时间并非计算量：消费级 GPU 上墙钟时间受限训练的缩放定律

Yi Liu 独立研究者
lylewis@outlook.com

arXiv:2603.28823v1 [cs.PF] 29 Mar 2026

摘要

缩放定律描述模型质量与计算预算（浮点运算次数）之间的关系，但实践者面临的是墙钟时间约束，而非计算预算。我们研究了在消费级 GPU（RTX 4090）上，时间预算固定为 5 分钟至 24小时的情况下，如何确定最优模型规模。在涵盖 50M–1031M 参数的 70+ 次运行中，我们发现：（1）每种时间预算下都会出现一条 U 形曲线：模型过小会过拟合，模型过大则会训练不足；（2）最优模型规模遵循 $N^* \propto t^{0.60}$，其增长速度快于Chinchilla 的 $N^* \propto C^{0.50}$，并且在所有敏感性分析中，$\alpha = 0.60 \pm 0.07$ 均稳健地优于计算最优规模；（3）存在一种双 *U* 形机制：短时间预算下的 U 形曲线源于计算瓶颈，而长时间预算下的 U 形曲线则源于数据瓶颈（过拟合）；两者之间还存在一个 U 形曲线暂时消失的中间区间。这些发现对使用消费级硬件进行训练的研究者具有直接意义，因为此时起决定性作用的约束是墙钟时间，而非浮点运算次数。我们公开所有代码、日志和70+实验配置。

1 引言

研究动机。 缩放定律是深度学习领域最具影响力的实证发现之一。Kaplan 等人 [2020] 的开创性工作确立了模型性能、模型规模、数据集规模与计算预算之间平滑的幂律关系。Hoffmann 等人 [2022] 将其改进为广泛采用的 Chinchilla 配方：对于计算预算 C（以浮点运算次数计），最优模型规模按 $N^* \propto C^{0.50}$ 的规律增长，且计算资源应大致平均分配给模型参数和训练词元。

然而，这些定律有一个关键假设：自变量是计算资源 （浮点运算次数）。在实践中，大多数研究人员——尤其是使用消费级硬件的研究人员——关注的是墙钟时间，而非浮点运算次数。实践者会问：“我有一台 *RTX 4090*，可以训练 *4*小时。应该训练多大规模的模型？”Chinchilla 无法直接回答这个问题，因为它忽略了一个关键因素：**吞吐量**。

吞吐量差距。 在固定硬件上，不同模型规模的吞吐量差异显著。在 RTX 4090 上，一个拥有 5,000 万参数的模型每秒处理 42.8 万个词元，而一个拥有 5.19亿参数的模型每秒只能处理 3.6 万个词元——差距达 $12\times$ 。在固定时间预算下，这意味着小模型能看到多得多的训练数据。计算最优分配（Chinchilla）隐含地将时间和计算资源视为可以互换的量。事实并非如此：在固定硬件上将模型规模翻倍，并不会 使消耗的计算资源翻倍——计算量仅增加 ∼15%（因为浮点运算次数 $\propto N$ 而吞吐量$\propto N^{-0.8}$，得出 $C \propto N^{0.2} \cdot t$）。
这种脱节会带来实际影响。研究者若遵循 Chinchilla 缩放定律，在 RTX 4090 上以 4 小时为预算，选择的模型规模约为 $\sqrt{C}$。但由于吞吐量并非线性变化，时间最优的选择可能会有很大不同。

贡献。 我们研究**时间约束缩放定律**：在消费级硬件上，给定实际时间预算 t，最优模型规模 $N^*(t)$ 是多少？我们在 $8\times$ RTX 4090 上进行了 70+ 次训练运行，涵盖 5000 万至 10.31 亿个参数，时间预算从 5 分钟到 24 小时不等。我们的主要发现如下：

1. **U 形最优　最优性曲线**覆盖了从 5 分钟到 24小时的各个预算，最优模型规模随着时间推移单调向右移动　t

：D8（50M，5 分钟）→ D10（86M，30 分钟）→ D14（201M，1 小时）→ D16（285M，4h）→ D20（519M，8h）→ D24（856M，12小时）→ D26（1031M，24小时）。

2. **缩放指数** $\alpha = 0.60 \pm 0.07$，稳健地超过 Chinchilla 的 0.50。95% 置信区间$[0.53, 0.67]$ 不包含计算最优值，表明时间约束下的最优规模设定是一个根本不同的机制。

3. **双 U 形机制：** 在较短的预算（≤8小时）下，U 形曲线源于计算瓶颈——大型模型无法处理足够的数据。在较长的预算（≥24小时）下，U 形曲线再次出现，源于数据瓶颈——中型模型在有限数据集上过拟合。在这两个机制之间（∼12小时），随着更大的模型持续单调改进，U 形曲线暂时消失。

4. **收益递减：** $R^2 = 0.971$。最初的 30 分钟带来 0.16 bpb 的改进；第 8–24 小时合计仅带来 0.022 bpb 的改进。

实际意义：</style>将时间预算加倍 $\rightarrow$，模型规模增加 $1.52\times$ ($2^{0.60}$)，而 Chinchilla 的增幅为 $1.41\times$ ($2^{0.50}$)。时间增加 $10\times$ 后，这意味着模型规模分别为 $3.98\times$ 和 $3.16\times$，相差 26%；随着预算延长，这一差距会不断累积。我们公开所有代码和训练日志，以便他人复现。

2 相关工作

神经网络缩放定律。 Kaplan 等人 [2020] 确立了语言模型损失随模型规模、数据集规模和计算资源呈幂律缩放的规律。他们的关键发现——更大的模型具有更高的样本效率——奠定了模型规模不断扩大的趋势。Hoffmann 等人 [2022] （Chinchilla）对此作了修正，指出 Kaplan 等人低估了数据的重要性：计算最优训练要求参数和词元大致同比例扩展，从而得到 $N^* \propto C^{0.50}$和 $D^* \propto C^{0.50}$。我们的工作将计算资源 C替换为墙钟时间 t作为自变量，扩展了这一框架，并发现 $N^* \propto t^{0.60}$。

数据约束缩放。 Muennighoff 等人 [2023] 研究了训练数据有限且必须重复使用的情形，指出数据重复会带来收益递减——这一发现与我们的设定直接相关，因为小模型会耗尽数据集，并多次循环使用其中的数据。Hernandez等人 [2022] 分析了迁移学习缩放；Clark等人 [2022] 提出了融合模型、数据和训练步数的统一缩放定律。

推理最优缩放。 Sardana 和 Frankle [2023] 研究了推理最优训练，其优化目标包含推理成本。他们的设定与我们的设定相呼应：二者都通过纳入实际约束而偏离了纯粹的计算最优性。不过，他们的约束是每次查询的推理成本，而我们的约束是实际训练时间。

消费级硬件上的高效训练。 Dettmers等人 [2023] 通过 4 位量化实现了单图形处理器微调，大幅降低了大语言模型训练的门槛。autoresearch 框架 [Karpathy, 2025] 提供了用于研究的小规模预训练流水线。Dao等人 [2022] 使注意力机制 $O(N)$在内存中运行，从而能够在有限显存上处理更长的序列。我们的工作处于这一生态系统中，研究消费级 GPU 上的预训练缩放。

架构比较。我们比较稠密 Transformer 与混合专家模型 [Jiang 等人，2024 年，Dai 等人，以及 2024],线性注意力变体（RetNet [Sun 等人、 2023], GLA [Yang 等人、 2023], RWKV [Peng 等人， 2023]），发现采用 Flash Attention 的稠密 Transformer 在消费级 GPU 的时间约束下表现占优。

批次大小与训练动态。McCandlish 等人 [2018] 分析了临界批次大小及其与计算–时间权衡的关系。Li 和 Gimpel [2024] 研究了训练过程中的“损失激增”现象。我们的工作对此进行了补充，给出了宏观层面的时间–性能–规模关系。

3 实验设置

3.1 硬件与软件

硬件。 我们使用一台配备 $8\times$ NVIDIA RTX 4090 GPU（每张显存为 24GB）、AMD EPYC 7542 32 核 CPU 和 503GB 内存的服务器。每张 GPU 都运行一个独立的训练任务——我们不 使用多GPU并行（数据并行或模型并行）。这种“易并行”的设置可最大限度地提高实验吞吐量：同时运行 8 种配置。RTX 4090 是 2026 年可用的最高端消费级 GPU，因此我们的结果与研究群体直接相关。

软件。 我们使用 autoresearch 框架 [Karpathy, 2025]搭配 Flash Attention v2、torch.compile（模式=default, BF16精度）以及带预热的余弦学习率调度。所有训练均采用单轮次范式，在训练预算内将学习率衰减至峰值的 10%，并设置 10% 的预热步数。

3.2 架构族

我们使用仅解码器Transformer，并采用单一的“深度”参数 D 来控制所有维度 :

$$n_layers = D, \quad model_dim = 64D, \quad n_heads = D, \quad head_dim = 64 \tag{1}$$

这种参数化方式确保不同规模下的架构比例保持恒定：纵横比（宽度/深度）始终为 64，注意力头的数量等于深度。这消除了架构选择带来的混杂变量，并隔离出规模的影响。

附录中的表 3 展示了所有配置。关键数据点：D8 有 5030 万个参数，吞吐量为 42.8 万令牌/秒；D14 有 2.009 亿个参数，吞吐量为 11 万令牌/秒；D20 有 5.19 亿个参数，吞吐量为 3.6 万令牌/秒；D26 有 $\sim$1031M 个参数，吞吐量为 $\sim$5K 令牌/秒。我们的模型族在吞吐量上的这一$85\times$ 跨度，正是导致时间约束缩放与计算约束缩放出现分歧的原因。

3.3 数据集

我们使用 FineWeb-Edu [Penedo 等人， 2024], 这一经过整理的网络语料库，并用 BPE（32K 词表）进行分词。我们的训练集包含 $\sim$48M 个词元。这个刻意缩小的数据集有助于研究**容量–数据–过拟合相互作用**：小模型会很快耗尽数据（D8 在 12小时内循环训练 250+ 次），而大型模型（D26，24小时）训练轮次不到 3 轮。我们全程报告验证集每字节比特数（BPB）。

为什么使用小数据集？ 虽然更大的数据集可以避免过拟合，但我们使用小数据集有一个重要目的：它构成了一个天然实验室，让我们能在切实可行的时间预算内，同时研究两种次优性——训练不足的大模型和过拟合的小模型。如果数据集足够大，U形曲线的左侧分支（过拟合）要到训练时间长得多时才会出现，这会让这一现象更难观察。

表1：不同模型规模和时间预算下的验证 BPB。**粗体** = 表示各时间预算下的最优结果。 ↑ = 表示过拟合（结果比训练时间更短时更差）。在12小时下，没有U形曲线：BPB随着模型规模增大而单调下降。在24小时下，由于数据瓶颈，U形曲线再次出现。

模型	参数	5min	30 分钟	1h	2h	4h	8h	12h	24h
D8	50M	**1.133**	0.977	0.979↑	0.906	0.925↑	0.886	0.919↑	—
D10	86M	1.178	**0.973**	0.976	0.906	0.892	0.886	0.885	—
D12	135M	1.363	1.001	0.991	0.904	0.878	0.873	0.871	0.870
D14	201M	1.578	1.016	**0.945**	**0.901**	0.866	0.854	0.852	0.857↑
D16	285M	1.566	1.026	0.951	**0.901**	**0.862**	0.844	0.841	0.851↑
D18	384M	—	—	—	—	0.866	0.837	0.833	0.845↑
D20	519M	1.804	—	1.009	—	0.872	**0.836**	0.828	0.838↑
D22	621M	—	—	—	—	—	—	0.826	0.829
D24	856M	1.854	—	—	—	0.896	0.845	**0.824**	0.817
D26	1031M	—	—	—	—	—	—	—	**0.814**

3.4 实验设计

我们针对 8 种时间预算训练模型：5 分钟、30 分钟、1 小时、2小时、4小时、8小时、12小时和24小时。在每种预算下，我们在多块 GPU 上并行运行 4–8 种模型规模的模型。除 DEPTH 以及因显存限制而需要减小批次大小的情况外（D26 使用批次大小 1），所有配置均采用相同的超参数（学习率 3×10^{-4}，批次大小为 64 个序列，每个序列包含 512 个词元）。学习率消融实验（第 4.5 节）证实了模型对这一选择具有稳健性。

总计，我们进行了 70+次独立训练运行，涵盖 $20\times$ 的参数范围（5000 万至 10.31 亿）和 $288\times$ 的时间范围（5分钟至24小时）。

4 结果

4.1 每种时间预算下的 U 形最优性

我们的核心实证发现是，在每种时间预算下，验证 BPB 随模型规模变化都会呈现典型的 U 形曲线。表1展示了涵盖全部 8 种时间预算和 10 种模型规模的完整结果。

U 形曲线源于两种相互竞争的力量。左侧分支（模型太小）：这些模型会耗尽训练数据并过拟合。D8 在 4 小时的每字节比特数为 0.925，差于其 2 小时的结果 0.906；到 12小时，其每字节比特数进一步恶化至 0.919，此前它已将 4800万 token 数据集循环训练了 250 多次。右侧分支（模型太大）：吞吐量不足意味着处理的数据太少，无法充分利用模型容量——D24 在 5 分钟内仅处理了 1300 万个词元，每字节比特数为 1.854，远差于 D8 处理 1.34 亿个词元时的 1.133。

最优模型会随时间单调向右移动：D8（5 分钟）$\rightarrow$ D10（30 分钟）$\rightarrow$ D14（1 小时）$\rightarrow$ D14=D16（2 小时）$\rightarrow$D16（4 小时）$\rightarrow$ D20（8 小时）$\rightarrow$ D24（12小时）$\rightarrow$ D26（24小时）。这种单调向右移动是拟合缩放定律所依据的关键经验规律。

4.2 时间约束缩放定律

对全部 8 个时间点（5 分钟–24小时）拟合幂律 $N^*(t) = a \cdot t^{\alpha}$：

$$N^*(t) = 14.20 \times t^{0.595} \quad (R^2 = 0.963) \tag{2}$$

其中， N^* 为最优参数（百万）， t 为时间（分钟）。指数 $\alpha = 0.595 \pm 0.067$ **比 Chinchilla 的 0.50 计算最优指数高约 20%**。

统计显著性。 α 的 95% 置信区间为 $[0.53, 0.67]$，不包含 Chinchilla 值 0.50。这证实，时间约束缩放代表着一个与计算约束缩放截然不同的机制。

实际解读。将时间预算翻倍，建议将模型规模扩大 $2^{0.60} = 1.52\times$ ，而 Chinchilla 的建议是 $2^{0.50} = 1.41\times$。在增加 $10\times$ 的情况下，这意味着模型规模分别为 $3.98\times$ 和 $3.16\times$ 。差异会不断累积：对于一位研究人员来说，从 1 小时的“快速实验”转为 24 小时的通宵运行，我们的定律建议使用大 $24^{0.60}/1^{0.60} = 7.2\times$ 的模型，而 Chinchilla 建议大 $24^{0.50} = 4.9\times$。

深度缩放。最优深度也遵循幂律： $D^*(t) = 4.97 \times t^{0.231}$ ($R^2 = 0.958$)。

敏感性分析。 我们通过改变拟合选择来验证结果的稳健性（完整结果见附录表 4）。在 2 小时时，D14 和 D16 的每字节比特数均为 0.901，并列最佳；我们测试了使用平均值（243M）、仅使用 D16、仅使用 D14，以及排除 2 小时数据点这几种情况。在所有变体中， α 都稳健地超过 0.50（$\alpha \in [0.60, 0.81]$）。值得注意的是，加入 24 小时数据后， α 从 0.747（7 个数据点）降至 0.595（8 个数据点）。我们将在第 5 节讨论这种收敛行为。

4.3 损失缩放

各时间预算下可达到的最佳损失如下：

$$L^*(t) = 1.223 \times t^{-0.061} \quad (R^2 = 0.971) \tag{3}$$

收益递减现象十分显著（完整细分见附录表 5）： 前 30分钟带来 0.16 bpb 的改善（每小时 0.38 bpb），而第 8 至 24 小时合计仅改善 0.022 bpb（每小时 0.001 bpb）。对于数据有限的消费级 GPU，第一小时就能获得大部分可实现的改善——快速探索性运行（1–4 小时）便能提供大部分信号。

4.4 双重U形：两个瓶颈，一条曲线

我们在24小时实验中发现了一个引人注目的现象：出现了第二种 U 形曲线，其背后的机制与第一种截然不同。

短预算 U 形曲线（5 分钟–8小时）：计算瓶颈。 在预算为 ≤8小时时，U形曲线的右侧曲线段之所以出现，是因为大型模型的吞吐量不足以处理足够的数据。D24（8.56亿）在4小时内达到的每字节比特数为0.896，远未发挥其潜力，因为它只处理了 ∼13百万个词元。

U形曲线消失（∼12小时）：过渡阶段。在 12小时 时，出现了一个引人注目的现象：U形曲线消失了。每字节比特数从 D10（0.885）到 D24（0.824）单调下降。最大的模型提升最为显著（$\Delta_{8h \to 12h} = -0.021$），这表明真正最优值超过了我们的显存容量。

表 2：30分钟多随机种子验证。每个模型使用三个随机种子。最大值为 $\sigma = 0.003$ 每字节比特数，模型间差距为 0.002–0.056 每字节比特数。D10 在所有随机种子下均为最优。

模型	参数	随机种子 42	随机种子 123	随机种子 456	均值 ±标准差	变异系数 (%)
D8	50M	0.977	0.975	0.975	0.976 ± 0.001	0.13
D10	**86M**	**0.973**	**0.974**	**0.973**	$\mathbf{0.973 \pm 0.000}$	**0.04**
D14	201M	1.016	1.022	1.017	1.018 ± 0.003	0.29
D16	285M	1.026	1.032	1.029	1.029 ± 0.003	0.31

长预算 U 形曲线（24小时）：数据瓶颈。在 24小时 时，U形曲线回归了——但其左侧分支出现了新的成因。中型模型（D14–D20）此前在较短预算下表现最优，如今却过拟合：D14 的每字节比特数从 12小时 时的 0.852 恶化至 24小时 时的 0.857；D18 则从 0.833 恶化至 0.845（+0.012）。只有 D24 和 D26 继续改善（完整细分见附录中的表 6）。24小时 时 U形曲线的左侧分支由数据重复驱动，而非计算资源不足。

这种双 U 形机制意味着，在较长的预算下，近乎最优的模型规模"安全区"会变窄：在 4h 时，D14–D18 的表现都与最优值相差不超过 0.004；在 24小时 时，只有 D24–D26 能避免过拟合。

4.5 消融研究

学习率敏感性。我们在两个模型规模（D8 和 D14）上进行学习率消融实验，训练 1 小时，测试学习率取值从 1.5×10^{-4} 到 6×10^{-4}。D14 的默认学习率（3×10^{-4}）接近最优，与最佳学习率（6×10^{-4}，每字节比特数为 0.944，而非 0.945）相差不到 0.001。即使采用最优学习率，D8（每字节比特数为 0.948）的表现也不及 D14（0.945），这证实缩放定律对中等程度的学习率变化具有鲁棒性。

多随机种子验证。 我们对四种模型规模（D8、D10、D14、D16）分别使用 3 个随机种子（42、123、456）进行 30 分钟训练（表 2）。结果非常稳定：随机种子间的标准差至多为0.0032 每字节比特数，而模型间差距达到 0.056 每字节比特数——相差 17× 倍。所有模型的变异系数均低于 0.31%。重要的是，在所有随机种子下，D10 在 30 分钟时仍是最优模型（均值 $= 0.9734 \pm 0.0004$），且每个 D10 随机种子的结果都优于每个 D8 随机种子的结果，这证实 Chinchilla 交叉点对随机性具有稳健性。这些结果验证了我们的主要单随机种子结果能够代表潜在趋势。

多架构比较。 在 RTX 4090 上训练 5 分钟后，我们将稠密 Transformer 与四种替代方案进行比较：混合专家模型（64 个专家，top-1，细粒度）、采用 FLA Triton 内核的 RetNet、采用 FLA 内核的 GLA，以及采用 FLA 内核的 RWKV6。结果：稠密 Transformer（1.133）$\gg$ 混合专家模型（1.143）$\gg$ RetNet（2.216）$>$ GLA（2.249）$>$ RWKV6（2.258）。线性注意力变体的表现落后近 2×，尽管其模型浮点运算利用率（MFU：20–27%，而稠密模型为 10%）更高。架构比较证明了我们聚焦于稠密 Transformer 的合理性。

5 讨论

5.1 为什么时间 $\neq$ 计算资源

时间–计算量分歧的数学起源非常直接。对于规模为 N、训练时间为 t 的模型，消耗的计算资源为：

$$C(N, t) = 6N \times \tau(N) \times t \tag{4}$$

其中， $\tau(N)$ 表示吞吐量（令牌/秒），系数 6 计入了前向和反向传播。我们在硬件上对 $\tau(N) \propto N^{-0.8}$ 进行了实证测量（表 3），结果为：

$$C \propto N \times N^{-0.8} \times t = N^{0.2} \times t \tag{5}$$

在 计算资源约束 （固定 C）下，将 N 加倍会使可用词元减半： $T = C/6N$。增大模型规模的代价是直接且线性的。在 时间约束 （固定 t）下，将 N 加倍只会使词元减少 $2^{0.8} \approx 1.74\times$: $T = \tau(N) \times t$。这种代价是间接的，通过吞吐量产生影响，而且是次线性的。较弱的代价意味着，相比计算资源约束，时间约束更有利于较大的模型——这正是 $\alpha = 0.60 > 0.50$ 所揭示的。

5.2 与 Chinchilla 的比较

$\alpha = 0.60$ 与 0.50 的定量差异会随着预算延长而累积。将时间加倍时，规定的 $1.52\times$ 模型规模（而 Chinchilla 规定的是 $1.41\times$）；时间增加 $10\times$ 时： $3.98\times$ 与 $3.16\times$；对于从 1 小时扩展到 24 小时这一常见情形： $7.2\times$ 与 $4.9\times$ （见附录表 7）。
这种分歧背后的机制是数据重复惩罚。在时间约束下，小模型会反复处理相同数据，因重复而遭受累积损失 [Muennighoff 等人， 2023]。这会产生一种推动模型变大的“推力”，而在数据不受限的计算约束环境中则不存在这种情况。

5.3 α的演变

一个耐人寻味的现象是，随着我们加入更长的时间点， α 呈现出非单调收敛 ts:

- 5 个时间点（5 分钟–4 小时）： $\alpha = 0.44$——低于 Chinchilla
- 6 个时间点（+8小时）： $\alpha = 0.55$ ——略高于 Chinchilla
- 7 个时间点（+12小时）： $\alpha = 0.75$ ——远高于 Chinchilla
- 8 个点（+24小时）： $\alpha = 0.60$ ——已稳定在 Chinchilla 之上

7 个点的估计值（$\alpha = 0.75$）偏高，因为 12 小时是 U 形曲线消失且 D24（最大的模型）变为最优模型的首个预算时长——这带来了过大的杠杆效应。加入 24 小时的数据，并以 D26（1031M）作为新的最优模型后， α “回落”至较为中等的 0.60。我们认为 $\alpha \approx 0.60$ 是一个合理的估计值，不过随着预算时长进一步延长，该数值可能还会继续变化。

α 呈现振荡而非单调收敛，这一事实表明， N^* 与 t 之间的关系在所有时间尺度上可能并非纯粹的幂律。双 U 形机制（第 4.4 节）引入了机制转换，单一幂律无法完美捕捉这些转换。尽管如此， $R^2 = 0.963$表明，在我们观察到的范围内，拟合效果极佳。

5.4 消失又重现的 U形曲线

U形曲线在不同时间预算下的演变揭示了三个截然不同的区间：
区间 1：计算受限（5 分钟–8 小时）。U形曲线具有明显的最低点。右侧曲线段是由吞吐量限制导致的。最优模型规模远低于显存容量。
阶段 2：过渡阶段（~12小时）。U形曲线消失：每字节比特数随模型规模增大而单调下降。这代表着“最佳点”：即使是能够装入显存的最大模型（D24，8.56亿）也能从延长训练中获益，而不会过拟合。真正的最优值可能超出单 GPU 显存的限制——这正是自然过渡到多 GPU 训练的节点。

阶段 3：数据受限（≥24小时）。U形曲线再次出现，但成因恰好相反：左侧分支如今是由过拟合（在有限数据上训练过多轮次）造成的，而非计算资源不足。模型 D14–D20 的每字节比特数比其 12小时时更差。此阶段仅适用于有限数据集；若数据量更大，曲线会向右移动。
这种三阶段结构具有实际意义：对于每种数据集规模，都存在一个“黄金时长”；超过这一时长后，若要延长训练时间，就需要增加数据，或将模型规模扩大到超出单 GPU 容量。

5.5 实用指南

经验法则： 时间翻倍 $\rightarrow 1.52\times$ 模型规模（$2^{0.60}$）。在探索性工作中，训练 1–4 小时即可获得可实现改进的 80%+。对于通宵运行（12–24小时），应在显存限制内尽可能增大模型规模。附录中的表 8 针对 RTX 4090 给出了各时间预算的具体建议。

5.6 局限性

我们的研究存在若干有待未来工作解决的局限性：

1. **仅使用一种 GPU**。 RTX 4090 具有特定的吞吐量与规模缩放关系。A100/H100 GPU 由于内存带宽、计算比率和多 GPU 互连不同，可能会产生不同的 α 。2. **数据集规模较小**。 我们的 4800万 token 数据集放大了数据重复效应。对于更大的数据集（数十亿词元），数据受限阶段（第 4.4 节）会向右移动，可能使 α 降至 0.50。3. **仅使用一种架构**。 虽然我们确认稠密 Transformer 在消费级 GPU 上优于其他方案，但其他架构的具体 α 值可能有所不同。4. **多随机种子覆盖有限**。 尽管我们在30分钟时进行的多随机种子实验证实种子间方差可以忽略不计（$\sigma < 0.003$），但在全部 8 种时间预算下进行完整的多随机种子验证，将进一步增强置信区间的可靠性，尤其是在较长的时间预算（12–24小时）下，因为单次运行的成本较高。5. **未进行多 GPU 训练**。 在 12–24小时的时间预算下，最优模型的规模接近或超过单 GPU 显存容量，这表明更长的时间预算需要进行多 GPU 训练。由于通信开销，多 GPU 阶段的缩放定律可能会呈现不同的行为。

6 结论

我们提出了**面向消费级 GPU 上语言模型训练的时间约束缩放定律**。我们的核心发现是，最优模型规模按 $N^*(t) \propto t^{0.60}$ 缩放，超过了 Chinchilla 的计算最优指数 0.50。这种差异源于次线性吞吐量缩放（$\tau \propto N^{-0.8}$），它打破了时间与计算资源作为优化目标时的等价关系。

我们的 70+训练运行呈现出三个关键现象：

1. **每种时间预算下的 U 形最优性曲线**，最优模型规模从 50M（5 分钟）到 1031M（24小时）单调向右移动。2. **双 U 形机制**：短时间预算下的 U 形曲线源于计算瓶颈，长时间预算下的 U 形曲线源于数据瓶颈，而在过渡阶段，U 形曲线会消失。3. **收益严重递减**：第一小时带来了大部分可实现的提升；训练时间从 8 小时延长到 24 小时，BPB 仅改善 0.022——这一发现应为研究人员分配 GPU 时间提供参考。

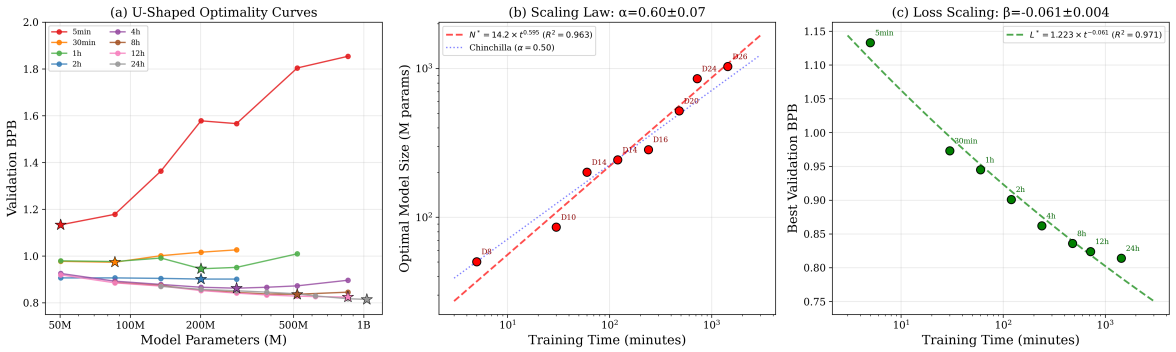


图1：**主要结果**。（**a**）8 种时间预算（5 分钟–24小时）下的 U 形最优性曲线。星号标示各时间预算下的最优模型。（**b**）最优模型规模与时间的关系，使用 $N^* = 14.2 \times t^{0.595}$ ($R^2 = 0.963$) 拟合。蓝色虚线展示了 Chinchilla 的 $\alpha = 0.50$，以供参考。（**c**）可达到的最佳每字节比特数与时间的关系，使用 $L^* = 1.22 \times t^{-0.061}$ ($R^2 = 0.971$) 拟合，显示收益递减十分显著。

大多数从业者面临的是时间约束，而非计算预算。以墙钟时间（而非太 FLOP）为单位提供指导，能让缩放定律对研究群体更具实践指导意义。

更广泛的影响。我们的发现表明，“越大越好”的说法虽然方向正确，但夸大了时间约束下模型规模带来的益处。对于日益壮大的消费级硬件训练研究者群体而言，了解时间约束缩放规律，可以避免数千 GPU小时的计算资源被错误分配。

未来工作。自然的拓展方向包括：(1)在 12小时+ 预算下进行多 GPU 训练，此时最优模型规模会超过单 GPU 显存容量；(2)使用吞吐量特征各异的数据中心GPU（A100/H100）；(3)采用更大的数据集，以区分吞吐量效应与数据重复的影响；(4)测试 α 是否会随着时间点增多而收敛至稳定值；以及 (5) 将研究拓展至大语言模型流程中的微调和基于人类反馈的强化学习阶段。

参考文献

Aidan Clark、Diego De Las Casas、Aurelia Guy、Arthur Sherburn、Katie Millican、BenSherburn 和 Tom Sherburn。路由语言模型的统一缩放定律。国际机器学习会议，2022 年。

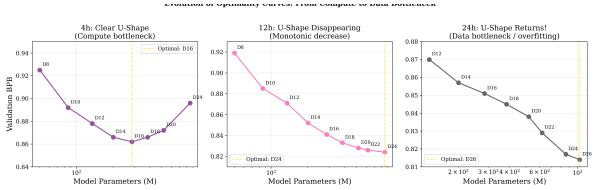
Damai Dai、Chengqi Deng、Chenggang Zhao、R X Xu、Huazuo Gao、Deli Chen、Jiashi Li、Wangding Zeng、Xingkai Yu、Y Wu 等。Deepseekmoe：迈向混合专家语言模型中的终极专家专门化。*arXiv* 预印本 *arXiv:2401.06066*，2024 年。

Tri Dao、Daniel Y Fu、Stefano Ermon、Atri Rudra 和 Christopher Ré。FlashAttention：具备 I/O 感知能力的快速、节省内存的精确注意力机制。神经信息处理系统进展，35：16344–16359，2022 年。

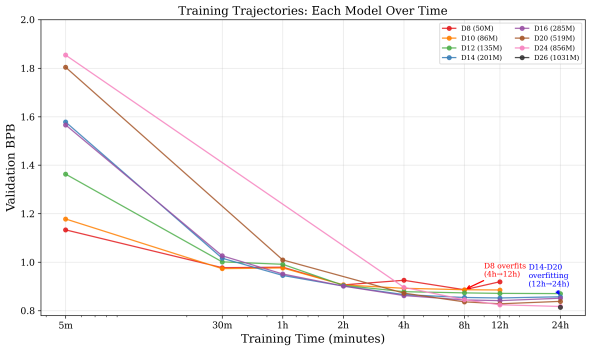
Tim Dettmers、Artidoro Pagnoni、Ari Holtzman 和 Luke Zettlemoyer。QLoRA：高效微调量化语言模型。神经信息处理系统进展，36，2023 年。

Danny Hernandez、Jared Kaplan、Tom Henighan 和 Sam McCandlish。《迁移缩放定律》。*arXiv* 预印本 *arXiv:2102.01293*，2022 年。

Jordan Hoffmann、Sebastian Borgeaud、Arthur Mensch、Elena Buchatskaya、Trevor Cai、Eliza Rutherford、Diego de Las Casas、Lisa Anne Hendricks、Johannes Welbl、Aidan Clark 等。训练计算最优的大语言模型。*arXiv* 预印本 *arXiv:2203.15556*，2022 年。



(a) U形曲线演变：4h 时呈现清晰的 U 形曲线（计算瓶颈），12h 时单调下降（过渡阶段），24h 时 U 形曲线再次出现（数据瓶颈）。



(b) 训练轨迹。D8 在 2 小时后过拟合；D14–D20 在 12h 至 24h 之间过拟合；D24 和 D26 在 24h 时仍持续改善。

图2：**关键现象**。 左：三种状态下的双 U 形机制。右：训练轨迹展现了模型特有的过拟合动态。更多图表（α 收敛图、热力图、双 U 形细节）见附录。

Albert QJiang、Alexandre Sablayrolles、Antoine Roux、ArthurMensch、BlancheSavary、Chris Bamford、DevendraSingh Chaplot、Diego de las Casas、Emma Bou Hanna、Florian Bressand 等。Mixtral 专家模型。*arXiv* 预印本 *arXiv:2401.04088*，2024 年。

Jared Kaplan、Sam McCandlish、Tom Henighan、Tom B Brown、Benjamin Chess、Rewon Child、Scott Gray、Alec Radford、Jeffrey Wu 和 Dario Amodei。神经语言模型的缩放定律。*arXiv* 预印本 *arXiv:2001.08361*，2020年。

Andrej Karpathy。autoresearch：自动化大语言模型研究框架。https://github.com/karpathy/ autoresearch，2025。

Xian Li 和 Kevin Gimpel。大语言模型训练中的损失尖峰。*arXiv* 预印本 *arXiv:2401.10330*，2024。

Sam McCandlish、Jared Kaplan、Dario Amodei 和 OpenAI Dota Team。大批量训练的经验模型。*arXiv* 预印本 *arXiv:1812.06162*，2018。

Niklas Muennighoff、Alexander M Rush、Boaz Barak、Teven Le Scao、Nouamane Tazi、Aleksandra Piktus、Sampo Pyysalo、Thomas Wolf 和 ColinRaffel。扩展数据受限语言模型。神经信息处理系统进展，36，2023。

Guilherme Penedo、Hynek Kydlí˘cek、Loubna Ben Allal、Anton Lozhkov、Margaret Mitchell、Colin Raffel、Leandro Von Werra 和 Thomas Wolf。FineWeb 数据集：提炼网络内容，以大规模获取最优质的文本数据。*arXiv* 预印本 *arXiv:2406.17557*，2024。

Bo Peng、Eric Alcaide、Quentin Anthony、Alon Albalak、Samuel Arcadinho、Stella Biderman、Huanqi Cao、Xin Cheng、Michael Chung、Matteo Grella 等。《RWKV：为 Transformer 时代重新发明循环神经网络》。*arXiv* 预印本 *arXiv:2305.13048*，2023。

Nikhil Sardana 和 Jonathan Frankle。《超越 Chinchilla 最优：将推理纳入语言模型缩放定律的考量》。*arXiv* 预印本 *arXiv:2401.00448*，2023。

Yutao Sun、Li Dong、Shaohan Huang、Shuming Ma、Yuqing Xia、Jilong Xue、Jianyong Wang 和 Furu Wei。RetNet：面向大语言模型的 Transformer 后继架构。*arXiv* 预印本 *arXiv:2307.08621*，2023。

Songlin Yang、Bailin Wang、Yikang Shen、Rameswar Panda 和 Yoon Kim。采用硬件高效训练的门控线性注意力变换器。*arXiv* 预印本 *arXiv:2312.06635*，2023。

A 模型配置

表 3：按深度参数化的完整模型配置。吞吐量在单张 RTX 4090 上测得。

深度	层	Dim	参数（M）	令牌/秒	模型浮点运算利用率 (%)
8	8	512	50.3	428K	10.4
10	10	640	85.9	252K	10.7
12	12	768	135.3	160K	10.9
14	14	896	200.9	110K	11.4
16	16	1024	285.2	78K	11.4
18	18	1152	∼384	56K	∼11
20	20	1280	519.0	36K	∼11
22	22	1408	∼621	27K	∼11
24	24	1536	855.6	20K	∼11
26	26	1664	∼1031	∼5K	∼3

B 敏感性分析

表 4：缩放指数 α 对拟合选择的敏感性。所有变体均得出 $\alpha > 0.50$。

变体	α	95% 置信区间	R^2
8 点（基线）	0.595 ± 0.067	$[0.53, 0.67]$	0.963
7 点（不包括 24小时）	0.747 ± 0.107	$[0.64, 0.86]$	0.957
7 点（2 小时 = 285M）	0.706 ± 0.110	$[0.60, 0.82]$	0.948
6 点（不包括 2 小时）	0.805 ± 0.145	$[0.66, 0.95]$	0.960

C 边际收益

Ta表 5：每个时间增量的 BPB 边际提升。1 小时后，收益迅速递减。

区间	Δt	Δ 每字节比特数	每小时 BPB	效率
5 分钟 → 30分钟	25 分钟	-0.160	-0.384	非常高
30 分钟 → 1h	30 分钟	-0.028	-0.056	High
1 小时 → 2h	1h	-0.044	-0.044	中等
2 小时 → 4h	2h	-0.039	-0.020	Low
4 小时 → 8h	4h	-0.026	-0.007	极低
8小时 → 12小时	4h	-0.012	-0.003	递减
12小时 → 24小时	12h	-0.010	-0.001	极小

D 过拟合分析(12小时 → 24小时)

表 6：从 12小时到 24小时的每字节比特数变化。正值表示过拟合。只有 ≥856M 模型仍在持续改进。

模型	参数	每字节比特数$_{12h}$	每字节比特数$_{24h}$	Δ
D12	135M	0.871	0.870	-0.001 (饱和)
D14	201M	0.852	0.857	$+0.005$ (过拟合)
D16	285M	0.841	0.851	$+0.010$ (过拟合)
D18	384M	0.833	0.845	$+0.012$ (过拟合)
D20	519M	0.828	0.838	$+0.010$ (过拟合)
D22	621M	0.826	0.829	$+0.003$ (边际)
D24	856M	0.824	0.817	-0.007 （改进中）
D26	1031M	—	0.814	首次运行

E 与 Chinchilla 的比较

表 7：模型规模建议：时间约束（我们的方法， $\alpha = 0.60$）与计算约束（Chinchilla，$\alpha = 0.50$）的比较。

时间倍数	我们的方法（$2^{0.60}$）	Chinchilla（$2^{0.50}$）
2× 时间	1.52× 模型	1.41× 模型
4× 时间	2.30× 模型	2.00× 模型
10× 时间	3.98× 模型	3.16× 模型
24× 时间（1 小时→24小时）	7.22× 模型	4.90× 模型

F 附加图表

G 多随机种子验证详情

多随机种子实验是一项关键的稳健性检验。在 30分钟的预算下——D10（86M）和 D8（50M）的每字节比特数仅相差 0.002——所有三个 D10 随机种子的每字节比特数都低于所有三个 D8 随机种子。所有 $3 \times 3 = 9$ 组比较均呈现这种成对优势，这有力地表明，30分钟时的 Chinchilla 交叉点是真实存在的现象，而非统计伪象。

与 D14 和 D16（$\sigma \approx 0.003$）相比，D10 的方差极低（$\sigma = 0.0004$），这可能反映出 D10 在 30 分钟时处于一个自然的“最佳点”：它已处理了足够多的数据（∼450M 个词元，∼5× 数据集），从而产生稳定的表征；而 D14 和 D16 相对于其容量所见的词元更少，因此对初始化更为敏感。

H 实用指南

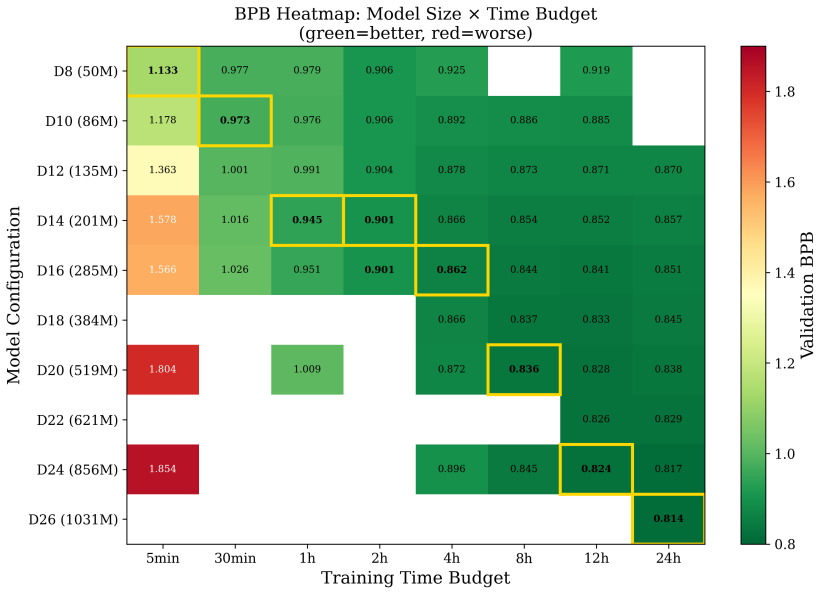


图 3：不同模型规模和时间预算下的 BPB 热力图。最优值构成的对角带描绘了受时间约束的缩放定律。金色方框标出了最优配置。

表 8：使用 FineWeb-Edu 规模数据在 RTX 4090 上训练时的推荐模型规模。

时间预算	模型规模	深度	预期 BPB
5分钟	50M	D8	1.133
30分钟	86M	D10	0.973
1小时	200M	D14	0.945
2小时	2亿至2.85亿	D14至D16	0.901
4小时	285M	D16	0.862
8小时	519M	D20	0.836
12小时	856M	D24	0.824
24小时	∼1B	D26	0.814

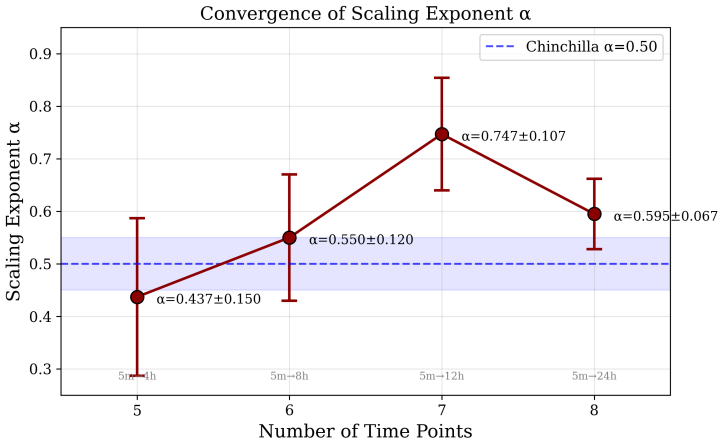


图4：随着时间点增加，α 的演变：0.44（5个点）$\rightarrow$ 0.55（6个点）$\rightarrow$ 0.75（7个点）$\rightarrow$ 0.60（8个点）。非单调收敛反映了机制转换（第5节）。

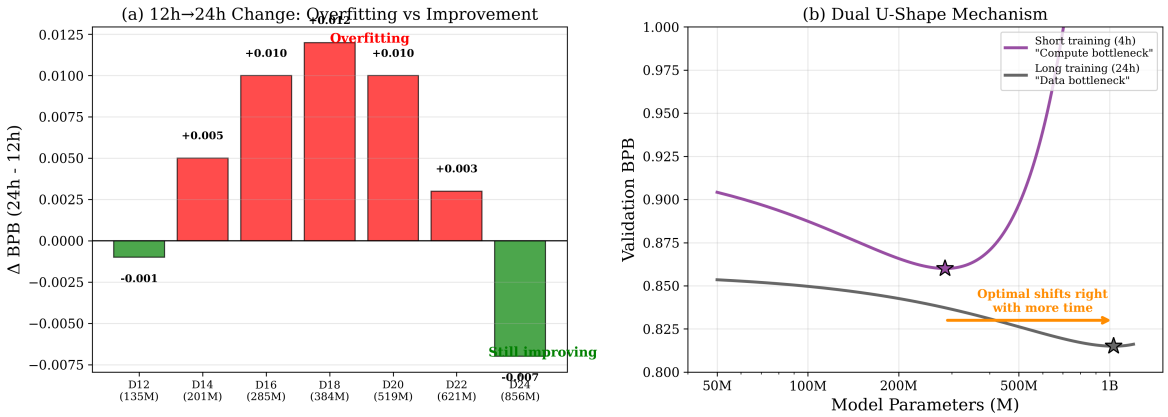


图5：**双重U形机制。(a)** 各模型从12小时到24小时的每字节比特数变化。模型 D14–D20 过拟合（红色柱，正值 Δ），而 D24–D26 仍在持续改善（绿色柱，负值 Δ）。**(b)** 双重U形的概念示意图。

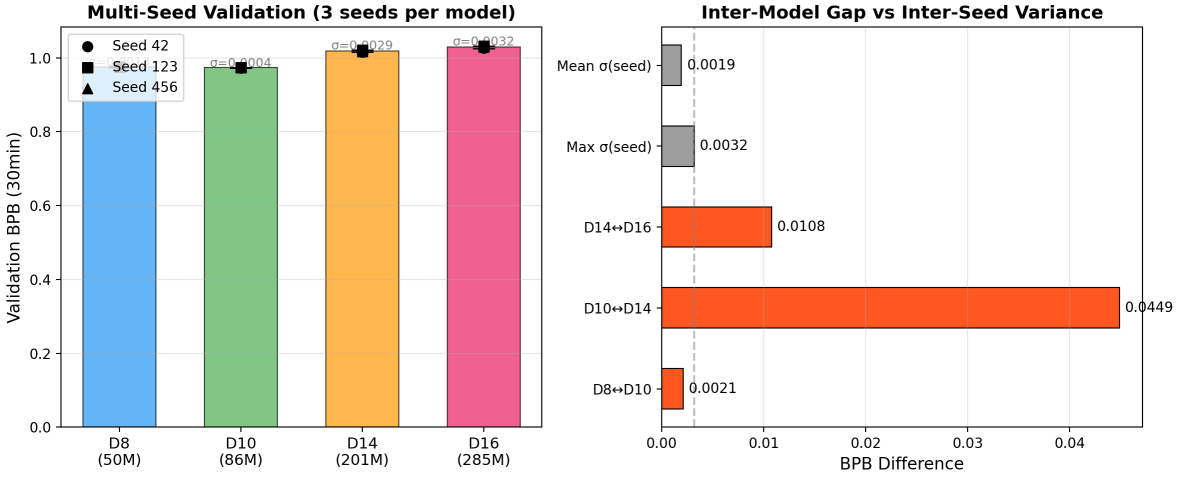


图 6：**多随机种子验证。(a)**每个模型使用 3 个随机种子训练 30分钟后的验证 BPB。误差棒表示 $\pm 1\sigma$。各随机种子的数值（标记）几乎完全重合。**(b)** 模型间差距（红色）与种子间方差（灰色）。模型选择的影响比随机种子大几个数量级。