每个外层折对 48 个未触及文本进行评分。温度使用另外 48 个（T_{48}）或 140 个拟合文本（T_{140}），额外加入 92 个独立文本。灵活映射使用 120 个内层拟合文本和 20 个内层验证文本（其中 8 个为原始文本，12 个为新增文本）。验证过程选择恒等映射、温度映射或灵活映射，以及其正则化和停止轮次；选定的方案随后在全部 140 个文本上重新拟合。向量和多层感知机分别报告，且不根据评分结果进行选择。

记 $\ell = \log q$ 和 $\beta = 1/T$，温度映射为 $q_T = \mathrm{softmax}(\beta\ell)$，其中 $\beta \in [e^{-2}, e^{2}]$。向量映射为

$$q_{\mathrm{vec}} = \mathrm{softmax}\big(\beta e^{a} \odot \ell + b - \bar{b}\big)\,,$$

在拟合温度下初始化。其惩罚项为 $\lambda(\|a\|_2^2 + \|b - \bar{b}\|_2^2)/(2n_{\mathrm{fit}})$，拟合令牌数为 n_{fit} 和 $\lambda \in \{1, 10, 100\}$。残差多层感知机为

$$q_{\mathrm{MLP}} = \mathrm{softmax}\left(\beta\ell + f_\theta\left(\frac{0.1\max(\ell, -30)}{\sqrt{V}}\right)\right),$$

其中 V 是词汇表大小，f_θ 包含 32 个 GELU 隐藏单元，第一层权重的初始值乘以 $\sqrt{V}$，输出层初始化为零。其惩罚项为 $\lambda\,\mathbb{E}\,\mathrm{KL}(q_{\mathrm{MLP}} \| q_T)$，其中 $\lambda \in \{0.1, 1, 10\}$。两种映射均通过最小化惩罚交叉熵进行拟合，并将 β 固定为其拟合集估计值。

两者均使用学习率 10^{-3}、25 个训练轮次的余弦调度、由 256 个微批次累积而成的 8,192 个词元批次，以及随机种子 0。向量缩放使用不带权重衰减的 Adam；多层感知机使用 AdamW

表2 在2,097,152个留出字节上对我们的字符级模型进行校准评估。 $N = 12hd^2$ 是用于字符模型缩放的解析参数计数，隐藏宽度为 d，并有 $h = d/32$ 层。每个条目都是以毫比特/字节为单位的有符号留出损失降低量。列 T、V 和 M 分别表示对几何集成进行温度、向量和 MLP 重新校准；成员列则对各个独立的 MLP 重新校准结果取平均。最后一列将集成MLP增益除以其在 $E = 0.593341$ 比特/字节处的可约损失。全部24个预选单元格均予以保留。

N (M)	字节/N	集成 T	集成 V	集成 M	成员 M	增益 / 可约损失
0.79	35.4	0.154	5.452	9.047	8.677	0.671%
0.79	70.7	0.465	5.253	8.409	8.232	0.652%
0.79	141.4	0.251	5.370	8.401	8.029	0.677%
0.79	282.8	0.280	4.197	7.038	7.174	0.588%
2.65	35.4	0.466	3.358	4.656	4.335	0.425%
2.65	70.7	0.288	2.466	3.700	3.468	0.365%
2.65	141.4	0.161	2.973	4.172	4.148	0.432%
2.65	282.8	0.074	2.837	4.250	4.379	0.454%
6.29	25.0	0.212	1.522	1.677	1.420	0.179%
6.29	50.0	0.371	1.603	1.210	0.990	0.138%
6.29	100.0	0.159	1.539	1.598	1.401	0.192%
6.29	200.0	0.346	1.374	0.998	0.830	0.126%
21.23	25.0	0.140	0.879	0.812	0.641	0.107%
21.23	50.0	0.234	0.676	0.370	0.272	0.052%
21.23	100.0	0.154	0.603	0.352	0.195	0.052%
21.23	200.0	0.105	0.850	0.709	0.579	0.111%
50.33	25.0	0.174	0.581	0.285	0.178	0.044%
50.33	50.0	0.193	0.510	0.221	0.124	0.037%
50.33	100.0	0.128	0.475	0.310	0.220	0.055%
50.33	200.0	0.075	0.564	0.470	0.381	0.087%
169.87	25.0	0.083	0.544	0.431	0.379	0.083%
169.87	50.0	0.128	0.434	0.246	0.175	0.051%
169.87	100.0	0.072	0.421	0.333	0.279	0.074%
169.87	200.0	0.121	0.276	0.091	0.052	0.022%

，权重衰减为 10^{-2}。验证检查第 1、2、4、8、16和25 个训练轮次；重新拟合保留该调度和所选训练轮次。

梯度拟合使用转换为32位浮点数并重新归一化的float16对数概率缓存。改进后的温度拟合、验证和评分使用原始的32位浮点数预测，并采用float64损失归约；每项增益都比较来自相同预测的校正结果。七项原始 Pythia T_{48} 评估则改用float16存档；针对新鲜的32位浮点数预测进行的四文本诊断无法提供涵盖所有文本的误差界。

条件95%区间使用1,000次完整来源的自助法重采样（随机种子2026110601），将重采样后的总增益除以总字节数。它们固定拟合映射，不计入拟合和选择不确定性，并且不在各检查点之间同时成立。

表3 公开模型的常见重新校准测试。 增益表示双参数顶级词元质量映射（见附录 A.2）所带来的留出集下一个词元对数损失下降，单位为毫比特/字节，且带有正负号。相邻列将相同的增益表示为原始对数损失的百分比。每个模型使用交替源文本进行拟合和评分，然后交换各折。所有 45 个检查点及带符号结果均予以保留。C4 和 Dolma 表示 DataDecide 训练语料库；其标签统计总参数量。

模型	Gain	% 损失	模型	Gain	% 损失
Pythia-14M	+0.226	+0.0176	DataDecideC4 7.0M	+0.040	+0.0026
Pythia-31M	+0.014	+0.0012	DataDecideC4 28.8M	+0.046	+0.0041
Pythia-70M	+0.107	+0.0098	DataDecideC4 76.4M	+0.001	+0.0001
Pythia-160M	+0.004	+0.0004	DataDecideC4 190.5M	+0.016	+0.0018
Pythia-410M	+0.099	+0.0118	DataDecideC4 371.5M	+0.040	+0.0049
Pythia-1B	+0.150	+0.0187	DataDecide C4 597.7M	+0.060	+0.0076
Pythia-1.4B	+0.129	+0.0167	DataDecide C4 1279.9M	+0.051	+0.0070
Pythia-2.8B	+0.131	+0.0178	DataDecide Dolma 190.5M	+0.030	+0.0034
Pythia-6.9B	+0.209	+0.0291	DataDecide Dolma 371.5M	+0.041	+0.0050
Pythia-12B	+0.310	+0.0440	DataDecide Dolma 597.7M	+0.049	+0.0063
Qwen2.5-0.5B	+0.174	+0.0216	DataDecide Dolma 1279.9M	+0.031	+0.0042
Qwen2.5-1.5B	+0.132	+0.0187	Delphi-447M	+0.133	+0.0139
Qwen2.5-3B	+0.073	+0.0108	Delphi-550M	+0.125	+0.0139
Qwen2.5-7B	+0.070	+0.0109	Delphi-837M	+0.083	+0.0096
Qwen2.5-14B	+0.034	+0.0056	Delphi-998M	+0.144	+0.0171
Qwen2.5-32B	+0.044	+0.0073	Delphi-1.2B	+0.122	+0.0150
Qwen2.5-72B	+0.007	+0.0011	Delphi-1.4B	+0.070	+0.0089
Qwen3-0.6B	+0.054	+0.0069	Delphi-1.9B	+0.060	+0.0078
Qwen3-1.7B	+0.057	+0.0080	Delphi-2.5B	+0.035	+0.0047
Qwen3-4B	+0.015	+0.0023	Delphi-2.9B	+0.041	+0.0056
Qwen3-8B	+0.043	+0.0069	Delphi-3.4B	+0.045	+0.0064
Qwen3-14B	+0.021	+0.0034	Delphi-9.7B	-0.019	-0.0029
Qwen3-30B-A3B	+0.031	+0.0052			

表4 公开模型的全分布重新校准。 在原始96篇源文本上，以毫比特/字节计的有符号留出对数损失降幅。 T_{48} 和 T_{140} 每个外层折分别使用48篇和140篇拟合文本；后者额外加入92篇独立文本。Vector和多层感知机报告基于20篇内层验证文本选择的校正结果，候选项包括恒等映射和温度，然后在全部140篇拟合文本上重新拟合。评分折未被使用。Vector和多层感知机单元格给出点增益，以及条件95% 区间；这些区间通过整源自助重采样获得，并固定拟合映射，排除拟合和选择不确定性；区间并非跨模型同时成立。 N 统计总参数量（百万）。正文中报告的百分比，是将相同增益除以原始总对数损失所得。较小的已实现增益为全分布校准误差提供下界。

模型	N (M)	T_{48}	T_{140}	向量	MLP
Pythia					
14M	14.1	-0.006	-0.001	+2.192 [+1.484, +2.896]	+1.975 [+1.652, +2.313]
31M	30.5	+0.113	+0.110	+2.144 [+1.533, +2.748]	+1.564 [+1.307, +1.817]
70M	70.4	-0.010	-0.014	+3.614 [+2.856, +4.312]	+1.828 [+1.569, +2.089]
160M	162.3	+0.414	+0.407	+2.753 [+2.139, +3.361]	+1.689 [+1.404, +1.977]
410M	405.3	+0.100	+0.103	+0.948 [+0.632, +1.254]	+0.647 [+0.484, +0.800]
1B	1,011.8	+0.073	+0.089	+0.643 [+0.404, +0.861]	+0.497 [+0.355, +0.632]
1.4B	1,414.6	+0.093	+0.100	+0.574 [+0.366, +0.772]	+0.441 [+0.319, +0.563]
2.8B	2,775.2	+0.080	+0.082	+0.699 [+0.427, +0.971]	+0.354 [+0.227, +0.477]
6.9B	6,857.3	+0.154	+0.159	+0.718 [+0.505, +0.929]	+0.476 [+0.303, +0.650]
12B	11,846.1	+0.265	+0.263	+0.784 [+0.567, +1.013]	+0.570 [+0.387, +0.764]

续见下页

表4续

模型	N (M)	T_{48}	T_{140}	向量	MLP
DataDecide C4					
4M	7.0	+0.016	-0.013	+2.311 [+1.225, +3.352]	+3.594 [+2.871, +4.361]
20M	28.8	+0.050	+0.047	+1.055 [+0.314, +1.801]	+1.252 [+0.911, +1.631]
60M	76.4	+0.005	+0.006	+0.906 [+0.332, +1.471]	+0.856 [+0.582, +1.184]
150M	190.5	-0.006	-0.001	+0.819 [+0.271, +1.375]	+0.511 [+0.372, +0.643]
300M	371.5	-0.002	-0.002	+0.930 [+0.434, +1.429]	+0.594 [+0.367, +0.862]
530M	597.7	+0.001	+0.014	+1.029 [+0.562, +1.501]	+0.588 [+0.368, +0.846]
1B	1,279.9	+0.036	+0.042	+1.182 [+0.690, +1.680]	+0.541 [+0.341, +0.777]
DataDecide Dolma					
150M	190.5	-0.029	-0.011	+0.317 [+0.156, +0.471]	+0.306 [+0.194, +0.416]
300M	371.5	-0.004	-0.001	+0.265 [+0.132, +0.391]	+0.327 [+0.217, +0.424]
530M	597.7	-0.005	-0.001	+0.273 [+0.054, +0.482]	+0.290 [+0.179, +0.394]
1B	1,279.9	-0.000	+0.004	+0.208 [+0.119, +0.300]	+0.270 [+0.176, +0.361]
Qwen2.5					
0.5B	494.0	+0.072	+0.080	+0.233 [+0.090, +0.380]	+0.198 [+0.092, +0.307]
1.5B	1,543.7	+0.120	+0.123	+0.226 [+0.099, +0.344]	+0.199 [+0.112, +0.286]
3B	3,085.9	+0.110	+0.110	+0.243 [+0.113, +0.356]	+0.192 [+0.102, +0.270]
7B	7,615.6	+0.080	+0.081	+0.192 [+0.081, +0.298]	+0.111 [+0.054, +0.165]
14B	14,770.0	+0.045	+0.042	+0.187 [+0.081, +0.285]	+0.145 [+0.073, +0.216]
32B	32,763.9	+0.053	+0.053	+0.189 [+0.085, +0.281]	+0.084 [+0.031, +0.135]
72B	72,706.2	+0.021	+0.023	+0.124 [+0.024, +0.214]	+0.116 [+0.059, +0.171]
Qwen3-Base					
0.6B	596.0	+0.045	+0.044	+0.169 [+0.071, +0.263]	+0.143 [+0.046, +0.236]
1.7B	1,720.6	+0.029	+0.030	+0.094 [+0.019, +0.165]	+0.068 [+0.002, +0.133]
4B	4,022.5	+0.040	+0.040	+0.151 [+0.074, +0.227]	+0.081 [+0.036, +0.125]
8B	8,190.7	+0.036	+0.036	+0.169 [+0.092, +0.242]	+0.136 [+0.067, +0.206]
14B	14,768.3	+0.053	+0.053	+0.162 [+0.087, +0.232]	+0.129 [+0.068, +0.196]
Qwen3 MoE					
30B-A3B	30,532.1	+0.052	+0.052	+0.192 [+0.110, +0.270]	+0.180 [+0.098, +0.257]
Delphi					
447M	447.2	+0.039	+0.038	+2.972 [+2.223, +3.787]	+2.123 [+1.540, +2.861]
550M	550.3	+0.008	+0.017	+2.625 [+1.988, +3.367]	+1.688 [+1.234, +2.207]
837M	837.0	+0.020	+0.018	+2.302 [+1.751, +2.916]	+1.264 [+0.956, +1.629]
998M	998.0	+0.004	+0.013	+1.978 [+1.450, +2.537]	+0.917 [+0.646, +1.244]
1.2B	1,180.0	+0.037	+0.033	+1.734 [+1.254, +2.248]	+0.885 [+0.667, +1.144]
1.4B	1,384.6	-0.009	+0.003	+1.606 [+1.177, +2.068]	+0.721 [+0.545, +0.899]
1.9B	1,934.7	+0.010	+0.006	+1.535 [+1.116, +1.976]	+0.615 [+0.463, +0.773]
2.5B	2,544.6	+0.022	+0.018	+1.551 [+1.147, +1.968]	+0.588 [+0.460, +0.721]
2.9B	2,895.2	+0.025	+0.014	+1.391 [+1.007, +1.793]	+0.580 [+0.399, +0.765]
3.4B	3,383.1	+0.006	+0.004	+1.390 [+1.019, +1.774]	+0.579 [+0.431, +0.726]
9.7B	9,714.7	-0.004	-0.002	+0.864 [+0.560, +1.155]	+0.409 [+0.265, +0.555]

A.3.1 Pythia 模型

我们使用最后一步为 143,000 的默认发布版本。表5还保留了一项较早的七模型研究，该研究使用字符模型的 256 隐藏单元多层感知机，并通过单独的验证来选择恒等映射或学习率。上文的共享研究则在全部十种规模中使用正则化的 32 隐藏单元映射。

A.3.2 DataDecide 模型

我们在种子比较中使用经过匹配的 C4 和 Dolma 训练检查点的默认种子（表8）。其步骤列于附录 B.2 中；其中有几个步骤早于计划的最后一步。

表5 仅基于 Pythia 的预测进行校正。增益是以毫比特/字节表示的有符号留出损失降低量。即使会使损失变差，我们的字符级模型所采用的 MLP方案仍会保留。嵌套验证会在每个拟合折内分别选择恒等映射或学习率；外部评分文本从不选择校正方式。最后几列给出了每一折对应的选择。

N (M)	温度	原始MLP	验证后MLP		第0折	第1折
14.1	-0.006	+0.528	+0.528	10^{-3}	10^{-3}	
30.5	+0.113	-238.249	+0.000	恒等映射	恒等映射	
70.4	-0.010	-245.615	+0.000	恒等映射	恒等映射	
162.3	+0.414	+0.361	+0.135	恒等映射	10^{-3}	
405.3	+0.100	+0.203	+0.000	恒等映射	恒等映射	
1011.8	+0.073	-191.980	+0.000	恒等映射	恒等映射	
1414.6	+0.093	-90.873	+0.000	恒等映射	恒等映射	

表 6 **公开模型的置信度校准**。 范围表示最终单个模型的 15 箱期望校准误差（ECE），单位为百分点。Pythia 每个模型使用 92,160 个预测；Qwen 使用 89,485 个预测，基于相同的 96 条源文本。规模依据模型发布名称标注。

族	模型	模型规模	ECE 范围
Pythia	10	14M–12B	0.436–1.642
Qwen2.5 稠密模型	7	0.5B–72B	0.566–1.329
Qwen3 稠密模型	5	0.6B–14B	0.626–0.912
Qwen3 MoE	1	30B-A3B	0.743

表 7 族内 Qwen置信度ECE 的变化，较大模型减去较小模型，单位为百分点。范围采用相同 96 条源文本进行配对重抽样得到；这些范围仅作描述之用，未针对多重比较进行调整。

比较	变化	配对 95% 范围
Qwen2.5-0.5B →Qwen2.5-14B	-0.689	[-0.911, -0.319]
Qwen2.5-0.5B →Qwen2.5-72B	-0.764	[-0.948, -0.396]
Qwen2.5-14B →Qwen2.5-72B	-0.074	[-0.294, +0.157]
Qwen3-0.6B → Qwen3-14B	-0.247	[-0.420, +0.107]

A.3.3 Qwen 模型

我们评估了最终的 Qwen2.5 和 Qwen3 基础模型，包括 Qwen3-30B-A3B。更大型模型使用全新的 32位浮点数评估，并直接记录预测令牌 ID。

A.4 置信度校准

对于 Pythia 和 Qwen，我们还报告了最高概率令牌置信度期望校准误差：在 15 个等宽分箱中，按预测加权的置信度与准确率绝对差的均值（表 6）。配对 95%区间通过对相同的 96 个源文本进行 1,000 次重采样计算，并保留每个文本的所有预测。表 7 报告了通义千问系列内部比较。

B 模型、评估与大小–数据缩放

本附录介绍第 3.1 节和第 3.2 节中的评估与拟合细节。公开模型规模统计所有参数。除非另有说明，损失和离散度均使用自然对数（如对数似然）除以

计分字节数（按字符取平均）。我们按照第 2.2 节对相关量进行 $S/(S-1)$ 重缩放。

B.1 集成系列

独立随机种子。 我们使用了来自 Pythia（1400万–4.1亿，保留 6–10 个成员）、DataDecide（700万–12.8亿，三个成员）和 Marin Delphi（4.47亿–97亿，十一个规模、每个规模三个成员）的独立种子集成（Biderman 等人，2023年；van der Wal 等人，2025年；Magnusson 等人，2025年；Held，2026年；Marin Community，2026年）。它们的训练语料库分别为 Pile——由 22 个来源组成的集合（Gao 等人，2020年）；C4——经过清理的网页文本（Raffel 等人，2020年）；或 Dolma——由网页文本、书籍、代码及其他来源混合而成的数据集（Soldaini 等人，2024年）；以及由网页、代码和数学内容组成的共享混合语料库。DataDecide 使用公开发布的 OLMo Transformer 系列（Groeneveld 等，2024）。我们还纳入了一个来自去重版 Pile v0 发布版本的独立公开 69 亿参数 Pythia 随机种子标注对。

发布版本重跑和小幅配方变更。 我们结合了在标准 Pile 或去重后的 Pile 上训练的 Pythia v0/v1 发布版本配对、由不同发布版本组成的 a6.9B 三元组，以及每种大小下的五种 DataDecide 训练配方。我们还纳入了哈勃近似副本，其训练文本的差异约占 0.02–0.08% 的词元；以及 OLMo-7B 与其单独训练的 Twin，二者的词元预算均为两万亿。它们的成员总参数量大致覆盖 7000 万至 120 亿。

更大规模的训练或架构变更。 我们配对了 Qwen2.5 和 Qwen3 代际中大小相近的稠密模型（Yang 等人，2025），以构建集成，同时还比较了在标准 Pile 和去重后的 Pile 上训练的 Pythia 模型。我们还纳入了 Qwen2.5-32B/Qwen3-30B-A3B 稠密模型/混合专家（MoE）模型配对，并将其与五组稠密 Qwen 配对组分开处理。我们最大的集成是 DeepSeek-V3/V3.1 配对（DeepSeek-AI，2024），每个成员的参数量达到 6710 亿；其成员共享早期训练阶段。我们还将 Nemotron-3 Nano/3.5 Lightning（英伟达，2025年、2026年）以及 Gemma-4 31B/26B-A4B 稠密模型/MoE 模型配对（Gemma Team，2026年）作为集成纳入其中。我们汇总对齐后的词元分布，并按字节计算损失和离散度。表 9 列出了图 2b 中的 72 项公开模型比较。

我们的字符级模型系列。 我们在 FineWeb-Edu 上训练字节级 Transformer。FineWeb-Edu 是根据教育质量筛选出的网页文本（Penedo 等人，2024年），模型总参数量为 0.86M–170M。主要的离散度研究采用五种规模，范围为每个解析参数 50–400 个字节；最小规模保留两个随机种子，其余规模各保留三个（图5a；图12）。这些分析使用留出的 FineWeb-Edu 文本。图2在近期文本上评估相同的检查点。训练自有模型使每个数据预算都拥有独立的学习率冷却阶段；现有公开检查点中缺少这一点——例如，Pythia 的中间检查点不会针对每个预算分别设置冷却阶段。此外，包含 256 个字节值的词汇表使参数量低于百万的模型成为可能。这一规模使我们能够在计算预算内开展后期分支实验。我们对超参数进行了细致优化，从而为该模型系列获得清晰的缩放定律。附录 B.7介绍了训练流程。

B.2 预训练随机种子群体

主要的随机种子比较在每个集成内部共享训练文本和分配方案：Pythia 使用标准的、未经去重的 Pile，DataDecide 使用 C4 或 Dolma，而 Delphi 使用 Nemotron-CC、StarCoderData 和 ProofPile 2 的混合数据（Biderman 等人，2023年；van der Wal 等人，2025年；Magnusson 等人，2025年；Held，2026年；Marin Community，2026年）。表8列出了所有保留的规模及成员数量。在不同规模下，Pythia 采用统一的 3000亿词元预算；DataDecide 遵循其发布的、依赖规模的训练计划，并在每次随机种子比较中于匹配的步数进行比较；Delphi 的训练词元数量则从 12.2亿到 1603.7亿不等。规模定律描述了每个族观测到的分配计划，包括训练预算随规模增长时规模与数据量的联合变化。DataDecide 按 C4 规模递增顺序对应的检查点步数为

表8 近期文本种子测量，包括单独标注的公开 6.9B 去重 v0 配对。该配对不纳入主要拟合。损失和校正几何离散度的单位为纳特/字节；分数使用以 V4-Pro 为锚定的假设下限 $E = 0.33372841$。每一行均评估完整保留池，包括 14M 的六个成员和 31M 的九个成员。有限池损失 $L(q_S^{\mathrm{G}})$ 为直接实测；只有离散度会进行 $S/(S-1)$ 校正。

族	N (M)	S	$\bar{L}_S$	$L(q_S^{\mathrm{G}})$	$\widehat{\rho}_\infty^{\mathrm{G}}$	$\widehat{f}_\infty^{\mathrm{G}}$
Pythia	14.1	6	0.91467	0.87543	0.04708	8.10%
Pythia	30.5	9	0.82338	0.78469	0.04352	8.89%
Pythia	70.4	10	0.75479	0.71678	0.04224	10.03%
Pythia	162.3	9	0.65751	0.62693	0.03440	10.62%
Pythia	405.3	8	0.58780	0.56178	0.02974	11.70%
Pythia 去重版 v0	6857.3	2	0.49224	0.48028	0.02391	15.09%
DataDecide C4	7.0	3	1.06419	1.03862	0.03836	5.25%
DataDecide C4	28.8	3	0.78230	0.75651	0.03868	8.62%
DataDecide C4	76.4	3	0.68822	0.66579	0.03364	9.49%
DataDecide C4	190.5	3	0.61666	0.59844	0.02732	9.66%
DataDecide C4	371.5	3	0.57274	0.55625	0.02474	10.35%
DataDecide C4	597.7	3	0.54724	0.53105	0.02428	11.37%
DataDecide C4	1279.9	3	0.50981	0.49594	0.02082	11.82%
DataDecide Dolma	190.5	3	0.61719	0.59931	0.02681	9.46%
DataDecide Dolma	371.5	3	0.56958	0.55390	0.02352	9.97%
DataDecide Dolma	597.7	3	0.54308	0.52778	0.02295	10.96%
DataDecide Dolma	1279.9	3	0.50481	0.49158	0.01984	11.60%
Delphi	447.2	3	0.66510	0.64449	0.03091	9.33%
Delphi	550.3	3	0.62298	0.60436	0.02793	9.66%
Delphi	837.0	3	0.59963	0.58203	0.02640	9.93%
Delphi	998.0	3	0.58321	0.56647	0.02511	10.07%
Delphi	1180.0	3	0.56536	0.54939	0.02396	10.35%
Delphi	1384.6	3	0.55064	0.53537	0.02291	10.56%
Delphi	1934.7	3	0.53123	0.51686	0.02155	10.91%
Delphi	2544.6	3	0.51976	0.50599	0.02066	11.10%
Delphi	2895.2	3	0.50448	0.49138	0.01964	11.50%
Delphi	3383.1	3	0.49200	0.47933	0.01900	12.01%
Delphi	9714.7	3	0.44627	0.43545	0.01622	14.41%

5,725、14,584、29,042、37,500、45,000、51,250 和 69,369。Dolma 在四个最大规模下使用相同的步数；在计划的最后一步之前还有若干个检查点。

如果某个 Pythia 成员的近期文本损失比套件中位数高出超过 0.01 纳特/字节，我们就将其排除。在每个大小下，我们评估所有保留成员的几何池，包括由六个成员组成的 14M 池和由九个成员组成的 31M 池。实测有限池损失为 $L(q_S^{\mathrm{G}}) = \bar{L}_S - \rho_S^{\mathrm{G}}$；只有显示的离散度会进行 $S/(S-1)$ 总体重缩放。筛选可能会降低观测到的变异；附录 B.6 将检验其影响。

单独的大型 Pythia 配对使用 pythia-6.9b-deduped-v0 及其 seed42 对应版本。两者都属于较早的去重版 Pile 发布版本。seed-42 检查点没有模型卡，存档评估也未记录修订哈希。因此，我们将这对模型视为单独的公开比较，并将其排除在主要的大小、难度和恢复份额拟合之外。

B.3 评估数据与单位

图 10 比较了近期网页文本和 C4 验证集上的离散度。附录 I 给出了主要图表对应的 C4 版本。诸如 DataDecide-C4 这样的模型标签表示训练语料库；两组模型梯度都在两种文本样本上进行了评估。

近期样本来自 2026年8月 Common Crawl，即 CC-MAIN-2026-34 数据集。我们要求日期表述为 6 月至 8 月，并采用类似 C4 的清洗方式，以保留较长的英语散文。为减少出版商差异，我们在 C4 验证集或抽样训练分片中每个域组仅保留一份文档。通过固定的随机打乱方式选取评估文档。这些选择旨在减少可能的训练重叠以及与 C4 的差异，但无法保证文档的发布日期，也无法排除较早的引文和模板文本。因此，语料库差异并不能单独揭示记忆化。

近期源文本片段包含每份所选文档开头的 1,025 个 GPT-NeoX 词元；C4 样本还包含 96 个片段。每个公开模型都会对解码后的文本重新分词，并处理最多 1,024 个词元，将前 64 个词元用作上下文，然后进行评分。损失和几何离散度均按计分字节数归一化。兼容的集成成员会预测对齐的词元标识；跨分词器比较时，同一源文本片段可能会对不同前缀进行评分。难度分析则在共享的词元空间中以纳特/词元为单位。

B.4 更广泛的公开模型比较

表 9 列出了图 2b 中的公开模型集成。圆形表示随机种子比较，方形表示发布版本重跑和小幅训练方案变更，三角形表示更大规模的训练或架构变更。后两组包含随机种子之外的变异。其总体重缩放仅用于描述；实际的有限离散度通过乘以 $(S-1)/S$ 获得。

当成员的大小不同时，图中所示的数量采用其几何平均数。稀疏模型使用总参数而非激活参数。较小的水平偏移用于分开重叠的符号；拟合使用未偏移的参数数量。一个权重几乎相同的 Pythia 标准版/去重版二元组被排除在外；跨发布版本的比较仍不纳入仅基于随机种子的拟合。

B.5 离散度拟合及其不确定性

我们依据表 10 中的系列约定，对校正后的离散度的对数与参数数量的对数进行回归。表 11 给出了主要的随机种子定律。DataDecide 的拟合省略了最小的几种规模，因为这些规模处的斜率发生了变化。图 2b 分别对主要的 DataDecide C4 和 Dolma 随机种子比较进行拟合。Qwen 的拟合使用五个稠密跨代配对。表 16 给出了相应的拟合指数，以及算术平均离散度和 BALD 的结果。图 10 比较了两个评估语料库上的拟合趋势。

对于区间覆盖率，将不同规模按 $x_i = \log N_i$、$i = 1$、……、m 排序。令 $b_0 = x_1$、$b_m = x_m$，并令 $b_i = (x_i + x_{i+1})/2$ 适用于 $1 \le i < m$，再定义 $v_i = (b_i - b_{i-1})/(x_m - x_1)$。因此，规模权重之和为一。将某一设置的权重分摊给其观测成员，可以避免保留的随机种子总体较大而获得额外的损失拟合权重。在同一族内，将每个权重乘以相同的常数不会改变拟合系数。

单模型损失拟合。 图 2a 使用每个检查点自身的损失和参数总量。六个拟合的公开模型序列包含表10中按族列出的66个检查点。另外11个检查点仅作为测量值出现。在固定 E 的情况下，我们使用这些权重对 $\log(L-E)$ 关于 $\log(N/10^9)$ 进行回归，然后恢复 E。中心下限和范围 0.31727 以及 $[0.26019, 0.33373]$ 纳特/字节，来自附录 E.1 中独立的28设置分量定律剖面。每个彩色带包络了重新拟合的损失定律，涵盖

表 9 图2 的集成清单。 规模按总参数量计算，包括嵌入；成员数量不等的池采用成员数量的几何平均数，而非总和。对于混合专家模型，这里统计的是总参数量，而非激活参数量。每个 Pythia 随机种子套件都使用其完整保留池，其中 14M 有六个成员，31M 有九个成员。三组分别包含 28、27 和 17 个公开模型比较。

模型系列	集成构建方式	绘图规模 N
种子比较		
Pythia，标准 Pile	每种大小保留 6–10 个随机种子	14.1–405M
Pythia，去重版 v0	单独的公开随机种子标注配对	6.86B
DataDecide-C4	每种大小使用三个随机种子	6.96M–1.28B
DataDecide-Dolma	每种大小使用三个随机种子	191M–1.28B
Marin Delphi	每种大小使用三个随机种子	447M–9.7B
发布版本重跑、配方变体和近似副本		
Pythia 发布重跑	v0/v1 配对、标准 Pile 或去重后的 Pile	70.4M–11.85B
Pythia 混合发布	同一大小的三个成员	6.86B
DataDecide 配方	每种大小对应五种训练方案	191M–1.28B
哈勃	每组 2–12 个近似副本	1.18–8.27B
OLMo-7B / Twin	配对使用匹配的 2T 词元预算	6.89B
更大规模的训练或架构变更		
Pythia Pile / 去重版	改变训练数据的二元组	70.4M–11.85B
Qwen2.5 / Qwen3，稠密	五个跨代配对	543M–14.77B
Qwen2.5 / Qwen3，稠密/混合专家	一组代际对	31.63B
Gemma-4	稠密/混合专家对	27.83B
Nemotron-3 / 3.5	成对改变训练流程	31.58B
DeepSeek-V3 / V3.1	采用共享早期训练的成对模型	671.03B

表10 规模拟合的系列约定。 设置权重适用于扩散定律，并在每个设置内的观测检查点之间等分，用于单模型损失拟合。等权重在各设置之间保持不变；区间权重描述对数规模的覆盖范围。分量定律拟合对每个集成设置使用一个成员平均损失，并采用附录 E.1 中更窄的队列。参数数量指总参数量。

族	设置权重 v_i	单模型损失观测值
Pythia	各规模相同	每个规模一个已发布检查点（10 个检查点）
DataDecide C4	各规模相同	每个规模的默认检查点（7 个检查点）
DataDecide Dolma	各规模相同	每种规模的默认检查点（共 4 个检查点）
稠密 Qwen2.5/3	各规模相同	每一代分别进行拟合（总共 12 个检查点）
Delphi	对数规模区间覆盖率	全部 33 个已发布检查点；每种大小使用三个随机种子

表11 独立的真实种子规模拟合。表16给出了三种池化指标的公开模型参考定律及系数。规模范围按总参数量计算。拟合形式为 $\hat{\rho}_{\infty}^{\mathrm{G}} = c\,(N/10^{9})^{-\gamma}$。拟合使用附录 B.5 中的族权重。标准误和对数均方根误差描述的是在选定模型和文本条件下的加权回归残差；它们并非完整的抽样区间。

族	规模	N (M)	γ	SE	对数均方根误差
Pythia	5	14.1–405	0.138	0.020	0.041
DataDecide C4	5	76.4–1280	0.162	0.016	0.027
DataDecide Dolma	4	190.5–1280	0.152	0.018	0.017
Delphi	11	447.2–9715	0.200	0.008	0.022

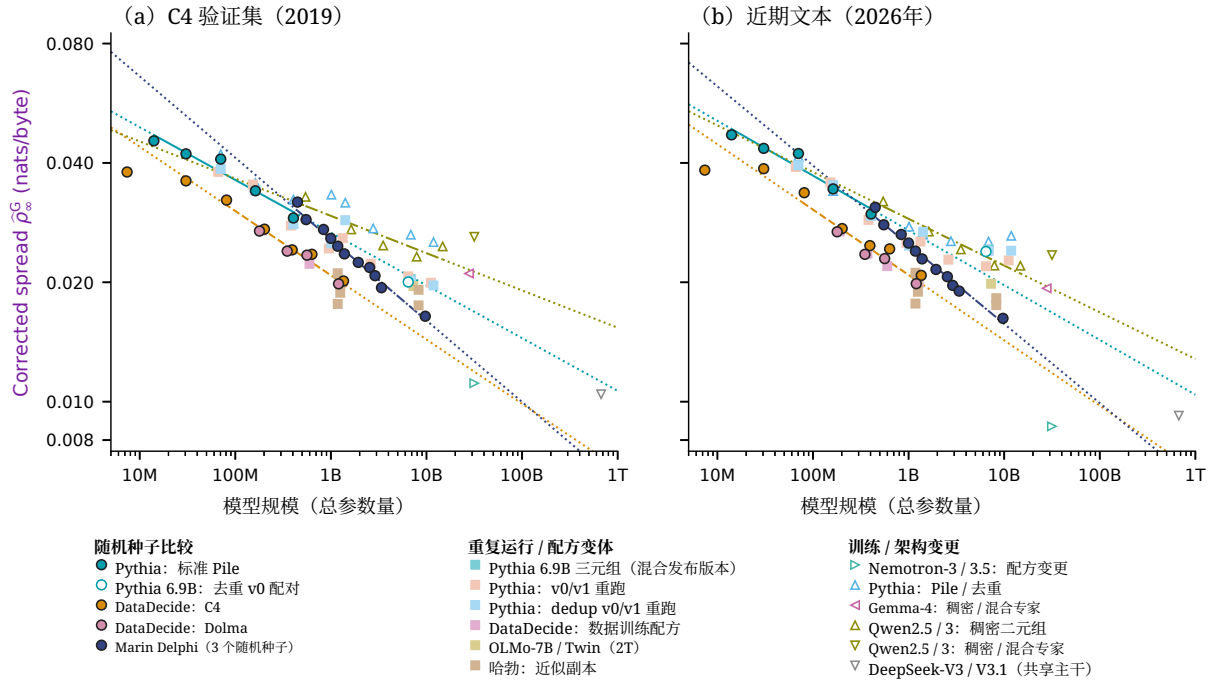


图 10 C4 验证集和近期文本上的离散度缩放。 图 2b 中的 72 项公开比较均在每个语料库上进行评估。标记类别和大小偏移沿用主要比较。每个语料库上的曲线均重新拟合；点线段表示外推。DataDecide曲线结合了随机种子和配方比较。Pythia筛选分别应用于每个语料库，并且每种大小都使用其完整的保留数据池。两个语料库的拟合均采用表10中的约定。这些测量未纳入任何不可约损失估计。

该范围；虚线曲线外推至该族最大实测大小之外，直至 1万亿。这些区间反映的是下限敏感性，而不是提供检查点置信区间。不会对单个检查点损失应用总体重缩放。

专门评测领域。 我们评估每个领域中固定的 32 段文本：来自 MoleculeNet/ChEMBL 脂溶性数据集的分子 SMILES 字符串（Wu 等，2018 年）、采用 PGN 记谱法记录的 2013 年 1 月 Lichess 对局（Lichess，2013 年）、采用 ABC 记谱法记录的诺丁汉民歌曲调（Foxley，2011 年），以及来自 mathlib4 的 Lean 形式数学来源（mathlib 社区，2020 年）。原始的 96 段 C4 样本提供了图 11 中的参考分布；这些专门来源构成领域偏移，但我们并未假设它们未出现在预训练数据中。

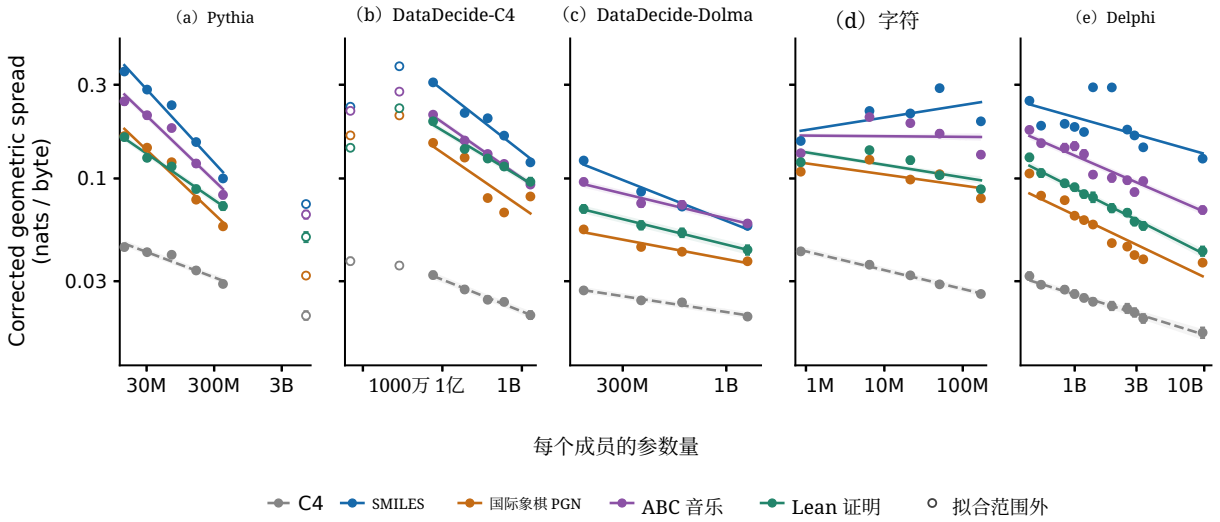


图 11 各评估领域中的种子集成离散度。（a）Pythia，（b、c）在 C4 和 Dolma 上训练的 DataDecide，（d）每个解析参数对应 100 个训练字节的字符模型，以及（e）Delphi 的全部 11 种规模。各点使用每种规模的完整保留种子池和 $S/(S-1)$ 校正，且不同语料库（包括 C4（灰色））中的成员保持不变。除 Delphi 使用对数规模区间权重外，其他情况的独立双对数拟合均采用相等的规模权重，并在各自观测到的拟合范围内停止。空心标记不纳入拟合：单独的 6.9B Pythia 配对模型以及两个最小的 DataDecide-C4 规模。柱线和带状区域表示 95% 区间，基于 3,000 次段落自助法重采样，并以检查点为条件。

表12 使用按字节归一化的扩散量时，成员选择和规模选择对几何扩散拟合的敏感性。DD 表示 DataDecide；Recent 表示 2026年8月的文本样本。经过筛选的 Pythia 拟合会排除损失超过套件中位数 0.01 纳特/字节以上的运行结果。报告的衰减指数 γ 为正值。删除范围每次省略一个规模；它们是敏感性检查，而非置信区间。

语料库	选择	规模	γ	对数均方根误差	删除范围
C4	Pythia，可用	5	0.154	0.035	$[0.138, 0.180]$
C4	Pythia，筛选后	5	0.132	0.041	$[0.109, 0.152]$
C4	DD C4，所有规模	7	0.127	0.048	$[0.118, 0.154]$
C4	DD C4，$N \geq 60$M	5	0.162	0.018	$[0.152, 0.167]$
C4	DD Dolma	4	0.155	0.020	$[0.126, 0.160]$
最近	Pythia，可用	5	0.169	0.027	$[0.161, 0.185]$
近期	Pythia，筛选后	5	0.138	0.042	$[0.118, 0.158]$
近期	DD C4，所有规模	7	0.128	0.062	$[0.118, 0.163]$
近期	DD C4，$N \geq 60$M	5	0.162	0.027	$[0.137, 0.168]$
最近	DD Dolma	4	0.152	0.017	$[0.140, 0.156]$

B.6 对成员和规模选择的敏感性

表 12 重复了过滤和不过滤情况下的拟合，并改变了纳入的大小范围。这项检查在每个 Pythia 大小下对校正后的成对离散度取平均，而主要分析则在每个大小下使用完整保留池。因此，其系数可能与表11不同。缓慢下降的趋势仍然存在，但过滤会实质性地改变指数。

B.7 我们的字符级模型

表 13 按模型规模划分的字符模型超参数。 各项研究所选用的批次大小和峰值学习率。批次以 512 字节的序列计数。三列学习率分别适用于隐藏矩阵（Muon）、输入嵌入（AdamW）以及输出头和归一化参数（AdamW）。权重衰减适用于隐藏矩阵和输出头。

宽度 d	层数 h	N (M)	批次	Muon学习率	嵌入学习率	输出头/归一化学习率
Spread 与全参数分支实验						
128	4	0.786	16	0.006087	0.066236	0.001775
192	6	2.654	32	0.006087	0.066236	0.001775
256	8	6.291	64	0.006087	0.066236	0.001775
384	12	21.234	128	0.006087	0.066236	0.001775
512	16	50.332	256	0.006087	0.066236	0.001775
768	24	169.869	512	0.006087	0.066236	0.001775
损失律比较						
128	4	0.786	1	0.001522	0.016559	0.000444
192	6	2.654	2	0.000761	0.008280	0.000222
256	8	6.291	8	0.001522	0.016559	0.000444
384	12	21.234	32	0.003044	0.033118	0.000888
512	16	50.332	256	0.006087	0.066236	0.001775
768	24	169.869	512	0.006087	0.066236	0.001775

词汇表 / 上下文长度	256 字节 / 512 字节
预热 / 冷却阶段占比	18 步 / 最后 20%
缩放预算（字节/N）	6.25–400，以 $\sqrt{2}$ 为倍数递增
Muon 动量	0.95
AdamW（β_1，β_2）/ ϵ	(0.9,0.95) / 10^{-10}
权重衰减	0.001423

B.7.1 训练设计和参数数量

我们的字符级模型是仅解码器 Transformer，在 FineWeb-Edu 上训练，每个词元一个字节。模型采用 RMSNorm、SwiGLU、旋转位置嵌入、查询/键归一化、64 维注意力头、不使用偏置，并且输入和输出嵌入不共享权重。随着训练预算增加，训练会处理更多语料位置。

令 d 表示隐藏宽度，$h = d/32$ 表示层数，D 表示训练字节数。字符缩放和分支采用解析计数 $N = 12hd^2$，该计数近似表示块矩阵，并排除归一化和词汇表权重。表 13 列出了模型规模和训练网格。

训练先进行线性预热，然后保持恒定学习率。在每个实测预算下，模型和优化器状态的一个副本都会单独完成线性冷却阶段。随后，恒定学习率轨迹从保存的状态继续运行。这样，每个预算都对应一个已完成的学习率计划，相较于从一次长时间衰减中取得的中间检查点，这使数据对比更容易解读。

表 13 列出了各模型规模对应的批次大小和学习率。在每项研究中，这些设置在不同训练预算下保持不变。附录 H 介绍了分支和适配器流程。

B.7.2 离散度与可约损失拟合

为了比较离散度和可约损失随规模下降的速度，我们在图12所使用的相同模型规模、训练预算和留出的 FineWeb-Edu 字节上，对离散度、不可见部分和总可约损失进行拟合。离散度研究保留了宽度为 128 时的随机种子 0 和 2，以及宽度为 256、384、512 和 768 时的三个随机种子。由于训练轨迹异常，我们排除了宽度 192，以及宽度为 128 时的随机种子 1。这些拟合

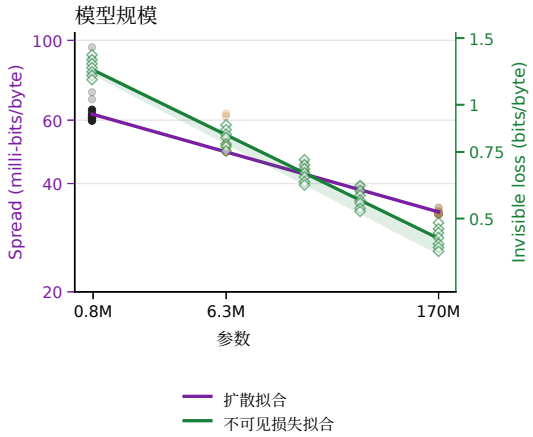


图12 字符模型扩散比不可见损失下降得更慢。 修正后的几何离散度（左轴，毫比特/字节）和有限池不可见损失 $L(q_S^{\mathrm{G}}) - E$（右轴，比特/字节），基于 524,288 个留出的 FineWeb-Edu 字节。拟合使用五种规模下相同的 35 组设置，每个解析参数对应 50–400 字节；颜色较浅的扩散数据点表示其余训练预算。绿色阴影表示所假设的不可约损失 E 从 0.593341 到 0.63764990比特/字节的变化，并展示下限敏感性。

表14 扩散比可约损失下降得更慢。 拟合使用与图12相同的 35 组设置：五种模型规模，以及每个解析参数对应 50–400 字节，并在 524,288 个留出字节上进行评估。对于 $Q(N) \propto N^{-\eta}$，指数越大，下降越快。各条目给出指数 $\pm$ 及其回归标准误。两列改变了所假设的不可约损失 E（单位为比特/字节）。扩散不依赖于 E，因此两项数值一致。

量 Q	$E = 0.593341$	$E = 0.637650$
几何离散度 $\widehat{\rho}_\infty^{\mathrm{G}}$	0.11654 ± 0.00163	0.11654 ± 0.00163
不可见部分 $L(q_S^{\mathrm{G}}) - E$	0.19140 ± 0.00512	0.20396 ± 0.00547
总可约损失 $\bar{L}_S - E$	0.18709 ± 0.00488	0.19887 ± 0.00521

使用解析计数 $N = 12hd^2$。将离散度针对总参数量重新拟合得到 $\gamma = 0.11844 \pm 0.00156$；将比特转换为纳特后，该指数保持不变。

对于图12中的不可见部分，我们首先撤销总体重缩放，以恢复实际的有限池损失：

$$L(q_S^{\mathrm{G}}) = \bar{L}_S - \frac{S-1}{S}\widehat{\rho}_\infty^{\mathrm{G}}.$$

然后，我们减去假定的不可约损失 E。表14比较了E的两种选项下的规模指数：在这两种情况下，离散度的下降速度都比不可见部分或总可约损失更慢。离散度本身不依赖于 E。这些拟合遵循每参数字节数的窗口，因此规模增大也会增加训练数据。连续的预算共享训练历史；回归误差并非独立运行的置信区间。

B.7.3 关于不可约损失的经验约束

工作中采用的不可约损失来自根据字符研究的上下文设置，对公开发布的模型损失进行外推。在这些设置下，Kimi-K2-Base（Moonshot AI，2025年）提供了已达到的最佳参考损失（表15）。我们将假定的不可约损失设定在前两个值之间；在这些设置下损失更高的 DeepSeek-V4-Pro-Base 则作为比较对象列出。

请注意，离散度本身不依赖于不可约损失；该值主要用于计算可见比例。附录E.3针对每个损失定律假设使用了相同的不可约损失选项。

C 池化、熵与固定难度组

表15 字符研究的不可约损失敏感性。 所有值均使用512字节上下文设置。达到的上端点向下取整后用于重新拟合。

选项	E (比特/字节)
联合外推的工作值	0.593341
最佳已达参考分布：Kimi-K2-Base	0.63764990
DeepSeek-V4-Pro-Base	0.68195265

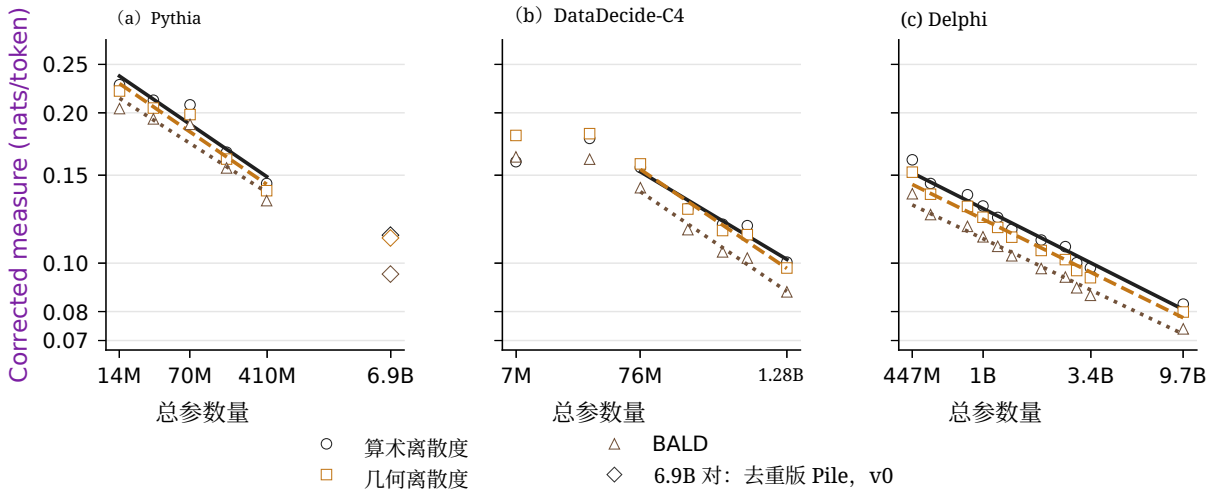


图13 在这两种损失池和 BALD 下，浅层的规模趋势仍然存在。 校正后的度量在每个族内使用匹配成员和近期文本位置。各点表示测量值；各线依据表10，对五个标准 Pile Pythia 套件、参数量超过 60M 的 DataDecide-C4 规模，以及十一组三元组 Delphi 进行拟合。空心菱形表示单独的 6.9B 去重 v0 配对，不纳入拟合。各线在其拟合范围处停止。

C.1 比较集成度量

我们在相同的成员和近期文本位置上比较第 2.2 节定义的度量。图13展示了 Pythia、DataDecide-C4 和 Delphi 种子集成。算术离散度和 BALD 的浅层大小趋势依然存在。

C.2 在全部三种集成度量下的缩放

表16给出了公开模型参考定律，以及在匹配成员总体上的拟合结果。每种度量均按计分字节进行归一化，并乘以 $S/(S-1)$。我们使用附录 B.5 中的族权重，在对数空间中通过最小二乘法拟合 $d(N) = c(N/10^9)^{-\gamma}$。标准误描述了在给定所选模型和文本条件下的回归残差。

三项度量的每次拟合都使用相同的 96 个近期文本窗口。对于 72 项比较中的每一项，几何平均和算术平均离散度都使用全部 96 个窗口。全窗口 BALD 可用于其中 65 项。由于 BALD 需要全词汇表预测，而这类预测占用大量存储空间，因此仅为其中 7 项比较保存了两个窗口的数据，导致这 7 项比较被排除在匹配度量拟合之外：标准 Pile Pythia 在 7000万、1.6亿、4.1亿和 69亿规模下的重复运行，以及 Nemotron、Gemma 和 DeepSeek 配对。

表16 公开模型参考定律和主要种子系列的缩放系数。 定律为 $d(N) = c(N/10^9)^{-\gamma}$。每个单元格在第一行给出 c（纳特/字节），第二行给出 γ，并在可估计时给出条件回归标准误；n 表示符合条件的比较数量。第一部分包含合并的 DataDecide参考。第二部分使用匹配的 96窗口队列，并在三项度量上采用全池估计量。图 2b 使用单独的 DataDecide种子拟合。拟合采用附录 B.5中的族权重。稠密 Qwen参考是描述性的跨代比较，而非独立种子估计。详情和排除项见 C.2节。

模型系列 / 参考分布	n	几何平均	算术平均	BALD
公开模型参考定律				
Pythia参考	5	0.02703 ± 0.00157	0.02802 ± 0.00183	0.02628 ± 0.00176
		0.138 ± 0.020	0.138 ± 0.023	0.129 ± 0.023
DataDecide参考	13	0.02087 ± 0.00031	0.02138 ± 0.00036	0.01924 ± 0.00019
		0.165 ± 0.012	0.153 ± 0.014	0.162 ± 0.008
稠密Qwen参考	5	0.02891 ± 0.00073	0.02904 ± 0.00076	0.02435 ± 0.00049
		0.118 ± 0.015	0.120 ± 0.016	0.117 ± 0.012
Delphi参考	11	0.02487 ± 0.00023	0.02613 ± 0.00027	0.02275 ± 0.00020
		0.200 ± 0.008	0.205 ± 0.009	0.194 ± 0.007
匹配系列定律：随机种子				
Pythia、standard-Pile种子	5	0.02703 ± 0.00157	0.02802 ± 0.00183	0.02628 ± 0.00176
		0.138 ± 0.020	0.138 ± 0.023	0.129 ± 0.023
DataDecide-C4种子	5	0.02163 ± 0.00050	0.02244 ± 0.00059	0.01946 ± 0.00036
		0.162 ± 0.016	0.143 ± 0.018	0.163 ± 0.013
DataDecide-Dolma 随机种子	4	0.02072 ± 0.00037	0.02079 ± 0.00033	0.01899 ± 0.00032
		0.152 ± 0.018	0.147 ± 0.016	0.152 ± 0.017
Delphi 随机种子	11	0.02487 ± 0.00023	0.02613 ± 0.00027	0.02275 ± 0.00020
		0.200 ± 0.008	0.205 ± 0.009	0.194 ± 0.007

C.3 保持难度组不变

OLMo-7B 的词元熵将每个近期文本位置分配到六个固定的百分位组之一。随后，我们比较 Pythia 不同规模模型中的相同词元。固定组别归属可以避免随着模型性能提升，仅因词元被移动到另一个组而使其看起来更容易。参考熵是难度代理指标，而不是实测的不可约随机性。对于每个组 b，我们以每个集成等权的方式拟合 $\log d_b(N) = \log A_b - \gamma_b \log(N/10^9)$。图14展示了三项度量中从容易到困难的对比。

D 改变可用上下文

这里我们详细介绍上下文长度在集成离散度中的作用。为此，我们改变前置上下文，同时保持模型权重和目标段落不变。

D.1 模型与评估

我们使用表 8 中的 Pythia、DataDecide 和 Delphi 规模，但排除单独的 6.9B Pythia 配对模型，以及三个最小的稠密 Qwen2.5/Qwen3 配对（0.5B/0.6B、1.5B/1.7B 和 3B/4B；附录 B.1）。Pythia 在每个规模下使用三个成员：默认发布版本以及随机种子 1–2；但在 160M 规模下，默认发布版本未通过损失筛选器（附录 B.2），因此改用随机种子 3。DataDecide 使用 default、small-aux-2 和 small-aux-3；在最大规模下，则将辅助运行替换为 large-aux-2 和 large-aux-3；Delphi 使用随机种子 0、42 和 62746。

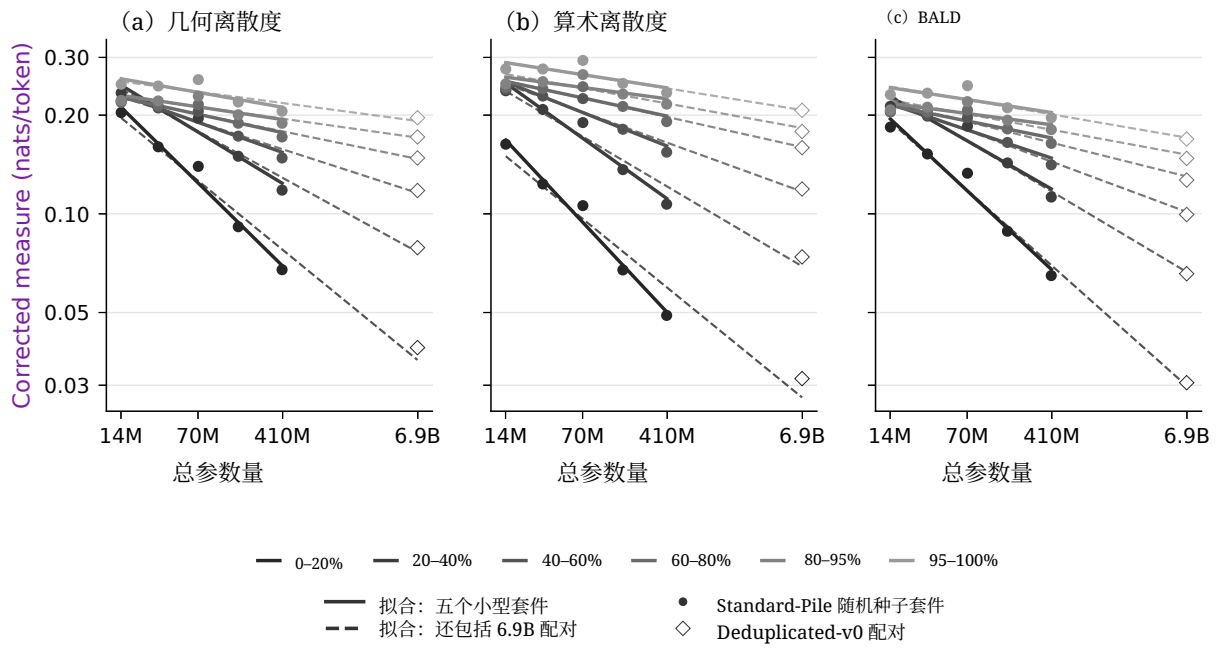


图14 在三项度量下，低熵词元的分歧下降得更快。 Pythia种子集成在相同的近期文本位置上进行评估，并根据 OLMo-7B 的预测熵划分为六个固定百分位组（从深到浅：熵值从最低到最高）。各面板展示了各组的 (a) 几何离散度、(b) 算术离散度和 (c) BALD 均值，并乘以 $S/(S-1)$，针对 S 个集成成员，单位为纳特/词元，并以双对数坐标轴表示其相对于总参数量的关系。实心圆表示参数量为 14M–410M 的五个 Standard-Pile 套件；空心菱形表示独立的 6.9B 去重 v0 配对。实线是对这五个套件拟合的幂律；虚线则加入 6.9B 配对后重新拟合，混合了两种训练方案。两种拟合都显示，低熵词元的下降更陡峭，而高熵词元的分歧持续时间更长。

分布使用 50,277 个已分配的 Pythia词元ID、50,280 个 DataDecide ID、151,665 个共享的 Qwen ID，以及 128,256 个已分配的 Delphi ID。Qwen 使用原生上下文限制，不进行位置扩展，也不使用聊天模板。我们从 77 本符合条件的 PG-19 测试书籍中采样 128 个段落（Rae等，2019）。每本书贡献 128 个连续目标，因此每种设置共有 16,384 个目标。每个族都会对相同的段落重新分词，并在不同上下文长度下固定其目标。目标区块彼此不重叠，但其上下文可能重叠。预训练重叠情况未知。

我们以 2 的幂次改变上下文上限 C，即用于预测的可用前置词元的最大数量；对于 Pythia、DataDecide 和 Delphi，最大可达 2,048 个词元，对于 Qwen 则最大可达 32,768 个词元。为提高效率，每次前向传播都会对一组连续的 B（C）个目标词元进行评分，其中

$$B(C) = \min\{128, \max[1, \lfloor C/16 \rfloor]\} \tag{12}$$

因此，每次遍历会对 1 到 128 个目标进行评分。在一个块内，可用上下文会随着每个目标增加一个词元：对于第一个目标，为 $C - B(C) + 1$ 个前置词元；对于最后一个目标，则为 C。我们绘制这一范围的几何平均中点 $c = \sqrt{(C - B(C) + 1)C}$。对于 $C \leq 16$，$B(C) = 1$，因此每个目标都恰好有 C个前置词元。我们会重新计算截断前缀，以防被丢弃的上下文残留在缓存中。

D.2 度量与拟合

离散度和 BALD 均遵循第 2.2 节中的定义及 $S/(S-1)$ 重缩放。数值单位为纳特/词元，每个分词器分别使用独立坐标轴。误差线表示各段落均值的一个标准差。对于每种度量，我们进行拟合

$$\bar{d}(c) = a + b\log_2(c/32) \tag{13}$$

采用非加权最小二乘法，在上下文限制 $C > 16$范围内进行拟合，并设置自由截距和斜率。这里的 b 是每次上下文加倍对应的变化量。拟合从 $C = 32$开始，因此斜率包含了初始收缩的一部分。我们通过对用于评估的 77PG-19 测试书籍进行有放回重采样，并重新拟合斜率 2,000 次，来估计 95% 斜率区间。每次重采样都保留一本书中的段落整体，并在不同上下文长度和模型设置之间重复使用。对于训练好的模型保持不变时，这些区间反映了评估文本之间的变异。

D.3 结果

图 15 比较了各模型家族中的三种分歧度量。在初始收缩之后，对于 Pythia、Qwen、DataDecide Dolma 和 Delphi，随着上下文进一步增加，几何离散度变化很小；DataDecide C4 的离散度略有上升。BALD 和算术离散度表现相似。在三个 Qwen 配对模型中，拟合的几何离散度斜率范围为 -1.888 至 -1.116 纳特/词元/每次上下文加倍。同样以这些单位计，Pythia-160M 的斜率为 -0.304 （配对 95%区间为 $[-0.822, 0.210]$），而 DataDecide-C4- 1.28B 为 $+3.919$ （$[2.480, 5.935]$）。DataDecide-Dolma 的几何斜率范围为 -0.852 至 -0.573。

E 可见份额缩放与损失定律比较

E.1 拟合分离的可见与不可见定律

我们在第 3.3 节中拟合分量分离描述，并将公式（8）特化为几何池化，$P = \mathrm{G}$。其可见分量为 $\rho_\infty^{\mathrm{G}}(N) = c_{\mathrm{G}} N^{-\gamma_{\mathrm{G}}}$，不可见分量为 $L(q_\infty^{\mathrm{G}}) - E = AN^{-\alpha}$。我们根据五个系列中 28 种近期文本设置，在图3中估计共享下限。这些设置包括三个规模高于 60M 的标准 Pile Pythia 套件、五个 DataDecide-C4 数据集、四个 DataDecide-Dolma 数据集、五组稠密 Qwen2.5/3 配对模型，以及十一个 Delphi 三元组。单独的 6.9B Pythia 配对模型也在图中展示，但未纳入这些拟合。我们统计总参数量，并对规模不等的成员使用其几何平均数。这些拟合样本范围窄于图 2b 中的样本；尤其是，此处的 Pythia fit 排除了 14M 和 31M 套件，因此其扩散指数与主拟合不同。

对于大小为 S 的测量集成，我们使用第 2.2 节中的校正：

$$\widehat{\rho}_\infty^{\mathrm{G}} = \frac{S}{S-1}\rho_S^{\mathrm{G}}, \qquad \widehat{f}_\infty^{\mathrm{G}}(E) = \frac{\widehat{\rho}_\infty^{\mathrm{G}}}{\bar{L}_S - E}. \tag{14}$$

相应的 $L(q_\infty^{\mathrm{G}}) - E$ 估计值是校正残差 $\bar{L}_S - E - \widehat{\rho}_\infty^{\mathrm{G}}$。它不同于实测有限集成的不可见损失 $L(q_S^{\mathrm{G}}) - E$。对于跨代 Qwen 配对模型，该校正仍然只是描述性重缩放。

对于每个公共族，我们使用 w_i 表10中的归一化设置权重 ρ_∞^{G}，在双对数空间中通过最小二乘法将N拟合为 $\widehat{\rho}_\infty^{\mathrm{G}}$。随后固定这些可见律，并拟合一个公共不可约损失 E

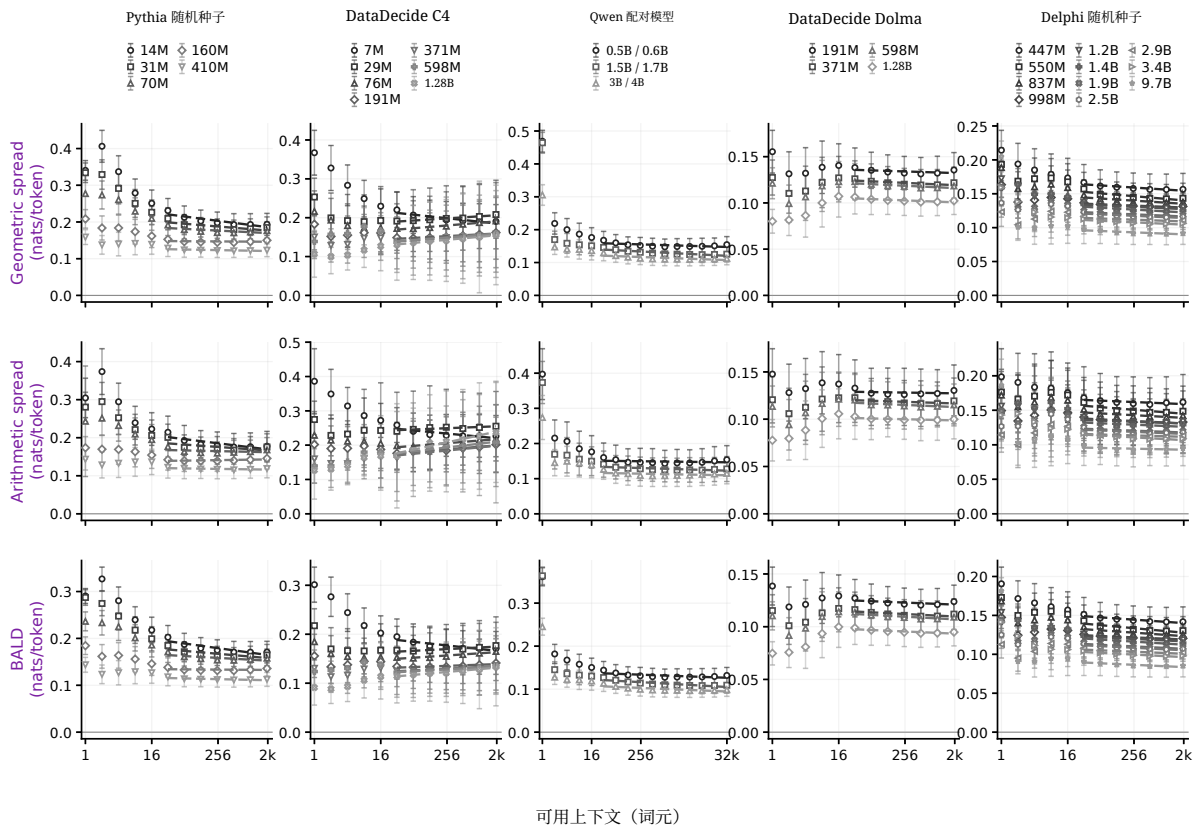


图 15 随着上下文增长的集成分歧。 各行分别展示几何离散度、算术离散度和 BALD；各列分别展示 Pythia、DataDecide C4、Qwen、DataDecide Dolma 和 Delphi。各点表示每个族在 16,384 个固定 PG-19 目标上的校正均值，并附带一个段落标准差。虚线拟合了高于 16 词元的上下文限制。

以及各族特有的不可见律。令 $g(i)$ 标识设置 i 所属的族，并令 L_i 表示其实测均值成员损失 L_S。损失拟合使下式最小化

$$\widehat{L}_i(E,\theta) = E + A_{g(i)} N_i^{-\alpha_{g(i)}} + c_{\mathrm{G},g(i)} N_i^{-\gamma_{\mathrm{G},g(i)}},$$
$$R(E,\theta) = \sum_{i=1}^{28} w_i \left[\log \widehat{L}_i(E,\theta) - \log \bar{L}_i\right]^2. \tag{15}$$

对于具有 n_g 个保留设置的族，我们将权重归一化为 $w_i = n_g v_i / \sum_{j:\ g(j)=g(i)} v_j$，使其总和等于该族的设置数量。每个集成都贡献一个平均成员损失。θ 中的十个系数为每个族给出一个不可见振幅和指数。

为评估对未知下限的敏感性，我们在 E 的候选值处重复进行损失拟合。对于每个候选值，我们保持可见扩散定律不变，并针对五个公开族分别重新拟合不可见振幅 A 和指数 α。这十个系数可根据扣除 E 后剩余的损失进行调整。由此得到的最小对数损失平方误差为 $R(E)$，而 $R_{\min} = \min_E R(E)$ 则是在所有候选下限中最佳拟合的误差。此过程称为对 E 进行剖面分析。

我们保留重新拟合误差足够接近 $R_{\min}$ 的下限，判断依据是名义95%残差F剖面。拟合包含十一个自由参数：一个共享下限和十个不可见律系数，因此剩余 $28-11=17$ 个残差自由度。判据为

$$R(E) \le R_{\min}\left[1+\frac{F_{1,17}(0.95)}{17}\right], \qquad R(E)=\min_{\theta} R(E,\theta). \tag{16}$$

此处$F_{1,\ 17}$(0.95) 是具有 1 个和 17 个自由度的 F分布的第95百分位数。固定E 会约束一个参数。我们将增量 $R(E)\ -R_{\min}$ 与估计残差方差$R_{\min}/17$进行比较，并保留 $R(E)\ \lesssim\ 1.26R_{\min}$。名义校准假设，在给定已拟合的扩散定律的条件下，加权残差 $\sqrt{w_i}$（log $L_i-\log L_i$）相互独立且近似为 ^ ¯高斯分布，并具有共同方差。由于设置权重包括覆盖权重，而不是估计的误差方差倒数，因此我们将该剖面解释为敏感性范围。

最佳拟合下限为 $E=0.31727$ 纳特/字节，未封顶剖面范围为 $[0.26019, 0.35190]$。随后，我们采用附录E.2中的经验上限：不可约损失不应超过评估分布上已经达到的最低损失。这会将保留范围截断为 $[0.26019, 0.33373]$纳特/字节。

图3将这一保留的下限范围传播到分量拟合和可见比例中。在每个候选下限下，五个公开族均使用相同的 E，并重新拟合其不可见律。在面板（b）中，每个阴影带涵盖这些重新拟合下、每个规模对应的公式（8）中拟合的 f_{∞}^{G}（N）的最小值和最大值。对于每个实测标记，损失 $\bar{L}_S$和校正后的离散度 $\hat{\rho}_{\infty}^{\mathrm{G}}$保持不变；其区间仅因 $\widehat{f}_{\infty}^{\mathrm{G}}$（$E$）$=\hat{\rho}_{\infty}^{\mathrm{G}}/$（$\bar{L}_S-E$）中的 E而变化。下限越大，分母越小，因此这一估计部分越大。这些阴影带和区间描述的是在拟合的扩散定律和观测集成条件下的下限敏感性。它们不包含扩散拟合不确定性、随机种子和文本采样以及系统误差，因此并非可见比例的完整95%不确定性区间。

对于图2和图3中的字符序列，我们在 Pythia 使用的432,878个近期文本目标字节上评估相同的35个规模—预算设置（98个检查点），并最多使用前置的512个字节。每个设置具有相同权重，并使用参数总数。对于单项损失拟合，其权重在各检查点之间平均分配。公开设置单独决定下限剖面；字符拟合将该剖面固定不变。其他字符实验使用留出的 FineWeb-Edu 文本。

在图3a中，可见标记为 $\hat{\rho}_{\infty}^{\mathrm{G}}$，不可见标记为 $\bar{L}_S-E-\hat{\rho}_{\infty}^{\mathrm{G}}$。每个组分及其拟合曲线，均以其所属族中最小拟合规模处该组分的实测值进行归一化。对于面板(a)中的带状区域，我们在每个候选 E 处重新拟合不可见曲线，并以同一 E下最小拟合规模处的实测不可见损失进行归一化。在每个规模处，该带状区域涵盖这些归一化曲线中的最小值和最大值。表17报告了 α、 γ_{G}以及预测的 f_{∞}^{G}。所有六个中心拟合均具有$\alpha>\gamma_{\mathrm{G}}$，因此在给定的下限选择下，拟合占比会随规模增加。附录E.2说明了这一趋势为何取决于 E。阴影带在每个族最小和最大的实测拟合规模处停止；在面板(b)中，只有中心曲线延伸到这些范围之外。

E.2 对未知不可约损失的依赖

表18给出了已达到的单模型损失。DeepSeek-V4-Pro-Base 实现了最低的近期文本损失，即 0.33372841 纳特/字节。这为近期文本下限设定了经验上限。Kimi-K2-Base 和 DeepSeek-V4-Pro-Base 使用官方原生 FP8 检查点，不使用聊天模板，并在与其他近期文本模型相同的源文本片段和上下文限制下运行。我们保留分词器默认行为，并按解码后的计分字节进行归一化。对于两个检查点，基于所有模型输出逻辑值进行归一化，与仅基于前$|\mathcal{V}_{\mathrm{tok}}|$ 个逻辑值进行归一化所得的损失相同（后者是其他模型所采用的过程）。

表17 近期文本上的独立分量拟合。 指数描述不可见损失 $L(q_\infty^{\mathrm{G}}) - E$ 以及可见离散度 ρ_∞^{G} 在公式（8）中的情况。28 个近期文本设置确定了共享下限 $E = 0.31727$ 纳特/字节。其未封顶剖面为 $[0.26019, 0.35190]$，根据经验上限截断为 $[0.26019, 0.33373]$。35 个字符设置使用相同的固定下限剖面。预测比例 f_∞^{G} 使用 10^{12} 总参数量。方括号给出保留的下限剖面上的范围。这些是外推，而非观测值。

族	不可见 α	可见 γ_{G}	$f_\infty^{\mathrm{G}}(10^{12})$ (%)
Pythia	0.28507	0.19988	19.3 [12.0, 22.6]
DataDecide-C4	0.24043	0.16184	17.0 [10.2, 20.2]
DataDecide-Dolma	0.25644	0.15189	19.3 [11.1, 23.4]
稠密 Qwen2.5/3	0.27564	0.11784	33.4 [17.4, 42.3]
Delphi	0.32614	0.19968	20.1 [10.6, 25.2]
字符	0.17036	0.08193	13.1 [10.5, 14.1]

表18 两个评估语料库上取得的单模型损失，单位为每个评分字节的纳特数。这两个语料库在这些模型中均使用相同的96个源文本片段，但会进行依赖分词器的截断。这些是实测交叉熵，而非条件熵估计值；两个语料库具有不同的文本分布。Kimi-K2-Base 和 DeepSeek-V4-Pro-Base 仅在近期文本上进行了评估；短横线表示未实测的 C4 损失。

模型	C4 验证集	近期文本
Qwen2.5-72B	0.479487	0.395215
DeepSeek-V3-Base	0.430436	0.365561
DeepSeek-V3.1-Base	0.442147	0.373619
GLM-4.5-Air-Base	0.450985	0.384772
GLM-4.5-Base	0.417401	0.352689
Kimi-K2-Base	—	0.340229
DeepSeek-V4-Pro-Base	—	0.333728

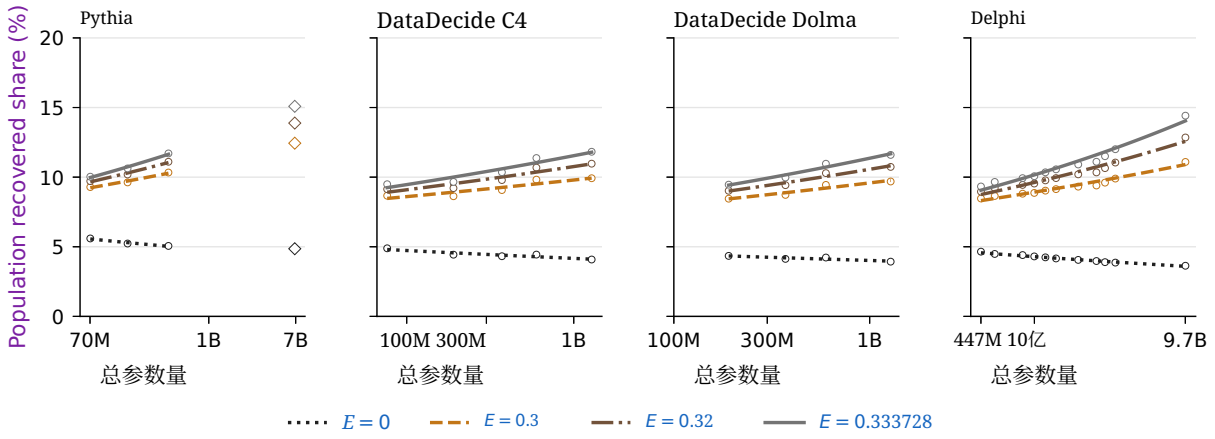


图16 恢复比例取决于假设的不可约损失。 曲线拟合的是校正几何离散度除以成员平均损失减去 E 在近期文本上的结果。公开曲线使用附录 B.5 中的族权重；空心菱形表示排除的 6.9B Pythia 配对模型。四种下限选择属于敏感性假设。

图16展示了 60M 参数以上的 23 个公开种子集的 E。公开拟合使用保留的种子族，并排除单独的 6.9B Pythia 配对模型。在图示的三个非零下限中，对于每个族，拟合得到的恢复份额都会随大小增加而上升。在 $E = 0$ 时，四个公开族的趋势均发生逆转。

表 19 直接恢复份额拟合的反转阈值。

族	E_{crit} (纳特/字节)
字符	-0.35597
Pythia	0.18996
DataDecide C4	0.20088
DataDecide Dolma	0.17154
稠密 Qwen2.5/Qwen3 配对	0.14397
Delphi	0.20384
Pythia（包括 6.9B；敏感性检查）	0.14014

表 19 给出了使 $\hat{\rho}_{\infty}^{\mathrm{G}}/(\bar{L}_S - E)$ 相对于 N 的拟合双对数斜率为零时的下限。字符模型的形式阈值为负值；对于允许的非负下限，其拟合恢复份额斜率仍为正。

以下损失律比较固定一个较早的 C4 渐近线，以比较两种对成员损失的描述。对于附录 I 中的验证结果，我们对五类模型拟合下限，并将其设为实测 C4 损失的最低值。图 10 中的语料库比较不需要假设下限。仅凭在不同语料库上损失更低，既不能证明记忆化程度更低，也不能证明熵估计更准确。

E.3 单一损失定律，还是分量定律之和?

第 3.3 节探讨了总可约损失或其不可见部分是否更适合用幂律描述。我们在同一目标上比较这两种描述：测得的成员平均损失。

拟合与验证。 方法A直接拟合传统定律：

$$\widehat{L}_{\mathrm{A}}(N, D) = E + AN^{-a} + BD^{-b}. \tag{17}$$

方法B对 $R_i = \bar{L}_i - \hat{\rho}_{\infty,i}^{\mathrm{G}}$ 拟合相同形式，单独拟合离散度，然后将二者的预测相加：

$$\begin{aligned} \widehat{R}(N, D) &= E + A'N^{-a'} + B'D^{-b'}, \\ \hat{\rho}_{\infty}^{\mathrm{G}}(N) &= cN^{-\gamma}, \qquad \widehat{L}_{\mathrm{B}} = \widehat{R} + \hat{\rho}_{\infty}^{\mathrm{G}}. \end{aligned} \tag{18}$$

最终检查点比较省略了数据项，因为它可以被吸收到常数中。残差 R_i使用近似修正来估计总体池损失，而不是经过评估的无限集成。它既包含共享数据和优化局限，也包含容量损失。

两种损失拟合都在对数残差上最小化 Huber惩罚，转折点为 0.01；离散度拟合使用双对数最小二乘。每个设置的权重相同。我们将方法A和方法B中的总和与原始成员损失进行比较，并使用原始单位下的均方根误差进行评分。方法B额外拟合了两个离散度系数。配对比较使用相同的设置、下限和留出规模。交叉验证省略整个字符宽度、Pythia规模或 Qwen 对，并重新拟合两个分量；它绝不会将被省略设置的实测离散度加入其预测中。表20给出了各下限，并包含一个带有 $0 \le E \le 1.5$ 的自由拟合字符下限控制项。

评估总体。 字符研究使用表 13 中的损失定律设置：六种宽度，每种宽度对应从每参数 6.25 到 400 字节的 13 个训练预算，并在 524,288 个留出字节上进行评估。除宽度为 192 时使用随机种子 0–1 外，其余均保留随机种子 0–2。其扩散定律基于这些相同的运行结果和训练预算进行拟合。

Pythia 在 96 个 C4片段上使用十个最终检查点，采用标准的 1,024 个词元上下文和 64 个词元预热。五种较小规模使用三个真实随机种子；较大规模使用 main/v0重跑对。此处 N

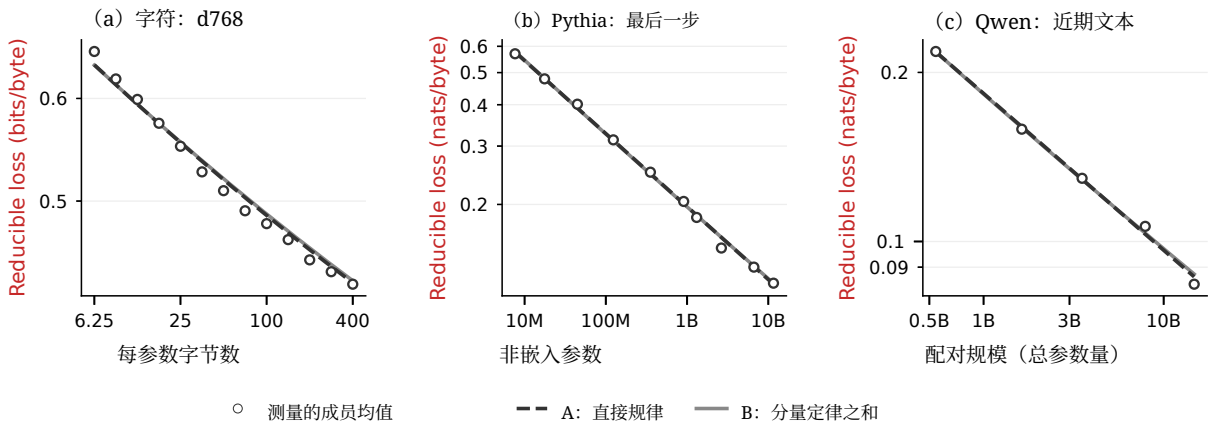


图17 直接规律和分量定律给出的成员损失预测相近。 (a)：隐藏宽度为 $d = 768$ （标记为 d768）的字符级模型在不同训练预算下的表现。(b)：Pythia 的十个最终检查点，包括规模更大的重新运行配对组。(c)：近期文本上的五组稠密 Qwen 配对模型。各点表示测得的成员平均损失减去所述不可约损失。曲线预测同一量：A 直接预测，B 则将独立拟合的汇总对数损失与离散度相加。对数坐标轴显示了相对于简单幂律的偏差；表20在未变换的成员损失上评估误差，并加入留出规模测试。

表20 直接拟合与分量拟合具有相近的预测误差，没有持续占优的一方。RMSE 在 A 和 B 使用的同一原始成员平均损失上进行评估。B 使用几何离散度和仅依赖规模的规律。留出误差汇总所有被省略的单元格；两个分量均在不包含留出规模的情况下重新拟合。字符模型使用 78 种 FineWeb-Edu 设置，损失以比特/字节计。采用自由估计的不可约损失的拟合结果共享边界 $0 \le E \le 1.5$。公共模型误差以纳特/字节计。

队列	E	单元格	A：拟合		B：拟合	A：留出	B：留出
字符模型：比特/字节，留出六种宽度中的一种							
字符，N, D	0.593341		78	0.0193	0.0198	0.0290	0.0294
字符，N, D	0.637650		78	0.0200	0.0203	0.0303	0.0310
字符，N, D	free		78	0.0191	0.0197	0.0304	0.0301
Pythia：纳特/字节，留出十种规模中的一种							
最终结果，仅大小	0.436514	10		0.0056	0.0061	0.0075	0.0085
Qwen2.5/3：纳特/字节，留出一对规模							
C4，稠密	0.436514		5	0.0028	0.0035	0.0047	0.0052
最近的，稠密	0.333728		5	0.0020	0.0022	0.0032	0.0032

不包括嵌入。中间检查点没有单独的冷却阶段，且无法很好地确定数据依赖性，因此最终步长比较是最清晰的检验。

Qwen 分别在 C4 和近期文本上使用五个稠密 Qwen2.5/Qwen3 配对，并采用成员总数的几何平均数。这些二元组混合了不同代际。其校正后的离散度仍然是一项描述性估计，并且每个语料库都使用各自声明的下限。

这两种描述给出了相似的有限范围成员损失预测，在留出误差方面没有一致的优势。即使总和预测几乎不变，移除缓慢下降的离散度也会改变拟合的规模指数。这促使我们单独测量离散度，同时仍不确定不可见部分的形式。特别是，总损失和离散度的两个幂律并不能强制它们的差值也服从幂律；如果该差值最终会变为负值，那么它们不相等的指数也不可能无限期成立。

F 上下文依赖的不确定性

F.1 目标与分数

图6探讨了某个分数能否识别模型自身的认知误差。由于真实的下一个词元分布未知，我们使用更强的参考分布 r_x 并测量

$$e_N(x) = \frac{1}{S}\sum_{s=1}^{S} \mathrm{KL}(r_x \| q_{N,s,x}). \tag{19}$$

如果结果遵循参考分布，这是成员的平均可约损失。分数是单个成员的熵 $h_s(x) = H(q_{N,\,s,\,x})$ 和几何离散度 $g(x) = {}_{S-1}^{S}[-\log Z_{N,\,x}]$，使用第 2.2 节中的校正。二者都只取决于给定上下文中的预测。算术离散度还需要下一个词元或目标分布，因此在此处不作为预测器使用。熵使用字符种子 0、默认的 Pythia 种子（160M 时为种子 1）、DataDecide 的默认种子、Qwen2.5，或 Delphi 种子 0。隐藏状态探针使用同一个成员。

F.2 模型、文本与参考分布

表 21 给出了图 6 和图 18 的评估总体。字符学生模型使用表 13 中的损失定律设置，并以字节数与解析参数之比相同的方式进行训练；因此，模型规模和训练数据同步增长。Pythia 学生模型则共享固定的训练预算，其成员关系见附录 D.1。随着学生模型逐渐接近参考模型，参考质量变得更加不确定，因此 DataDecide 略去了可用的最大大小。Qwen 使用具有几何平均参数数量的 densebase 模型对。每个族都使用表 21 中列出的参考分布和捕获的上下文总体。

表 21 上下文依赖的不确定性数据。 规模按总参数量计算。每个窗口在至少包含 64 个上下文字节或词元后，贡献 32 个采样目标。字符模型使用 512 字节的 FineWeb-Edu 窗口；公开模型最多使用 1,024 个近期文本词元。字符学生模型每个解析参数使用 50 个训练字节；Pythia 使用 300B 个训练词元。

学生模型	S	参考分布	窗口	训练/测试上下文
字符，0.86–50.7M	3	BLT-7B	256	4,096/4,096
Pythia，14–410M	3	MPT-30B	96	1,536/1,280
DataDecide C4，6.96–597.7M	3	OLMo-7B	96	1,536/1,280
Qwen，0.5/0.6–14/14B	2	Qwen2.5-72B	96	1,536/1,280
Delphi，447M–9.7B	3	Delphi 25B	96	1,536/1,280

图6中的字符-模型列使用留出的 FineWeb-Edu 文本，其中交替的来源区域定义了拟合部分和测试部分。Pythia、DataDecide、Qwen 和 Delphi 使用附录 B.3 中所述的近期网页文本；公开模型协议交替使用由四个完整窗口或文档组成的区块。MPT-30B 此前是根据前16个近期文本窗口上的下一词元损失选出的，因此在所有近期文本评估中，我们都将这些窗口从评分中省略。48个拟合窗口保持不变，而测试集包含其余40个窗口，即每个族1,280个上下文。在每个族和语料库内，图6中的比例和相关性，以及图18中的预测评分，均使用相同的测试上下文，并在不同模型规模之间保持固定。

图6的(f,l)面板复用了(b–e,h–k)列中的公开模型族近期文本评估。Pythia、DataDe-cide、Qwen 和 Delphi 使用相同的40个评分窗口或文档，不包括参考分布选择试验。每个族对每个窗口中的32个token位置进行评分，共计1,280个留出上下文。token位置和原始前缀取决于分词器；在每个族内，它们在不同模型规模和面板间保持固定。

参考分布和输出词汇表因族而异，因此这些是描述性比较，而非跨族汇总的缩放拟合。

在拟合不确定性预测器之前，我们固定参考分布；报告中的测试上下文不会用于选择参考分布。基于 BLT-7B 在留出集 FineWeb-Edu 上的损失低于 EvaByte，我们为字符模型选择了 BLT-7B（Pagnoni 等，2024；EvaByte，2025）。在预留的近期文本试点中，MPT-30B 的表现优于 GPT-NeoX-20B 和 DCLM-7B（Databricks AI 研究团队，2023；Li 等，2024）。原始 OLMo-7B 仍作为 DataDecide参考（Groeneveld 等，2024）。
在计算 KL 散度之前，我们先对齐输出词汇表。BLT 的预测以普通字节输出为条件；公开模型的预测保留已分配的词元 ID，并排除未分配的填充。Qwen 保留共享的特殊词元。公开参考模型使用 FP32 推理和 FP64 归约；原生前向检查用于验证归一化概率，BLT 也通过了因果注意力检查。条件设定很重要，因为字节模型无法预测 BLT 的特殊词元。

F.3 相关性与预测

图 6(g–l)中的斯皮尔曼相关系数在每个大小下分别计算，不拟合预测器，并使用表 21中的测试上下文。图18中的可见比例面板和预测面板在每个族内使用相同的上下文。某个分数可以很好地对上下文进行排序，却给出较差的数值估计，因此我们将预测与排序分开评估。

F.3.1 拟合与评估

每个标量分数都使用仿射最小二乘拟合。隐藏状态探针改编了 Ahdritz等（2024）的回归头流程：使用 Adam 训练仿射头，以预测参考 KL 散度。验证会在中间隐藏层和最终隐藏层之间进行选择，并确定停止轮次。最终状态取自输出归一化之后；中间状态来自模型的 h 个 Transformer 块中的 $\lceil h/2 \rceil$。表 22给出了拟合设置。

表 22 隐藏状态探针设置。 文本组在验证过程中保持完整。探针保留一个训练折用于验证，之后不再重新拟合。测试目标不选择层、停止轮次或系数。

设置	值
Loss	均方误差
Adam 学习率	10^{-5}
批次大小	64
训练折 / 验证折	3 / 1
耐心值 / 训练轮次上限	100 / 100,000
初始化种子	42
字符拟合 / 验证上下文	3,072 / 1,024
公开模型拟合 / 验证上下文	1,152 / 384

我们报告 $R^2 = 1 - \text{均方误差} \ / \ \text{方差}(e_N)$，两项均在测试上下文上进行评估 s.
图18显示，在更大的公开模型规模下，离散度比熵更有效。在最大的 Pythia、DataDecide、Qwen 和 Delphi 设置下，离散度分别得到 $R^2 = 0.351$、0.273、0.536 和 0.347；相比之下，熵的结果分别为 0.265、0.127、0.096 和 0.116，而探针的结果为 0.100、 -0.083、 -1.349 和 -1.591。在每个实测规模下，字符模型熵仍比离散度更准确。负的探针分数意味着其平方误差比预测测试集均值更大。这些比较均以固定的监督预算和拟合过程为条件。

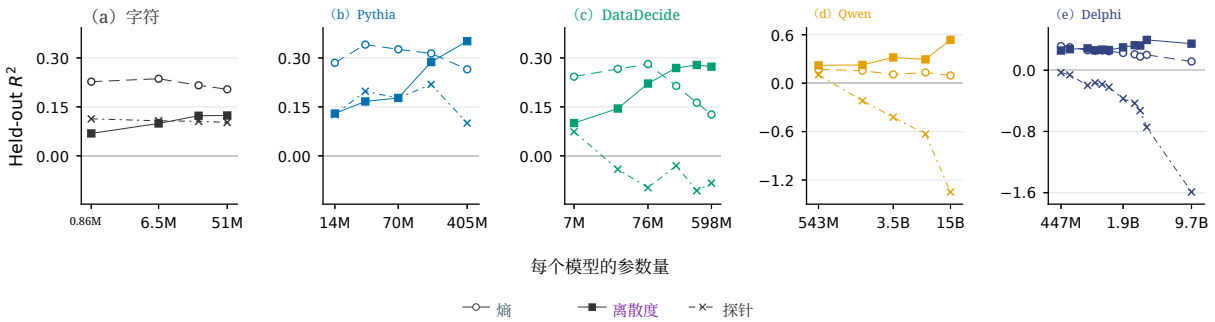


图18 预测上下文依赖的认知不确定性。 留出 R^2 用于从单模型 熵、校正后的几何平均 离散度或隐藏状态线性探针预测基于参考分布的可约损失。熵和离散度使用仿射最小二乘拟合；探针采用 Adam 进行训练，遵循改编自 Ahdritz 等人（2024 年）的协议。（a）在 4,096 个 FineWeb-Edu 测试上下文上的字符模型。（b–e）Pythia、DataDecide、Qwen 和 Delphi 族，每个族使用 1,280 个近期文本测试上下文。每个族使用与图6相同的测试上下文，并在不同规模间保持固定。保留负分数；各面板的纵轴刻度不同。

F.4 跨上下文的恢复比例

图6(a–f)使用与相应相关性面板相同的留出上下文、模型组和固定参考分布。单个族列对字符模型使用4,096个FineWeb-Edu上下文，对每个公共族使用1,280个近期文本上下文。最右列对每个公共族重新使用与单个族列相同的1,280个近期文本测试上下文。每个窗口贡献32个位置。所有测试上下文都纳入这些描述性比例；不拟合任何预测器，也不因分母较小但为正而排除任何上下文。

令 $g_N(x) = \frac{S}{S-1}[-\text{对数} Z_{N,x}]$ 表示图18中使用的校正后的几何离散度。对于 $e_N(x) > 0$，校正恢复比例和有限集成恒等式为

$$f_N(x) = \frac{g_N(x)}{e_N(x)}, \qquad e_N(x) = \frac{S-1}{S} g_N(x) + \mathrm{KL}(r_x \| q^{\mathrm{G}}_{S,N,x}). \tag{20}$$

实线曲线表示平均校正离散度与基于参考分布的可约损失之比：

$$F_N = \frac{\mathbb{E}_X g_N(X)}{\mathbb{E}_X e_N(X)} = \frac{\mathbb{E}_X[e_N(X) f_N(X)]}{\mathbb{E}_X e_N(X)}. \tag{21}$$

因此，总体是按可约损失加权的逐上下文比例平均值。如果参考分布等于真实条件分布，且上下文、成员和损失单位与损失评估相匹配，则 $\mathbb{E}_X e_N(X) = L_S - E$，因此该比值在真实不可约损失处与公式（14）中的校正总体比例一致。

浅色带表示来自2,000次完整的32位置测试窗口自助法重采样的95%百分位区间；对于每个族，在不同规模之间使用相同的重采样窗口。每次重采样都会重新计算平均离散度与平均可约损失之比。这些区间以所选成员和参考分布为条件。图6的面板（f）重复展示了面板（b–e）中的公共族比值和区间。有限成员校正将三成员集成的离散度乘以1.5，将 Qwen 配对模型的离散度乘以2，并适用于所有绘制的分数和区间。校正后的分数不一定低于1。

G 集成与更大型模型

G.1 交叉点主张的证明

在固定 ND的情况下，最小化 $L(N,\ D)$可得 $\alpha aN^{-\alpha} = \beta bD^{-\beta}$。当 $\alpha = \beta$时，规模项与数据项相等：$bD^{-\alpha} = aN^{-\alpha}$。固定 $S_2 > S_1 \geq 1$，令 $r = S_2/S_1$和 $u_i = 1 - S_i^{-1}$，并令（$N,\ D$）为该 S_2成员集成的计算最优成员分配。

在第3.4节的情景（1）中，每个 S_1 成员使用（$rN,\ D$），由于 $S_1 rN = S_2 N$，因此参数总量和训练计算量相匹配。在情景（2）中，每个成员使用（$\sqrt{r}N,\ \sqrt{r}D$），保持计算最优性，并且由于 S_1（$\sqrt{r}N$）（$\sqrt{r}D$）$= S_2 ND$，使总训练计算量相匹配。因此，相对于 S_2 集成的成员，成员平均损失的降低为

$$G_1 = a(1 - r^{-\alpha})N^{-\alpha}, \qquad G_2 = 2a(1 - r^{-\alpha/2})N^{-\alpha}.$$

S_2 集成相对于 S_1 集成的额外离散度为

$$H_1 = c_P(u_2 - u_1 r^{-\gamma_P})N^{-\gamma_P}, \qquad H_2 = c_P(u_2 - u_1 r^{-\gamma_P/2})N^{-\gamma_P}.$$

因此，在情景 j中，S_2 集成的损失减去 S_1 集成的损失为 $G_j - H_j$。$N^{-\gamma_P}$的两个系数均为正，因为 $u_2 > u_1 \geq 0$和 $r > 1$。由于 $\alpha > \gamma_P$，不等式 $H_j > G_j$恰好在 $N > N_j^\star$时成立，其中有限的正阈值见公式（10）。最后，

$$\left(\frac{N_2^\star}{N_1^\star}\right)^{\alpha-\gamma_P} = \frac{2}{1 + r^{-\alpha/2}} \frac{u_2 - u_1 r^{-\gamma_P}}{u_2 - u_1 r^{-\gamma_P/2}} > 1.$$

第一个因子大于 1，第二个因子至少为 1，因此 $N_2^\star > N_1^\star$。令 $S_1 = 1$可得 $u_1 = 0$和$r = S_2$，从而恢复与单模型的比较。因此，在所假设的规律下，任何固定的较大集成规模最终都会胜过较小规模，且在情景（2）中交叉点会延后。这一排序要求采用计算最优的平衡；仅使数据量随规模成比例增长是不够的。 □

G.2 图 7 的数据与拟合7

我们评估实际的有限几何池，不对其损失应用总体校正。Pythia 包含表8中保留的独立运行的所有二元组和三元组。字符模型和 DataDecide C4 在每个显示的规模下使用所有二元组和该三元组。DataDecide 保留完整的七规模单例拟合，其中包括显示范围之外的三个较小规模。Qwen 在每个规模下都有一个代际对，并为每一代分别设置一个单例参考。横坐标将成员的实际参数数量相加。字符损失使用 FineWeb-Edu，而公开模型损失使用近期文本，因此绝对损失应在各自的族内进行比较。

递增数据的字符轨迹采用每个解析参数100个训练字节数。对于其固定数据面板，表23给出了共享预算和损失。0.858M和6.461M模型均使用了三个延续，以达到精确的共享预算，并在与更大型模型相同的留出字节上进行评估。其中一个中间大小需要在记录检查点之间对数插值 D；不使用数据预算外推。我们保留了偏离平滑趋势的情况，包括Pythia的70M模型以及参数量接近6.5M的字符模型。

单模型基准。 上方曲线拟合 $H_S(B) = E_S + A_S(B/10^9)^{-\eta_S}$，其中 B 是总参数数量。我们对每种规模等权，并在对数超额损失上采用最小二乘法。固定数据增益将实测集成与 $H_1(B)$进行比较。对于递增数据面板，参考分布则拟合单模型损失与训练成本之间的关系，即 $J_1(C) = E + A(C/C_0)^{-\eta}$，并在集成的精确成本 $C = \sum_s N_s D_s$处对其进行评估。表 24 给出了数值设置。这些成本参考分布的参数数量通常与集成不同，因此下方标记并不是上方曲线之间的垂直间隔。

成员 N (M) 来源	L_1 ±标准差	L_2 ±标准差	L_3
0.858 实测	1.146265 ± 0.001574	1.120932 ± 0.000701	1.112706
6.461 实测	0.941672 ± 0.000186	0.923487 ± 0.000264	0.917532
21.441 实测	0.859157 ± 0.000824	0.843909 ± 0.000443	0.838903
50.678 插值	0.812848 ± 0.000357	0.799516 ± 0.000199	0.795133
170.303 实测	0.764301 ± 0.001086	0.752760 ± 0.000440	0.748968

表23 固定训练数据下的字符集成。 每个成员处理 $D_0 = 8$、493、465、600 字节。损失以纳特/字节表示，并给出成员或二元组之间的样本标准差；完整三元组没有子集选择标准差。50.7M 的损失在 对数 D 尺度上，在 7、117、996、032 和 10、066、329、600 字节之间进行插值。重叠二元组共享模型，因此这些标准差反映的是子集选择，而不是独立重复实验。

表 24 集成预算比较的参考设置。 不可约损失设置以纳特/字节表示。固定数据的字符设置来自相应的规模–数据拟合。 A 和 η 分别针对每条上包络曲线进行拟合；成本参考系数已明确列出。字符成本使用 $C_0 = 10^{18}$ 参数字节；DataDecide 使用 10^{19} 参数词元。

曲线	E	A	η
公开模型上包络曲线	0.33372841	—	—
字符递增数据上曲线	0.41127265	—	—
字符固定数据 , $S = 1$	0.65696935	—	—
字符固定数据 , $S = 2$	0.65553698	—	—
字符固定数据 , $S = 3$	0.65496251	—	—
字符单例成本参考	0.41127265	0.366901	0.092751
DataDecide 单例成本参考	0.33372841	0.240116	0.128495

预测的集成增益。 下方曲线使用独立于上方集成损失拟合的扩散定律。我们将每个有限扩散量除以 $1 - S^{-1}$，并在对数空间中对二元组和三元组联合拟合共同定律 $c(x/x_0)^{-\gamma}$，对每种规模和集成数量等权。结合相应的单例定律，可得

$$\widehat{\Delta}_S(x) = (1 - S^{-1})c(x/x_0)^{-\gamma} - A(1 - S^{-\eta})(x/x_0)^{-\eta}.$$

此处， x 表示固定数据下的成员规模，或递增数据面板中的成员训练成本。第一项是汇聚增益；第二项是将相同预算用于一个更大模型所带来的改进。通过将成员规模映射到成本的拟合关系，可将后者的曲线置于总参数轴上；标记保留精确记录的成本。

下限敏感性。 尽管在拟合损失之差中， E 会被显式抵消，但改变它会改变重新拟合的振幅和指数。图7传播了附录 E.1 和 B.7.3 中保留的公开下限范围 $[0.26019, 0.33372841]$纳特/字节，以及单独的字符范围 $[0.593341, 0.63764990]$ 比特/字节。对于每个候选下限，我们在各自原始队列上、使用相同目标重新拟合规模和成本定律。对于固定数据字符，我们首先在原始的 65 个单元格上重新拟合每个集成规模的规模–数据定律，然后在共享预算 D_0 下重新计算其有限数据偏移 $E_{D,S} = E + b_S D_0^{-\beta_S}$，再重新拟合五个规模中心。中央曲线保留表 24 中的设置。

阴影范围涵盖逐点最小值和最大值的重新拟合曲线。对于每个预测间隔，我们在取其包络之前，先在每个下限处计算上述完整表达式，同时固定独立拟合的离散度定律和分配映射。垂直区间则改变拟合的单例参考，同时保持每个实测集成损失不变，并对递增数据面板使用准确记录的 ND。上限损失标记与下限无关。公共范围来自名义 95% 的条件下限剖面，而字符范围是从参考值到上限的敏感性范围；这些传播得到的包络并非完整的

置信区间。它们省略了固定下限处的残差拟合不确定性、随机种子和文本采样、插值以及模型形式不确定性。
这些是经验性的预算比较。DataDecide 的数据分配是为过度训练而选择的，而 Qwen 各代在数据和训练配方上有所不同（Magnusson 等人，2025年；Qwen 团队，2024年；Yang 等人，2025年）。这些比较无法确定理论交叉规模，也无法确立其排序。虚线部分表示外推，而最大的固定数据字符集成所需的单例参考超出了实测规模范围。每种候选下限下的五种规模固定数据字符拟合各有三个残差自由度，且以对应的有限数据偏移为条件。参数匹配或 ND 也不能证明运行时间或延迟相同。

数值交叉示例。 不可约损失的不确定性会改变重新拟合得到的损失幅度和指数，从而影响公式（10）及其经验对应形式在匹配 ND 时得出的交叉预测。在保留拟合扩散定律的同时改变 $E \in [0.26019, 0.33373]$ 纳特/字节时，固定 D、并在匹配 ND 的情况下，每个 Pythia 成员约为 $[2.9 \times 10^{12}, 1.6 \times 10^{15}]$ 个参数、每个 DataDecide C4 成员为 $[2.0 \times 10^{10}, 7.5 \times 10^{11}]$ 时，会出现成对集成与单例的交叉（$S_1 = 1$， $S_2 = 2$）。若三模型集成要优于成对集成（$S_1 = 2$， $S_2 = 3$），相应的交叉阈值分别约为每个 Pythia 三元组成员 $[1.7 \times 10^{14}, 9.8 \times 10^{17}]$ 个参数，以及每个 DataDecide C4 三元组成员 $[1.3 \times 10^{12}, 6.2 \times 10^{14}]$ 个参数。这些是阈值范围；高于该阈值时，规模较大的集成具有更低的预测损失，其中 N 始终表示较大集成中每个成员的参数量。DataDecide 的范围依据其拟合得到的训练分配，但不能据此证明计算最优性。所有范围均描述条件下限敏感性，且未计入其他拟合和外推不确定性。

G.3 两种预算约束下的实测比较

我们将同一对预先选定的 DataDecide C4对与两个更大的检查点进行比较：一个使用相同的数据预算训练，另一个在总 ND几乎相同的情况下使用更多数据。表 25给出了分配和损失。单独来看，这两个成员都无法击败固定数据参考分布。池化带来了这一优势，而允许更大的模型使用更多数据，则会使优势缩小为尚未解决的近似平局。这说明了情景（2）中更强的参考分布，但并未确定规模交叉点。

该对模型采用名义 300M 默认配置，以及步数 45,000 时的 small-aux-2 运行。参考模型分别为步数 25,000 时的名义 750M 默认配置和步数 40,000 时的 600M 默认配置。检查点和随机种子在评估前已固定。所有模型均使用附录 B.3中的通用近期文本评估、float32 推理、检查点 RMS 归一化 epsilon 10^{-6}以及完整的输出词表。配对区间对 24 个连续四文档组进行重采样，共 5,000 次，并重新计算损失总和除以字节总和。它们量化了在这些模型条件下的文本变化。

约束	参考值（M）	D （B）	成本比	单个	增益	95%区间
固定数据	758.6	29.4912	1.02097	0.565021	+5.616	$[+3.502, +7.656]$
匹配成本	597.7	36.7002	1.00108	0.559427	+0.023	$[-1.198, +1.392]$

表 25 相同 C4 对在两种预算约束下的表现。 每个成员均有 371.492M 个参数；其几何池损失为 0.559404 纳特/字节。 D 表示以十亿计的参考训练词元。成本比为参考 ND 除以成对模型的词元总数 ND。单个损失的单位为纳特/字节；增益及其对应的 95%区间单位为毫纳特/字节。正增益表示该对模型表现更优。

这些模型处于其已发布学习率调度表的不同阶段，且参考分布并未针对这些预算单独进行退火。因此，该比较针对的是可用的检查点，而不是针对每个约束分别进行优化的模型。

G.4 专业领域比较

图19在专业数据集上重复了预算比较，这些数据集可能并不严格属于分布内。我们看到，在这些设置下，在两种不同的匹配预算中，许多集成都能够优于单模型，而此时离散度通常更高；有关这些领域中离散度如何变化，请参见图 11。

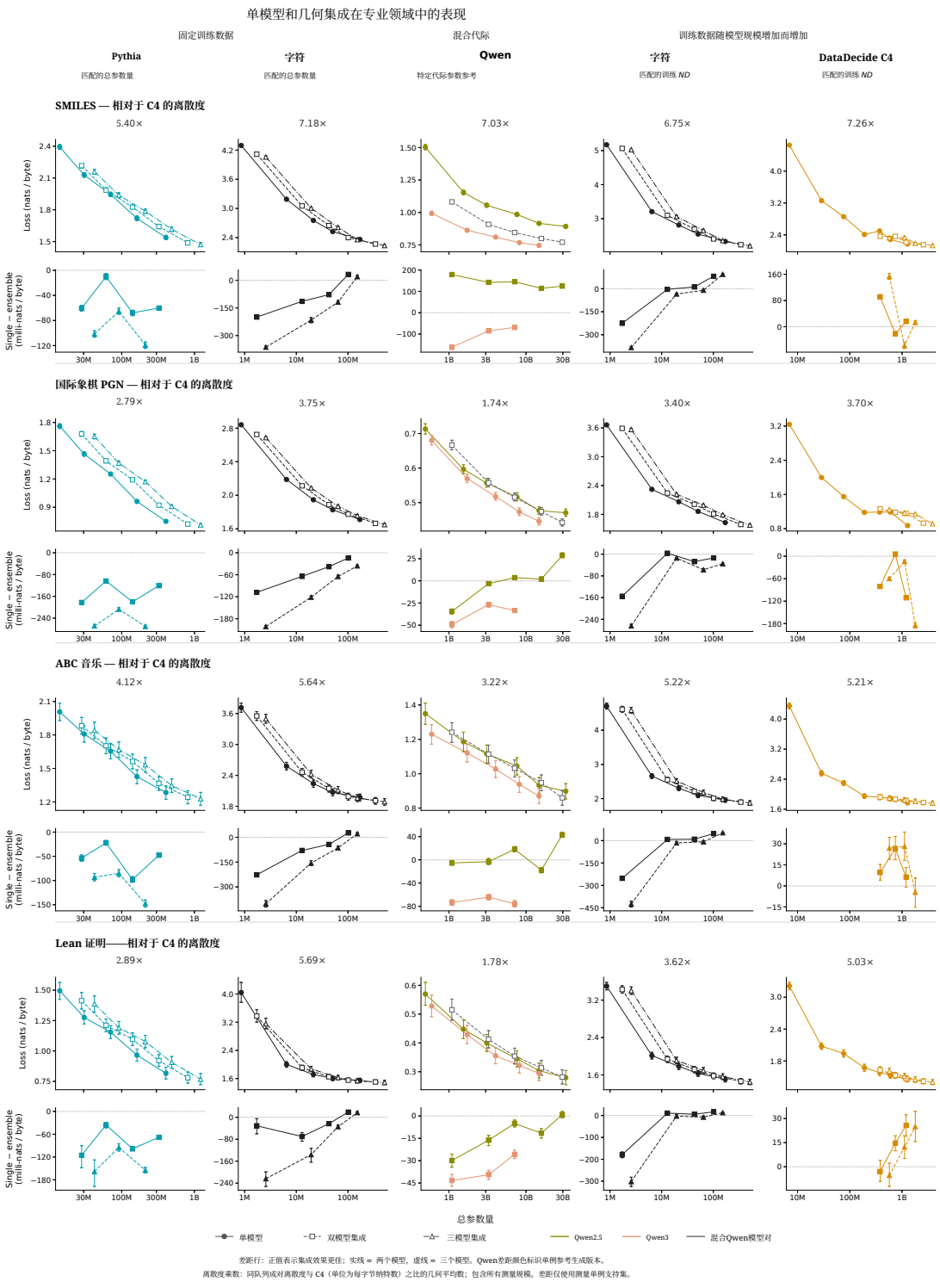


图19 专业领域中的单模型与几何集成。 每个数据集占据两行：上方为实测损失，下方为匹配预算的单模型损失减去集成损失（正值表示集成更优）。单例参考使用支持集内线性插值，插值在 对数 N 或 对数(ND)空间中进行，不设拟合损失下限；柱状图对应成对的段落自助法 95% 区间。乘数是在匹配模型规模下，以相对于 C4 的每字节成对离散度比率计算的几何平均数。Qwen 参考分布特定于生成版本，仅匹配参数数量；训练预算和训练配方有所不同。

H 共享训练和 LoRA 集成研究

H.1 匹配数据和训练状态

我们使用表 13 中的离散度和分支设置。表 26 列出了规模。令 $t = D/N$ 为评估预算，令 b 为分支点，二者均以每参数字节数表示。每个父模型、其延续模型及其独立参考模型之间的学习率、批次大小和冷却调度均保持匹配。

经过 18 步预热后，每个父模型沿恒定学习率轨迹训练至每参数 160 字节。衰减延续会追加此前轨迹的四分之一，并在 $T = 200$ 处结束。第 3.5 节中的所有分支点都发生在预热之后、衰减之前。分支会恢复保存的权重和优化器矩，不进行重置或额外预热。它会以分块方式排列剩余数据流，确保每个评估终点所消耗的语料位置保持不变。因此，两个成员看到的是相同的数据，只是在分支后采用了不同的顺序。

父模型使用初始化种子 0 和顺序种子 1；独立参考模型使用初始化种子 1 和顺序种子 2；分支使用顺序种子 7。父模型和参考模型从一开始就使用经过排列的数据流。图 8 中仅初始化的比较则是在共同的顺序数据流上，对原始随机种子的成对离散度取平均；它不是 R 的分母。

H.2 评估与拟合定律

表 26 和表 27 以及图 22 中的检查使用了 1,024 个留出的 FineWeb-Edu 序列，每个序列包含 512 字节。图 8的面板 (a,b) 在 432,878 个近期文本字节上评估相同的训练检查点。两项评估均显示$2\rho_2^{\mathrm{G}}$使用成对校正。该因子在 R 中相消；相关分支不会因重新缩放而变成独立的总体样本。表 26 保留了最终预算下的每个宽度。

表 26 在每参数 200 字节下保留的成对离散度。 评估使用留出的 FineWeb-Edu 文本。分支列给出 R 第 3.5 节中的结果。独立列报告 $2\rho_{2,\ \mathrm{ind}}^{\mathrm{G}}$ 单位为毫比特/字节。

宽度	N (M)	独立	$b = 1$	$b = 5$	$b = 25$	$b = 125$
128	0.79	76.72	1.043	0.971	0.722	0.281
192	2.65	60.65	0.899	0.837	0.659	0.247
256	6.29	51.21	0.949	0.885	0.700	0.284
384	21.23	43.90	0.944	0.866	0.687	0.307
512	50.33	40.04	0.926	0.846	0.678	0.318
768	169.87	34.12	0.905	0.830	0.688	0.383

我们在相同的五个宽度上（不包括最小宽度），于对数空间中使用最小二乘法拟合规模定律。最早的分支离散度大于独立成对样本，因此其损失份额无法纳入对数拟合。损失份额定律使用保留宽度上的所有分支点。表27列出了拟合结果。较早的分支大致保持了独立规模斜率；最新的分支斜率则较平缓。

平方根近似能够描述实测范围。但在完全共享时，该近似不成立，因为相同的最终状态必须具有零离散度；而共享训练为零的情况尚未经过测试。对中间预算的检查也表明，分别依赖分支点和评估预算优于仅依赖二者的比值。每个宽度对应一个父模型和一个独立参考分布；各分支比值共用同一个分母。因此，拟合所用的观测值并非独立的训练重复实验。

共享比例 $f = b/T$ 可为一对模型节省 $f/2$ 的训练工作，不包括用于中间测量的额外冷却过程。在拟合范围内，将分支对扩散除以拟合的保留份额，可用于估算独立对扩散。此校正尚未单独验证，而且保留份额较小时会放大其误差。

表27 FineWeb-Edu 上最终预算的分支拟合。 规模拟合使用 $2\rho_2^{\mathrm{G}} = cN^{-\gamma}$，覆盖宽度 192–768。不确定性是回归残差标准误，而非来自重复训练实验之间的变化。损失份额拟合使用了 20 个观测值，自然对数均方根误差为 0.160。

规模定律总体	γ	SE
独立初始化和顺序	0.134	0.008
仅初始化	0.127	0.005
在 $b = 1$ 处分支	0.135	0.002
在 $b = 5$ 处分支	0.140	0.001
分支点位于 $b = 25$	0.129	0.003
分支点位于 $b = 125$	0.039	0.010
损失份额拟合： $1 - R = 0.846(b/T)^{0.467\pm0.021}$		

H.3 最终权重相关性

我们比较每个父模型与其四个全参数分支在最终检查点处的情况，所有六种宽度均为每个解析参数 $T = 200$ 字节。参数按注册名称和形状对齐，并将所有唯一的已学习标量坐标串接成向量 w_{P}, $w_{\mathrm{B}} \in \mathbb{R}^M$。这里的 M 是已学习参数总数，包括嵌入、输出权重和归一化增益；解析计数 N 仍用于定义训练预算和图例。我们省略非参数缓冲区，并计算

$$r_w = \frac{\sum_{i=1}^{M}(w_{\mathrm{P},i} - \bar{w}_{\mathrm{P}})(w_{\mathrm{B},i} - \bar{w}_{\mathrm{B}})}{\sqrt{\sum_{i=1}^{M}(w_{\mathrm{P},i} - \bar{w}_{\mathrm{P}})^2}\sqrt{\sum_{i=1}^{M}(w_{\mathrm{B},i} - \bar{w}_{\mathrm{B}})^2}}, \tag{22}$$

其中，每个均值均取自完整拼接向量。因此，每个标量坐标都会参与计算；我们不会对分别针对单个张量计算的相关性取平均。归约使用 float64 算术，不进行坐标采样，也不执行神经元或注意力头置换对齐。

图20显示了 r_w 与共享训练比例及该对模型的预测扩散之间的关系。扩散度面板使用与图8相同的 432，878 个经过评分的近期文本字节，并以毫比特/字节显示 $2\rho_2^{\mathrm{G}}$。独立参照会同时改变初始化和数据顺序；它们没有分支点，仅为显示目的被置于单独的“独立”位置。它们的坐标相关性取决于所选参数化，分支的坐标相关性也是如此。

高方差参数组可能主导这种坐标相关性。作为敏感性检查，移除嵌入、输出头和一维归一化增益后，在 $b = 1$ 处的相关性为 0.147–0.216，在 $b = 125$处为 0.918–0.964；上升趋势依然存在。纳入所有参数时，对应范围分别为 0.365–0.575 和 0.937–0.974。这些是描述性比较，每个分支点和宽度仅对应一个父模型和一次延续。四个分支共享同一个父模型，这些结果并未估计多次训练运行之间的变化。

H.4 多样性惩罚与集成损失

不同模型规模下的结果。 在 $\lambda = 1$处，多样性惩罚使几何离散度增加 2.30–2.72 倍，而较高的成员平均损失抵消了新增离散度的 93–106%。因此，总集成损失的额外增益大多被抵消（表29；图21）。

我们测试显式奖励分歧能否恢复共享训练过程中损失的离散度。首先，我们在宽度为192的字节模型中扫描惩罚权重（6层，2.76M总参数量），从$b = 125$ 开始分支，最终达到每个解析参数 $T = 200$ 字节。随后，我们固定 $\lambda = 1$，并将比较扩展到宽度为256、384和512的模型，涵盖2.76–50.68M总参数量。我们保留每个检查点的优化器

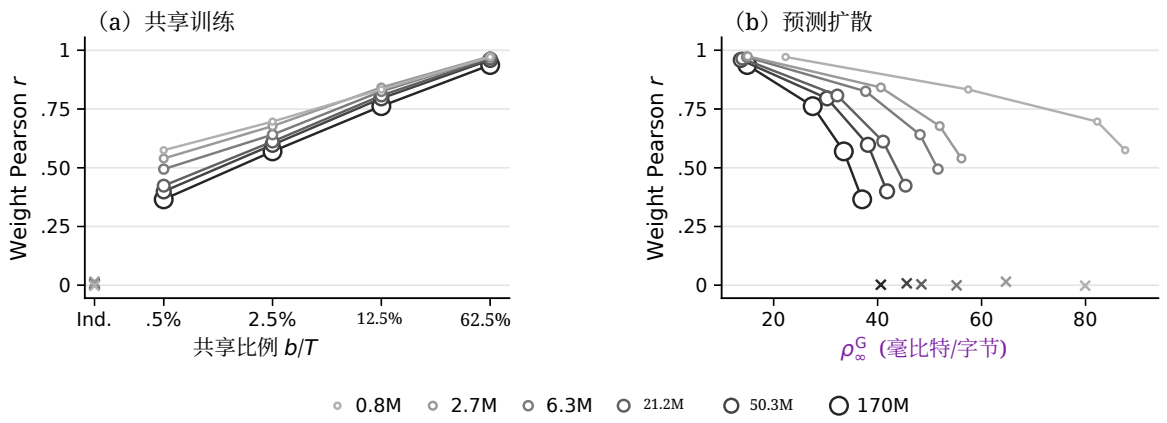


图20 后续分支保留了更为相似的最终权重。 在 $T = 200$处，完整参数向量的皮尔逊相关系数分别与 (a) 共享训练比例 b/T和 (b) 近期文本上经校正的几何配对扩散的关系。空心圆表示父模型–分支模型对；叉号表示独立初始化和数据顺序参考，单独显示在“Ind.”处，未分配 b/T。四个实测分支点由线连接。灰度和标记大小表示解析参数数量 N。

矩以及上述所述的批次大小、学习率、冷却阶段和数据顺序。父模型继续进行常规交叉熵训练；只有分支模型获得多样性奖励，该奖励基于其训练上下文，并针对父模型当前的预测进行评估：

$$\mathcal{J}_\lambda = \widehat{L}(q_{\mathrm{B}}) - 2\lambda\widehat{\rho}_2^{\mathrm{G}}, \qquad \widehat{\rho}_2^{\mathrm{G}} = -\widehat{\mathbb{E}}_X \log \sum_{y\in\mathcal{V}} \sqrt{q_{\mathrm{P},X}(y)q_{\mathrm{B},X}(y)}. \tag{23}$$

这里， q_{P} 和 q_{B} 表示父模型预测和分支模型预测，帽子表示训练小批次平均值。对该目标函数求导时，父模型的预测保持不变，而其自身更新不受正则化约束。负的几何扩散项会激励完整词汇表上的预测多样性。该项的系数为 2，使得在父模型保持不变时，分支梯度等于 $2[(1-\lambda)\widehat{L}_2 + \lambda\widehat{L}(q_2^{\mathrm{G}})]$, 的梯度，从而得到 Webb 等人（2021）联合训练目标的仅分支版本。因此， $\lambda = 1$直接针对分支最小化几何池损失。

Pair	$\bar{L}_2$	ρ_2^{G}	$L(q_2^{\mathrm{G}})$	$L(q_2^{\mathrm{A}})$
普通晚期分支 ($\lambda = 0$)	1.542097	0.007462	1.534636	1.534759
$\lambda = 0.25$	1.543177	0.008937	1.534241	1.534463
$\lambda = 0.50$	1.544696	0.011007	1.533689	1.534087
$\lambda = 1.00$	1.554386	0.020424	1.533962	1.536031
独立随机种子	1.542116	0.030323	1.511793	1.514174

表28 多样性奖励会增加离散度和成员平均损失。 所有数值均为比特/字节，基于相同的 1,024个留出序列，每个序列包含 512 字节。损失为普通交叉熵，不包含训练奖励。离散度是原始成对量 ρ_2^{G}，因此 $L(q_2^{\mathrm{G}}) = \bar{L}_2 - \rho_2^{\mathrm{G}}$；前面的分支图展示了 $2\rho_2^{\mathrm{G}}$。该 $\lambda = 0$ 行是一个新设的普通晚分支对照组。独立随机种子会改变初始化和数据顺序，但架构与训练预算相同。

在系数试验中，在每个测试的正权重下，成员平均损失和离散度都会增加（表28）。在 $\lambda = 0.5$ 处，相对于普通晚期分支，离散度上升了 47.5%，成员平均损失增加了 2.599 毫比特/字节，而几何集成损失仅下降了 0.946 毫比特/字节。在 $\lambda = 1$ 处，离散度是对照组的 2.74 倍，但算术集成损失增加了 1.272 毫比特/字节。

尺度研究在宽度为 192 时重新使用这一 $\lambda = 1$ 二元组，并在每个宽度下使用保存的普通晚分支二元组作为对照组。在宽度为 192 时，保存的对照组与表28中的新鲜对照组略有不同。所有模型都在相同的留出字节上进行评估。令 Δ 表示受惩罚配对的值减去其普通晚期分支对照组的值。额外几何集成增益恰好为

$$G^{\mathrm{G}} = L(q^{\mathrm{G}}_{2,0}) - L(q^{\mathrm{G}}_{2,1}) = \Delta\rho^{\mathrm{G}}_2 - \Delta\bar{L}_2, \tag{24}$$

其中，下标 0 和 1 分别表示普通配对和受惩罚配对。因此，只有当增加的离散度超过成员平均损失的增幅时，提高离散度才能改善集成总损失。

宽度	参数（M）	扩散比	$\Delta\bar{L}_2$	$\Delta\rho^{\mathrm{G}}_2$	G^{G}	G^{A}
192	2.76	2.724	+12.180	+12.927	+0.747	−1.214
256	6.46	2.550	+10.493	+11.278	+0.785	−0.916
384	21.44	2.375	+9.832	+9.269	−0.563	−1.735
512	50.68	2.297	+7.840	+8.253	+0.414	−0.758

表29 更高的成员损失抵消了跨模型规模的额外离散度。 在每个宽度下，相对于保存的普通晚分支对照组，$b = 125$、$T = 200$处 $\lambda = 1$ 的影响。参数数量为总数；扩散比为$\rho^{\mathrm{G}}_{2,1}/\rho^{\mathrm{G}}_{2,0}$。最后四列的单位为毫比特/字节。正的 $G^P = L(q^P_{2,0}) - L(q^P_{2,1})$ 表示更低的集成损失，对于$P \in \{\mathrm{G}, \mathrm{A}\}$而言。所有评估均使用相同的524，288个留出字节。

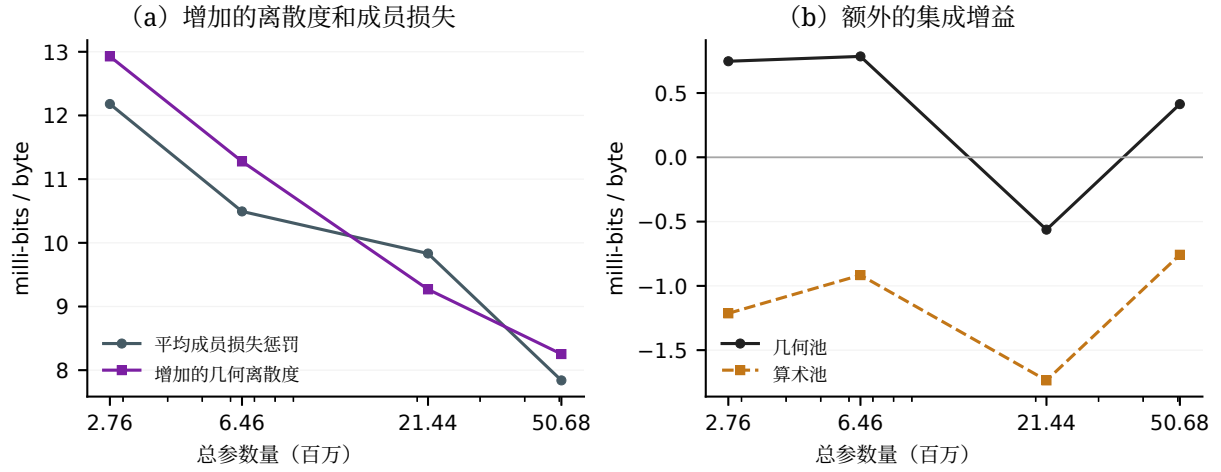


图21 大部分额外的集成增益都被更高的成员损失抵消了。 在 $\lambda = 1$下，(a) 增加的原始几何离散度与平均成员损失惩罚几乎一致；(b) 由此产生的几何池化增益很小，并且在不同规模之间不一致，而算术池化在每种规模下都变差。正的增益表示损失低于普通晚期分支。所有损失变化的单位均为毫比特/字节。

在四种规模下，惩罚项使离散度增加了2.30–2.72倍，同时成员平均损失增加了7.840–12.180毫比特/字节（表29）。这一成员损失惩罚抵消了新增几何离散度的93–106%（图21）。几何平均增益净值介于 -0.563 至 $+0.785$ 毫比特/字节之间，且不会随规模呈现一致的改善；在所有规模下，算术损失均增加了0.758–1.735毫比特/字节。独立随机种子集成在整个过程中仍具有更高的离散度和更低的池化损失。因此，更高的成员平均损失基本抵消了总集成损失中额外的池化收益，尽管相较于各自的成员平均损失，池化仍然改善了每一对模型的表现。

这些是使用每个宽度对应的一个父检查点和续训顺序进行的描述性比较。通常情况下，父模型损失与已保存参考值之间的差异范围为 -0.035 至 $+0.264$ 毫比特/字节，因此数值轨迹变化也与较小网络的汇聚增益有关。

H.5 LoRA 训练与秩扫描

我们固定父模型的参数，从 $b = 125$ 开始，针对每个秩训练两个适配器，并使用相同的剩余数据训练至 $T = 200$。宽度取自分支拟合中保留的宽度。嵌入、输出头和归一化参数保持冻结。我们在每个块中的组合查询/键/值投影、注意力输出以及全部三个前馈矩阵上添加 $W + BA$ 更新。 A 的条目采样自 $\mathcal{N}(0, 1/n_{\text{in}})$， B 从零开始，且 $\alpha/r = 1$。因此，两者开始时都采用父模型的预测。

秩选项为 8， $d/8$， $d/2$。适配器随机种子为 1 和 2；剩余数据顺序种子为 7 和 8。第一个成员与全参数延续使用相同的数据顺序。适配器使用全新的 AdamW 状态，学习率为 0.003，$(\beta_1, \beta_2) = (0.9, 0.95)$， $\epsilon = 10^{-10}$，权重衰减为零，并进行 100 步预热。该学习率通过在最小保留宽度下进行小范围试探选定，未在更大宽度下重新调优。全参数分支保留其优化器，不再重新进行预热。该比较同时改变了可训练子空间和优化处理方式，因此无法将离散度差异单独归因于秩。

H.6 LoRA 离散度与预测损失

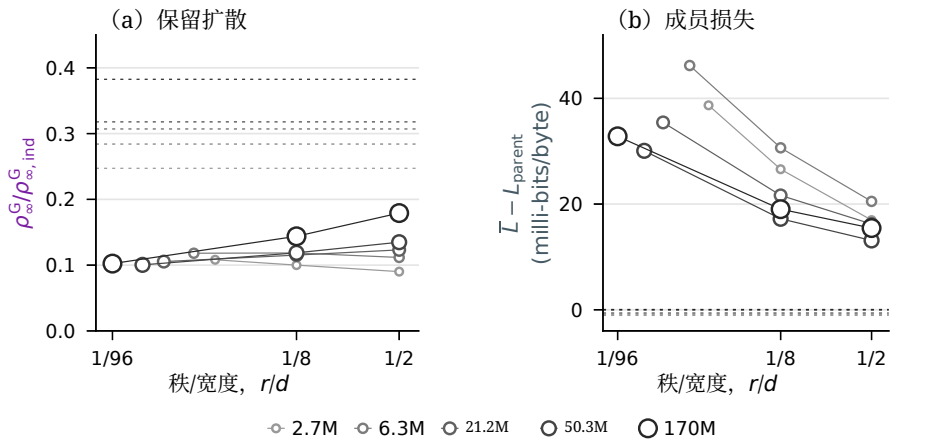


图 22 LoRA 进一步降低了 FineWeb-Edu 上的离散度。 在 $b = 125$, $T = 200$: (a) 保留扩散; (b) 高于父模型的成员平均损失。虚线表示全参数分支。

表30总结了秩扫描结果。在每种宽度下，提高秩都会改善成员损失，而其对离散度的影响则随宽度而变化。即使是最大的秩，其离散度仍低于完整分支离散度的一半。池化后的适配器的损失仍高于一个完整的延续：其较小的汇聚增益不足以抵消成员损失惩罚。因此，这里的较低离散度并不能证明总认知损失更低。

表30 共享训练后的 LoRA 秩扫描。 评估使用留出的 FineWeb-Edu 文本。三个秩列给出了 LoRA 二元组离散度除以独立二元组离散度的结果。最后一列给出了相对于一个完整延续的几何池损失范围（跨秩计算），单位为毫比特/字节。所有损失惩罚项均为正。

宽度 d	$r = 8$	$r = d/8$	$r = d/2$	池化损失惩罚项
192	0.108	0.100	0.090	15.30–36.47
256	0.118	0.119	0.112	18.22–43.76
384	0.105	0.115	0.123	14.31–33.92
512	0.100	0.119	0.135	10.48–28.13
768	0.102	0.144	0.179	12.34–31.02

适配器在评估前进行合并，使用与完整分支相同的留出的 FineWeb-Edu 文本。每对适配器的秩相同。图 8b 则使用近期文本，其中 $r = d/8$ 以及一个覆盖

整个秩扫描范围的带状区域。每种宽度下每个秩对应一个父模型和一对适配器，因此我们不拟合 LoRA 大小规律，也不报告训练总体的置信区间。该结果仅针对这一后期分支和优化器方案，而非所有共享骨干网络不确定性方法。

C4 验证结果

我们在附录 B.3 所述的 96 段英文 C4 验证集文本上，重复图 2、3、6 和 7 中的公开模型分析。对 28 种设置进行的五族拟合给出了$E = 0.38504$ 纳特/字节的 C4 不可约损失估计。其剖面区间为 $[0.31973, 0.42460]$，在应用实测最低 C4 损失（GLM-4.5-Base 的 0.41740 纳特/字节，见表18）之前。对于 C4 拟合，我们使用所得范围 $[0.31973, 0.41740]$。

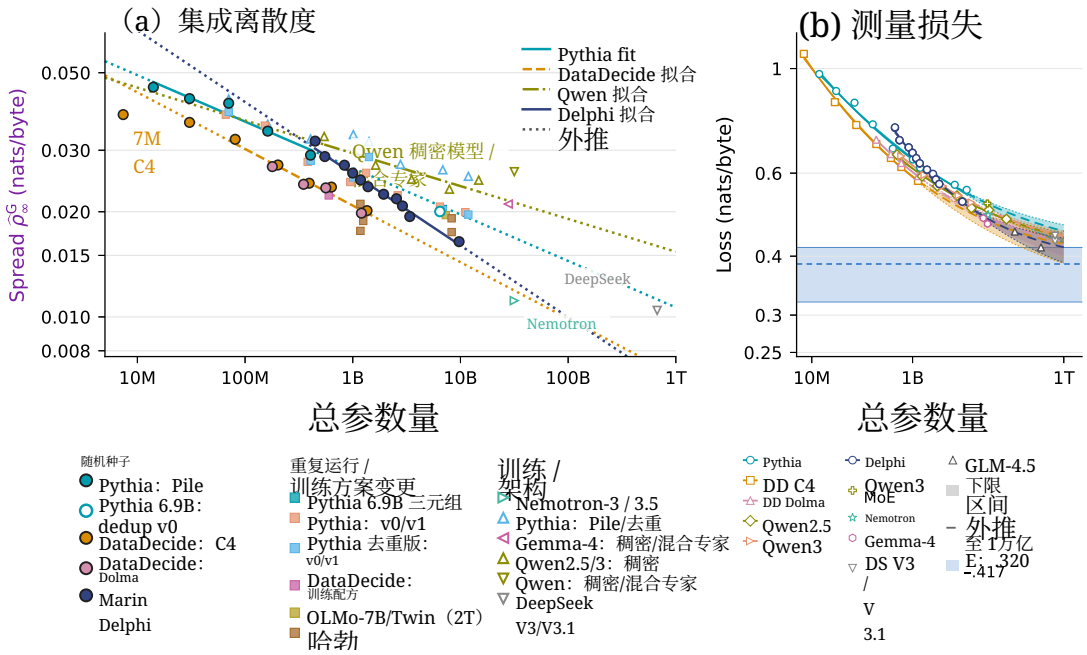


图23 C4 上的集成离散度与单模型损失（参见图2）。（a）在 72 次公开比较中，拟合得到的扩散指数分别为 0.133（Pythia）、0.162（DataDecide）、0.094（稠密 Qwen）和 0.205（Delphi）。（b）在具有 C4 得分的 75 个检查点中，有 66 个进入六族损失拟合。其指数范围为 0.196 至 0.307。Kimi-K2-Base 和 DeepSeek-V4-Pro-Base 仅在近期文本上有得分。

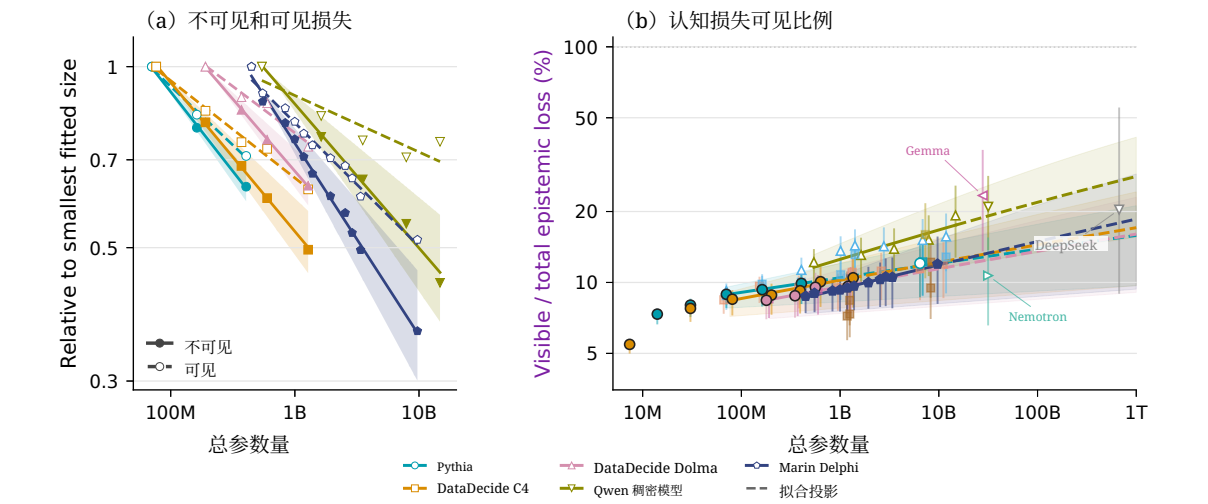


图24 C4 上的可见损失与不可见损失 (参见图3)。(a) 在每个族内，以拟合规模最小的情况对两个分量进行归一化。实线表示不可见损失，虚线表示离散度。(b) 在全部五个拟合序列中，不可见损失的下降速度都快于离散度，因此拟合的可见比例随规模增大而上升。两个面板中的阴影均表示在 C4 下限范围内重新拟合所得的结果。

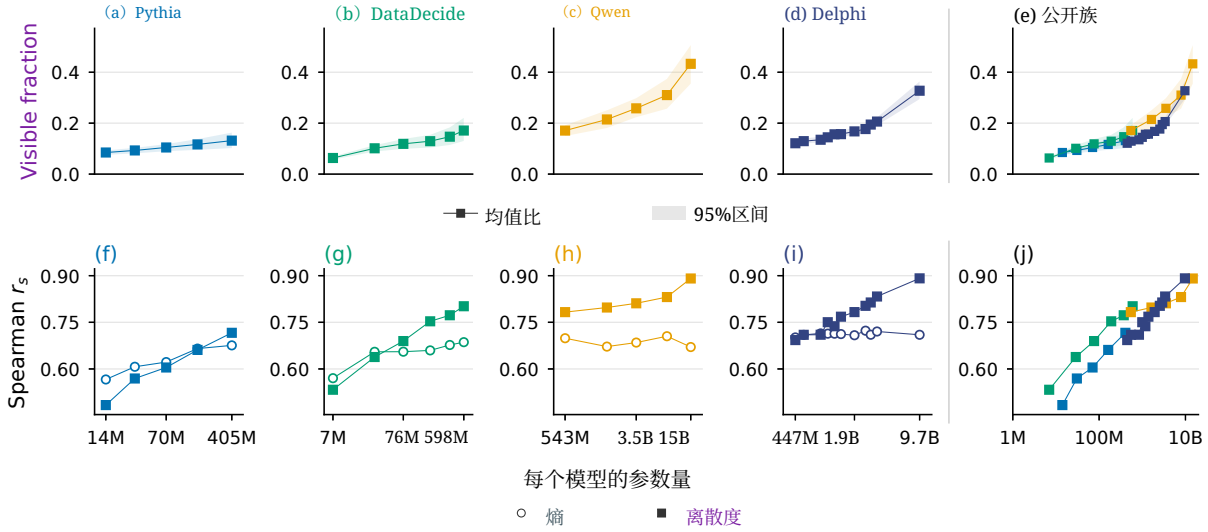


图25 C4 上的上下文依赖的不确定性 (参见图6)。Pythia、DataDecide、Qwen 和 Delphi 共用 40 个留出段落。每个族都在其原生词汇表中对每个段落的 32 个位置进行评分。顶行将平均校正离散度除以平均参考模型与成员模型之间的KL散度，95% 区间通过段落重采样得到。底行展示熵和离散度与参考模型与成员模型之间的KL散度之间的斯皮尔曼相关系数。参考分布见表 21。

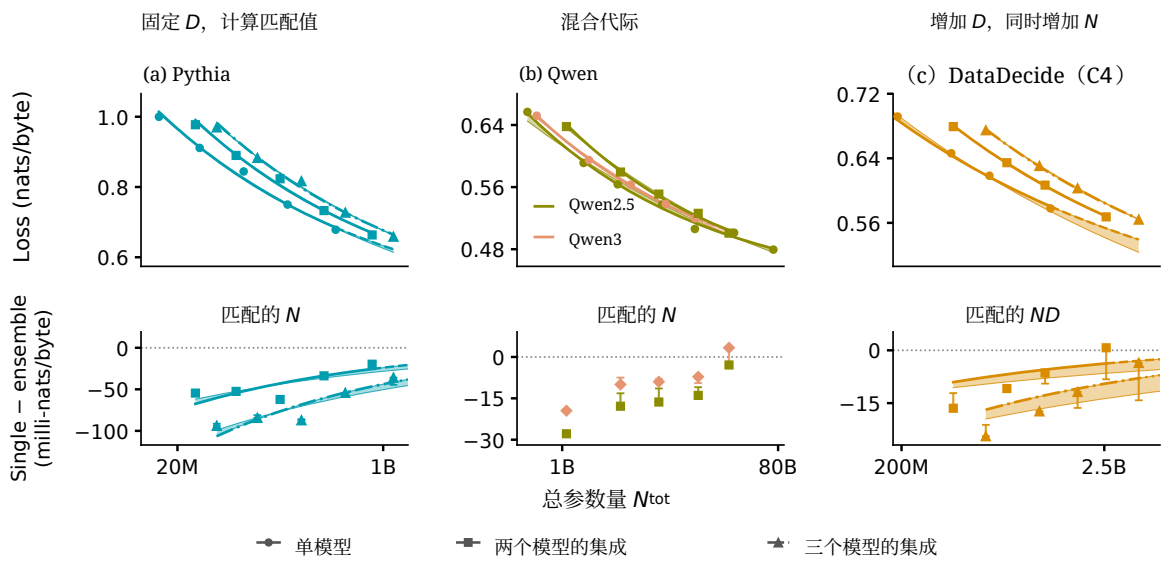


图 26 在 C4 上匹配预算时的集成与更大的单模型（参见图7）。上方面板显示了 Pythia、Qwen 和 DataDecide 的检查点损失与集成损失。下方面板分别在匹配参数量或训练计算量时，用拟合的单模型损失减去集成损失。曲线和阴影采用 C4 下限范围$[0.31973, 0.41740]$ 纳特/字节，中心曲线位于经验上限。