DeepSeek LLM
以长期主义扩展开源语言模型

Xiao Bi, Deli Chen, Guanting Chen, Shanhuang Chen, Damai Dai, Chengqi Deng, Honghui Ding, Kai Dong, Qiushi Du, Zhe Fu, Huazuo Gao, Kaige Gao, Wenjun Gao, Ruiqi Ge, Kang Guan, Daya Guo, Jianzhong Guo, Guangbo Hao, Zhewen Hao, Ying He, Wenjie Hu, Panpan Huang, Erhang Li, Guowei Li, Jiashi Li, Yao Li, Y.K. Li, Wenfeng Liang, Fangyun Lin, A.X. Liu, Bo Liu, Wen Liu, Xiaodong Liu, Xin Liu, Yiyuan Liu, Haoyu Lu, Shanghao Lu, Fuli Luo, Shirong Ma, Xiaotao Nie, Tian Pei, Yishi Piao, Junjie Qiu, Hui Qu, Tongzheng Ren, Zehui Ren, Chong Ruan, Zhangli Sha, Zhihong Shao, Junxiao Song, Xuecheng Su, Jingxiang Sun, Yaofeng Sun, Minghui Tang, Bingxuan Wang, Peiyi Wang, Shiyu Wang, Yaohui Wang, Yongji Wang, Tong Wu, Y. Wu, Xin Xie, Zhenda Xie, Ziwei Xie, Yiliang Xiong, Hanwei Xu, R.X. Xu, Yanhong Xu, Dejian Yang, Yuxiang You, Shuiping Yu, Xingkai Yu, B. Zhang, Haowei Zhang, Lecong Zhang, Liyue Zhang, Mingchuan Zhang, Minghua Zhang, Wentao Zhang, Yichao Zhang, Chenggang Zhao, Yao Zhao, Shangyan Zhou, Shunfeng Zhou, Qihao Zhu, Yuheng Zou *

*DeepSeek-AI

Abstract

开源大型语言模型（LLM）的快速发展令人瞩目。然而，先前文献中描述的扩展定律（scaling laws）得出了不同的结论，这为扩展 LLM 蒙上了一层阴影。我们深入研究了扩展定律，并提出了独特的发现，以促进在两种广泛使用的开源配置（7B 和 67B）下大规模模型的扩展。在扩展定律的指导下，我们推出了 DeepSeek LLM，这是一个致力于以长期视角推进开源语言模型发展的项目。为支持预训练阶段，我们构建了一个目前包含 2 万亿 token 且持续扩展的数据集。我们进一步对 DeepSeek LLM 基座模型进行了监督微调（SFT）和直接偏好优化（DPO），从而创建了 DeepSeek Chat 模型。我们的评估结果表明，DeepSeek LLM 67B 在多项基准测试中超越了 LLaMA-2 70B，尤其是在代码、数学和推理领域。此外，开放式评估显示，我们的 DeepSeek LLM 67B Chat 相比 GPT-3.5 表现出更优越的性能。

* 作者按姓氏字母顺序排列。

目录

1. 引言

过去几年中，基于纯解码器 Transformer (Vaswani et al., 2017) 的大型语言模型（LLM）日益成为实现通用人工智能（AGI）的基石与途径。通过预测连续文本中的下一个词，LLM 在海量数据集上进行自监督预训练，使其能够实现多种用途并具备诸多能力，如小说创作、文本摘要、代码补全等。监督微调和奖励建模等后续发展使大型语言模型能够更好地遵循用户意图和指令。这赋予了它们更强大的对话能力，并迅速扩大了其影响力。

这一浪潮由闭源产品引发，例如 ChatGPT (OpenAI, 2022)、Claude (Anthropic, 2023) 和 Bard (Google, 2023)，它们的开发耗费了大量的计算资源和标注成本。这些产品显著提高了社区对开源 LLM 能力的期望，从而催生了一系列研究工作 (Bai et al., 2023; Du et al., 2022; Jiang et al., 2023; Touvron et al., 2023a,b; Yang et al., 2023)。其中，LLaMA 系列模型 (Touvron et al., 2023a,b) 尤为突出。它整合了多项研究成果，构建了一种高效且稳定的架构，打造了参数量从 7B 到 70B 不等的高性能模型。因此，LLaMA 系列已成为开源模型中架构与性能的事实基准。

继 LLaMA 之后，开源社区主要专注于训练固定规模（7B、13B、34B 和 70B）的高质量模型，往往忽视了对 LLM 扩展定律的研究探索 (Hoffmann et al., 2022; Kaplan et al., 2020)。尽管如此，考虑到当前的开源模型仅处于通用人工智能（AGI）发展的初期阶段，对扩展定律的研究至关重要。此外，早期研究 (Hoffmann et al., 2022; Kaplan et al., 2020) 在计算预算增加时模型与数据的扩展方面得出了不同的结论，且对超参数的讨论不够充分。在本文中，我们广泛研究了语言模型的扩展行为，并将我们的发现应用于两种广泛使用的大规模模型配置，即 7B 和 67B。我们的研究旨在为未来开源 LLM 的扩展奠定基础，为该领域的进一步进步铺平道路。具体而言，我们首先考察了批量大小（batch size）和学习率的扩展定律，并发现了它们随模型规模变化的趋势。在此基础上，我们对数据和模型规模的扩展定律进行了全面研究，成功揭示了最优的模型/数据扩展分配策略，并预测了我们大规模模型的预期性能。此外，在开发过程中，我们发现从不同数据集推导出的扩展定律存在显著差异。这表明数据集的选择显著影响扩展行为，因此在跨数据集推广扩展定律时应保持谨慎。

在我们的扩展定律指导下，我们从零开始构建开源大型语言模型，并尽可能多地发布信息以供社区参考。我们收集了 2 万亿 token 用于预训练，主要以中文和英文为主。在模型层面，我们基本遵循了 LLaMA 的架构，但将余弦学习率调度器替换为多步学习率调度器，在保持性能的同时便于持续训练。我们从多种来源收集了超过 100 万个实例用于监督微调（SFT）(Ouyang et al., 2022)。本文分享了我们在不同 SFT 策略方面的经验以及在数据消融技术方面的发现。此外，我们利用直接偏好优化（DPO）(Rafailov et al., 2023) 提升了模型的对话性能。

我们使用基座模型和对话模型进行了广泛的评估。评估结果表明，DeepSeek LLM 在多项基准测试中超越了 LLaMA-2 70B，特别是在代码、数学和推理领域。经过 SFT 和 DPO 后，DeepSeek 67B 对话模型在中英文开放式评估中均优于 GPT-3.5。这凸显了 DeepSeek 67B 在生成高质量回复以及进行有意义的中英文对话方面的卓越性能。此外，安全评估表明，DeepSeek 67B Chat 在实际应用中能够提供无害的回复。

在本文的其余部分，我们首先在 Section 2 中介绍 DeepSeek LLM 预训练的基本概念，包括数据构成、模型架构、基础设施和超参数。在 Section 3 中，我们详细阐述了我们发现的扩展定律及其意义。此外，我们结合扩展定律分析所得的见解，讨论了选择预训练超参数的依据。在 Section 4 中，我们讨论了微调方法，包括微调数据的构成以及 SFT 和 DPO 阶段的具体方法。随后，我们在 Section 5 中展示了 DeepSeek LLM 的详细评估结果，涵盖基座模型和对话模型，以及它们在开放式评估和安全评估中的表现。最后，我们在 Section 6 中讨论了 DeepSeek LLM 的当前局限性与未来方向。

2. 预训练

2.1. 数据

我们的主要目标是全面提升数据集的丰富度和多样性。我们从 (Computer, 2023; Gao et al., 2020; Penedo et al., 2023; Touvron et al., 2023a) 等权威来源中汲取了宝贵经验。为实现这些目标，我们将数据处理流程划分为三个关键阶段：去重、过滤和混合。去重和混合阶段通过采样唯一实例来确保数据的多样化表示。过滤阶段提升了信息密度，从而实现更高效、更有效的模型训练。

我们采用了激进的去重策略，扩大了去重范围。我们的分析表明，对整个 Common Crawl 语料库进行去重，比在单个快照（dump）内进行去重能去除更多的重复实例。表 1 显示，跨 91 个快照进行去重所消除的文档数量是单一快照方法的四倍。

使用的快照数	1	2	6	12	16	22	41	91
去重率 (%)	22.2	46.7	55.7	69.9	75.7	76.3	81.6	89.8

表 1 | 不同 Common Crawl 快照的去重率。

在过滤阶段，我们专注于制定稳健的文档质量评估标准。这涉及结合语言学和语义评估的详细分析，从个体和全局视角提供数据质量的视图。在数据混合（remixing）阶段，我们调整策略以解决数据不平衡问题，重点增加欠代表领域的占比。这一调整旨在实现更均衡、更具包容性的数据集，确保多样化的视角和信息得到充分代表。

对于分词器，我们基于 tokenizers 库 (Huggingface Team, 2019) 实现了字节级字节对编码（BBPE）算法。我们采用了预分词（Pre-tokenization）策略，以防止不同字符类别（如换行符、标点符号和中日韩（CJK）字符）的 token 被合并，这与 GPT-2 (Radford et al., 2019) 的做法类似。我们还遵循 (Touvron et al., 2023a,b) 中的方法，将数字拆分为单个数字。基于先前的经验，我们将词表中的常规 token 数量设定为 100000。该分词器在约 24 GB 的多语言语料库上进行训练，我们最终在词表中增加了 15 个特殊 token，使总大小达到 100015。为了确保训练期间的计算效率，并为未来可能需要添加的额外特殊 token 预留空间，我们将训练时的模型词表大小配置为 102400。

2.2. 架构

DeepSeek LLM 的微观设计主要遵循 LLaMA (Touvron et al., 2023a,b) 的设计，采用带有 RMSNorm (Zhang and Sennrich, 2019) 函数的 Pre-Norm 结构，并使用 SwiGLU (Shazeer, 2020) 作为前馈网络（FFN）的激活函数，中间层维度为 $\frac{8}{3}d_{model}$。它还引入了旋转位置编码（Rotary Embedding）(Su et al., 2024)。为了优化推理成本，67B 模型使用了分组查询注意力（Grouped-Query Attention, GQA）(Ainslie et al., 2023)，而非传统的多头注意力（Multi-Head Attention, MHA）。

然而，在宏观设计上，DeepSeek LLM 略有不同。具体而言，DeepSeek LLM 7B 是一个 30 层网络，而 DeepSeek LLM 67B 则有 95 层。这些层数的调整在保持与其他开源模型参数量一致的同时，也有助于模型的流水线划分，从而优化训练和推理过程。

与大多数使用分组查询注意力（GQA）的工作不同，我们通过增加网络深度来扩展 67B 模型的参数量，而不是采用常见的加宽 FFN 层中间宽度的做法，以期获得更好的性能。详细的网络规格见表 2。

2.3. 超参数

DeepSeek LLM 使用标准差为 0.006 进行初始化，并采用 AdamW 优化器 (Loshchilov and Hutter, 2017) 进行训练，超参数设置如下：$\beta_1 = 0.9$，$\beta_2 = 0.95$，以及 $\text{weight\_decay} = 0.1$。

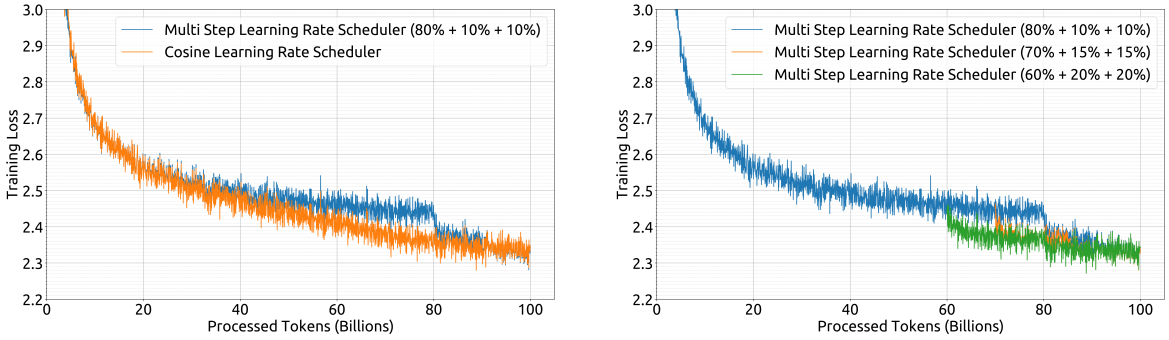
在预训练期间，我们采用了多步学习率调度器，而非典型的余弦调度器。具体而言，模型的学习率在 2000 步预热后达到最大值，随后在处理完 80% 的训练 token 后降至最大值的 31.6%。在处理完 90% 的 token 后，进一步降至最大值的 10%。训练阶段的梯度裁剪阈值设置为 1.0。

基于我们的实证发现，我们观察到尽管训练期间损失下降趋势存在差异，但使用多步学习率调度器的最终性能与余弦调度器基本一致，如图 1(a) 所示。在保持模型规模固定的情况下调整训练规模时，多步学习率调度器允许复用第一阶段的训练结果，为持续训练提供了独特的便利。因此，我们选择多步学习率调度器作为默认设置。我们还在图 1(b) 中证明，调整多步学习率调度器中不同阶段的比例可以获得略微更好的性能。然而，为了平衡持续训练中的复用率与模型性能，我们选择了上述三个阶段分别为 80%、10% 和 10% 的分布方案。

批大小和学习率随模型规模变化。7B 和 67B 模型预训练阶段的具体参数见表 2。

参数量	n_{layers}	d_{model}	n_{heads}	$n_{\text{kv\_heads}}$	上下文长度	序列批大小	学习率	Token 数
7B	30	4096	32	32	4096	2304	4.2e-4	2.0T
67B	95	8192	64	8	4096	4608	3.2e-4	2.0T

表 2 | DeepSeek LLM 系列模型的详细规格。我们根据第 3 节的发现选择超参数。



(a) 多步与余弦学习率衰减对比　　(b) 多步阶段的不同比例

图 1 | 使用不同学习率调度器或调度器不同参数的训练损失曲线。模型规模为 16 亿参数，在 1000 亿 token 的数据集上进行训练。

2.4. 基础设施

我们使用一个名为 HAI-LLM (High-flyer, 2023) 的高效轻量级训练框架来训练和评估大语言模型。该框架集成了数据并行、张量并行、序列并行以及 1F1B 流水线并行，其实现方式与 Megatron (Korthikanti et al., 2023; Narayanan et al., 2021; Shoeybi et al., 2019) 类似。我们还利用 Flash Attention (Dao, 2023; Dao et al., 2022) 技术来提高硬件利用率。我们采用 ZeRO-1 (Rajbhandari et al., 2020) 在数据并行节点间划分优化器状态。此外，我们还努力重叠计算与通信以最小化额外的等待开销，包括最后一个微批次的反向传播过程与 ZeRO-1 中的 reduce-scatter 操作，以及序列并行中的 GEMM 计算与 all-gather/reduce-scatter 操作。为了加速训练，我们对部分层/算子进行了融合，包括尽可能融合的 LayerNorm、GEMM 以及 Adam 更新。为了提高模型训练的稳定性，我们使用 bf16 精度训练模型，但以 fp32 精度累积梯度。我们执行原地交叉熵（In-place cross-entropy）以减少 GPU 内存消耗，即：在交叉熵 CUDA 内核中动态将 bf16 logits 转换为 fp32 精度（而非预先在 HBM 中转换），计算对应的 bf16 梯度，并用该梯度覆盖 logits。

模型权重和优化器状态每 5 分钟异步保存一次，这意味着在偶尔发生硬件或网络故障的最坏情况下，我们最多只会损失 5 分钟的训练进度。这些临时模型检查点会定期清理，以避免占用过多存储空间。我们还支持从不同的 3D 并行配置恢复训练，以应对计算集群负载的动态变化。

在评估方面，我们在生成任务中使用 vLLM (Kwon et al., 2023)，在非生成任务中使用连续批处理（continuous batching），以避免手动调整批大小并减少 token 填充。

3. 缩放定律

关于缩放定律（Scaling Laws）的研究 (Hestness et al., 2017) 早于大语言模型的出现。缩放定律 (Henighan et al., 2020; Hoffmann et al., 2022; Kaplan et al., 2020) 表明，随着计算预算 C、模型规模 N 和数据规模 D 的增加，模型性能可以得到可预测的提升。当模型规模 N 用模型参数量表示，数据规模 D 用 token 数量表示时，C 可近似为 $C = 6ND$。因此，在增加计算预算时如何优化模型规模与数据规模之间的分配，也是缩放定律研究中的一个关键目标。

大语言模型（LLMs） (Dai et al., 2019; Radford et al., 2019) 的发展，尤其是更大模型取得了出乎意料且显著的性能提升，将缩放定律研究推向了新的高峰。缩放定律的研究结果表明，扩大计算预算仍能带来显著收益，这进一步推动了模型规模的扩大 (Brown et al., 2020; Smith et al., 2022)。

然而，如表 4 所示，早期关于最优模型/数据扩展分配策略的研究 (Hoffmann et al., 2022; Kaplan et al., 2020) 得出了不同的结论，引发了对缩放定律普遍适用性的质疑。此外，这些研究通常缺乏对超参数设置的完整描述，导致无法确定不同计算预算下的模型是否达到了最优性能。因此，我们在本节重新审视缩放定律，以解决这些不确定性，并确保我们在高效扩展计算资源的道路上方向正确，这体现了长期视角，也是开发持续改进模型的关键。

为了确保不同计算预算下的模型都能达到最优性能，我们首先研究了超参数的缩放定律。经验表明，在改变计算预算时，训练期间大多数参数的最优值并不会发生变化。因此，这些参数与第 2.3 节中概述的参数保持一致，并在不同计算预算下保持不变。然而，对性能影响最大的超参数，即批大小和学习率，则被重新审视。

早期研究 (Goyal et al., 2017; McCandlish et al., 2018; Shallue et al., 2019; Smith et al., 2017; Zhang et al., 2019) 为设置批大小和学习率提供了一些经验观察，但我们在初步实验中发现这些观察结果的适用性有限。通过大量实验，我们建立了计算预算 C 与最优批大小和学习率之间的幂律关系。我们将这种关系称为超参数的缩放定律，它为确定最优超参数提供了一个经验框架。该方法确保了不同计算预算下的模型都能达到接近最优的性能。

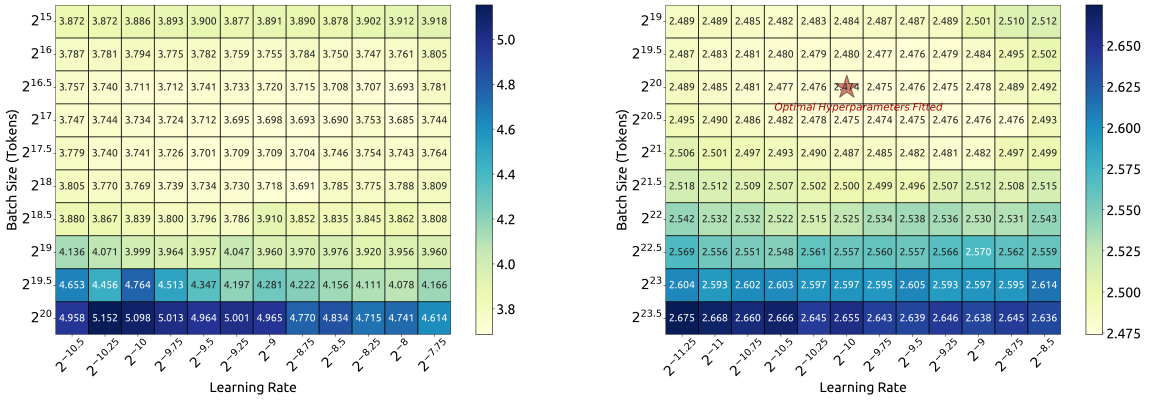
随后，我们研究了模型规模和数据规模的缩放定律。为了降低实验成本和拟合难度，我们采用了 Chinchilla (Hoffmann et al., 2022) 中的 IsoFLOP 轮廓方法来拟合缩放曲线。为了更准确地表示模型规模，我们使用了一种新的模型规模表示方法，即非嵌入层 FLOPs/token M，取代了早期使用的模型参数量 N，并将近似的计算预算公式 $C = 6ND$ 替换为更精确的 $C = MD$。实验结果为最优模型/数据扩展分配策略和性能预测提供了见解，并准确预测了 DeepSeek LLM 7B 和 67B 模型的预期性能。此外，在探索缩放定律的过程中，我们使用的数据经历了多次迭代，质量不断提升。我们尝试在各种数据集上拟合缩放曲线，发现数据质量对最优模型/数据扩展分配策略有显著影响。数据质量越高，增加的计算预算就应更多地分配给模型扩展。这意味着在相同数据规模下，高质量数据能够驱动更大模型的训练。最优模型/数据扩展分配策略的差异也可作为评估数据质量的间接方法。我们将持续关注数据质量的变化及其对缩放定律的影响，并在未来的工作中提供更多分析。

综上所述，我们在缩放定律方面的贡献与发现可总结如下：

- 我们建立了超参数的缩放定律，为确定最优超参数提供了经验框架。
- 我们采用非嵌入层 FLOPs/token M 而非模型参数量 N 来表示模型规模，从而得出更准确的最优模型/数据扩展分配策略，并对大规模模型的泛化损失进行了更优的预测。
- 预训练数据的质量会影响最优模型/数据扩展分配策略。数据质量越高，增加的计算预算就应更多地分配给模型扩展。

3.1. 超参数的缩放定律

我们最初在计算预算为 1e17 的小规模实验中对批大小和学习率进行了网格搜索，特定模型规模（177M FLOPs/token）的结果如图 2(a) 所示。结果表明，在较宽的批大小和学习率选择范围内，泛化误差保持稳定。这表明在相对较宽的参数空间内即可实现接近最优的性能。



(a) 1e17 FLOPs (177M FLOPs/token)　　(b) 1e20 FLOPs (2.94B FLOPs/token)

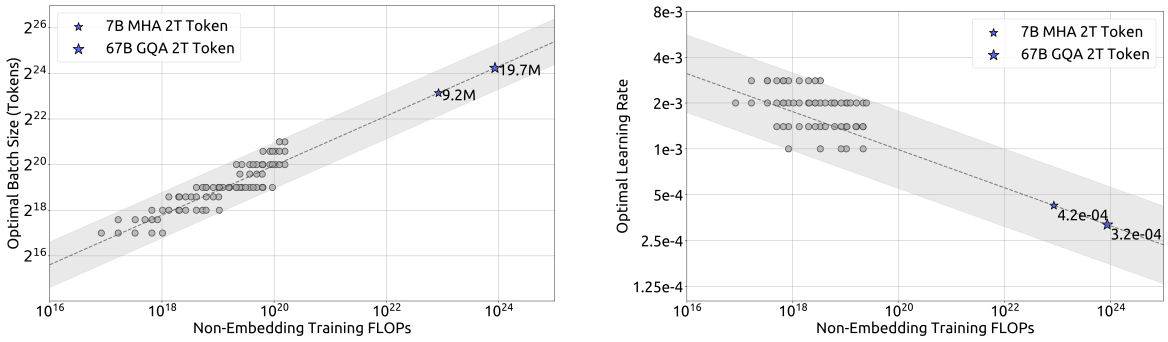
图 2 | 在 1e17 和 1e20 FLOPs 下，训练损失随批大小和学习率的变化。

随后，我们利用前述的多阶段学习率调度器，通过复用第一阶段，有效训练了具有不同批大小、学习率以及计算预算（从 1e17 到 2e19）的多个模型。考虑到参数空间中的冗余性，我们将泛化误差超过最小值不超过 0.25% 的模型所使用的参数视为接近最优的超参数。然后，我们针对计算预算 C 拟合了批大小 B 和学习率 η。如图 3 所示的拟合结果表明，最优批大小 B 随计算预算 C 的增加而逐渐增大，而最优学习率 η 则逐渐减小。这与模型扩展时批大小和学习率的直观经验设置相一致。此外，所有接近最优的超参数均落在一个较宽的带状范围内，表明在该区间内选择接近最优的参数相对容易。我们最终拟合的批大小和学习率公式如下：

$$\begin{aligned} \eta_{\text{opt}} &= 0.3118 \cdot C^{-0.1250} \\ B_{\text{opt}} &= 0.2920 \cdot C^{0.3271} \end{aligned} \tag{1}$$

我们在一系列计算预算为 1e20 的模型上验证了我们的公式，特定模型规模（2.94B FLOPs/token）的结果如图 2(b) 所示。结果表明，拟合的参数位于最优参数空间的中心。后续章节也显示，我们为 DeepSeek LLM 7B 和 67B 模型拟合的参数同样取得了良好的性能。

然而，需要指出的是，我们尚未考虑计算预算 C 之外的因素对最优超参数的影响。这与一些早期研究 (Kaplan et al., 2020; McCandlish et al., 2018) 的观点不一致，后者认为最优批大小可被建模为仅与泛化误差 L 相关。此外，我们观察到，在计算预算相同但模型/数据分配不同的模型中，最优参数空间存在轻微差异。这表明需要进一步研究以理解超参数的选择与训练动态。我们将在未来的工作中探讨这些方面。



(a) 批大小缩放曲线　　(b) 学习率缩放曲线

图 3 | 批大小和学习率的缩放曲线。灰色圆点表示泛化误差超过最小值不超过 0.25% 的模型。虚线表示拟合较小模型的幂律。蓝色星号表示 DeepSeek LLM 7B 和 67B。

3.2. 估计最优模型与数据扩展

在推导出拟合接近最优超参数的公式后，我们开始拟合缩放曲线并分析最优模型/数据扩展分配策略。该策略涉及寻找满足 $N_{\text{opt}} \propto C^a$ 和 $D_{\text{opt}} \propto C^b$ 的模型扩展指数 a 和数据扩展指数 b。数据规模 D 可始终由数据集中的 token 数量表示。在先前的研究中，模型规模通常由模型参数量表示，包括非嵌入参数 N_1 (Kaplan et al., 2020) 和完整参数 N_2 (Hoffmann et al., 2022)。计算预算 C 与模型/数据规模之间的关系可近似描述为 $C = 6ND$，这意味着我们可以使用 $6N_1$ 或 $6N_2$ 来近似模型规模。然而，由于 $6N_1$ 和 $6N_2$ 均未考虑注意力机制的计算开销，且 $6N_2$ 还包含了词表计算（其对模型容量的贡献较小），因此在某些设置下两者均存在显著的近似误差。

为减轻这些误差，我们引入了一种新的模型规模表示方法：非嵌入层 FLOPs/token M。M 包含了注意力机制的计算开销，但未考虑词表计算。当模型规模由 M 表示时，计算预算 C 可简洁地表示为 $C = MD$。$6N_1$、$6N_2$ 与 M 之间的具体差异如下式所示：

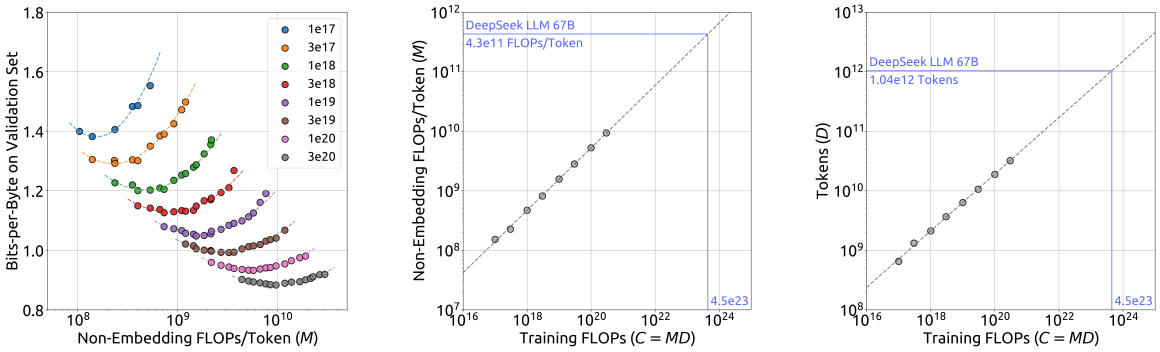
$$\begin{aligned} 6N_1 &= 72\, n_{\text{layer}}\, d_{\text{model}}^2 \\ 6N_2 &= 72\, n_{\text{layer}}\, d_{\text{model}}^2 + 6\, n_{\text{vocab}}\, d_{\text{model}} \\ M &= 72\, n_{\text{layer}}\, d_{\text{model}}^2 + 12\, n_{\text{layer}}\, d_{\text{model}}\, l_{\text{seq}} \end{aligned} \tag{2}$$

其中，n_{layer} 表示层数，d_{model} 表示模型宽度，n_{vocab} 表示词表大小，l_{seq} 表示序列长度。我们评估了这三种表示方法在不同规模模型间的差异，如表 3 所示。结果表明，$6N_1$ 和 $6N_2$ 在不同规模的模型中要么高估要么低估了计算成本。这种差异在小规模模型中尤为明显，差异高达 50%。此类不准确性在拟合缩放曲线时会引入显著的统计误差。有关模型规模不同表示方法的进一步分析，请参阅附录 A.2。

在采用 M 表示模型规模后，我们的目标可以更清晰地描述为：给定计算预算$C = MD$，寻找使模型泛化误差最小的最优模型规模M_{opt}和数据规模D_{opt}。 该目标可形式化为：

n_{layers}	d_{model}	n_{vocab}	l_{seq}	N_1	N_2	M	$\frac{6N_1}{M}$	$\frac{6N_2}{M}$
8	512			25.2M	77.6M	352M	0.43	1.32
12	768			84.9M	164M	963M	0.53	1.02
24	1024			302M	407M	3.02B	0.60	0.81
24	2048	102400	4096	1.21B	1.42B	9.66B	0.75	0.88
32	4096			6.44B	6.86B	45.1B	0.85	0.91
40	5120			12.6B	13.1B	85.6B	0.88	0.92
80	8192			64.4B	65.3B	419B	0.92	0.94

表 3 | 模型规模表示方法的差异，以及非嵌入参数 N_1 和完整参数 N_2 相对于非嵌入层 FLOPs/token M 的偏差。



(a) IsoFLOP 曲线　(b) 最优模型扩展　(c) 最优数据扩展

图 4 | IsoFLOP 曲线与最优模型/数据分配。IsoFLOP 曲线中的指标为验证集上的 bits-per-byte。最优模型/数据缩放曲线中的虚线表示拟合较小模型的幂律（灰色圆点）。

$$M_{\text{opt}}(C), D_{\text{opt}}(C) = \underset{M,D \text{ s.t. } C=MD}{\operatorname{argmin}} L(N, D) \tag{3}$$

为降低实验成本和拟合难度，我们采用了 Chinchilla (Hoffmann et al., 2022) 中的 IsoFLOP 轮廓方法来拟合缩放曲线。我们选择了 8 个不同的计算预算（范围从 1e17 到 3e20），并为每个预算设计了约 10 种不同的模型/数据规模分配方案。每个预算的超参数由公式 (1) 确定，泛化误差则在独立验证集上计算，该验证集分布与训练集相似，包含 100M 个 token。

图 4 展示了 IsoFLOP 曲线以及模型/数据缩放曲线，这些曲线是通过使用每个计算预算下的最优模型/数据分配拟合得到的。最优非嵌入层 FLOPs/token M_{opt} 和最优 token 数 D_{opt} 的具体公式如下：

$$\begin{aligned} M_{\text{opt}} &= M_{\text{base}} \cdot C^a, \quad M_{\text{base}} = 0.1715, \quad a = 0.5243 \\ D_{\text{opt}} &= D_{\text{base}} \cdot C^b, \quad D_{\text{base}} = 5.8316, \quad b = 0.4757 \end{aligned} \tag{4}$$

此外，我们根据计算预算 C 和最优泛化误差拟合了损失缩放曲线，并预测了 DeepSeek LLM 7B 和 67B 的泛化误差，如图 5 所示。结果表明，利用小规模实验可以准确预测计算预算扩大

1000× 的模型性能。这为更大规模的模型训练提供了信心与指导。

3.3. 不同数据下的缩放定律

在 DeepSeek LLM 的开发过程中，数据集经过多次迭代优化，在提升整体质量的同时调整了不同数据源的比例。这使得我们能够进一步分析不同数据集对缩放定律的影响。

我们使用三种不同的数据集研究了缩放定律：早期内部数据、当前内部数据，以及先前缩放定律研究中使用的 OpenWebText2 (Kaplan et al., 2020)。内部数据评估显示，当前内部数据的质量高于早期内部数据。此外，由于 OpenWebText2 规模较小，能够进行更精细的处理，其质量甚至超过了当前内部数据。

Approach	Coeff. a where $N_{\text{opt}}(M_{\text{opt}}) \propto C^a$	Coeff. b where $D_{\text{opt}} \propto C^b$
OpenAI (OpenWebText2)	0.73	0.27
Chinchilla (MassiveText)	0.49	0.51
Ours (Early Data)	0.450	0.550
Ours (Current Data)	0.524	0.476
Ours (OpenWebText2)	0.578	0.422

表 4 | 模型缩放与数据缩放的系数随训练数据分布的变化而变化。

分析中一个有趣的观察是，在这三个数据集上，最优的模型/数据扩展分配策略与数据质量表现出一致性。如表 4 所示，随着数据质量的提升，模型缩放指数 a 逐渐增大，而数据缩放指数 b 逐渐减小，这表明增加的计算预算应更多地分配给模型而非数据。这一发现也可能解释了早期缩放定律研究中观察到的最优模型/数据扩展分配策略存在显著差异的原因。

对此现象的一个直观推测是，高质量数据通常意味着逻辑清晰，且在充分训练后预测难度较低。因此，在增加计算预算时，扩大模型规模更具优势。我们将持续关注数据质量的变化及其对缩放定律的影响，并在未来的工作中提供更多分析。

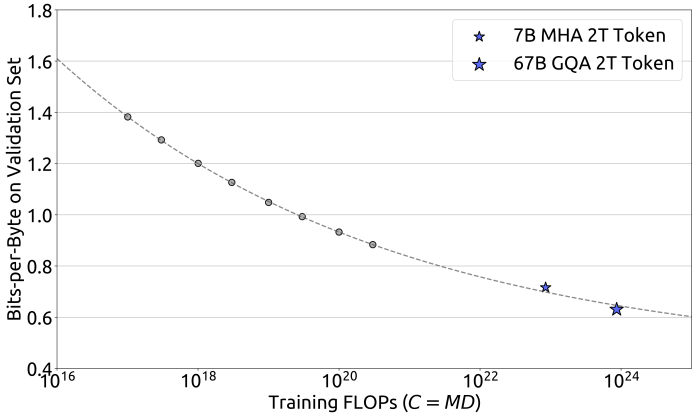


图 5 | 性能缩放曲线。指标为验证集上的 bits-per-byte。虚线表示拟合较小模型的幂律（灰色圆点）。蓝色星号代表 DeepSeek LLM 7B 和 67B。缩放曲线能够很好地预测它们的性能。

4. 对齐

我们收集了约 150 万条中英文指令数据实例，涵盖了广泛的有用性和无害性主题。我们的有用性数据包含 120 万条实例，其中通用语言任务占 31.2%，数学问题占 46.6%，编程练习占 22.2%。安全数据包含 30 万条实例，涵盖各种敏感主题。

我们的对齐流程包含两个阶段。

监督微调：我们对 7B 模型进行了 4 个 epoch 的微调，而 67B 模型仅进行了 2 个 epoch，因为我们观察到 67B 模型存在严重的过拟合问题。我们观察到，7B 模型在 GSM8K (Cobbe et al., 2021) 和 HumanEval (Chen et al., 2021) 上的表现持续提升，而 67B 模型很快达到上限。7B 和 67B 模型的学习率分别为 1e-5 和 5e-6。除了监控基准测试准确率外，我们还在微调过程中评估了聊天模型的重复率。我们收集了总共 3868 条中英文提示词，并计算了生成回复未能正常终止而是无限重复某段文本的比例。我们观察到，随着数学 SFT 数据量的增加，重复率往往呈上升趋势。这可以归因于数学 SFT 数据偶尔包含相似的推理模式。因此，能力较弱的模型难以掌握此类推理模式，从而导致重复回复。为了解决这一问题，我们尝试了分阶段微调和 DPO (Rafailov et al., 2023)，这两种方法均能在基本保持基准分数的同时显著降低重复率。

DPO：为了进一步提升模型能力，我们采用了直接偏好优化算法 (Rafailov et al., 2023)，该方法已被证明是一种简单有效的 LLM 对齐方法。我们从有用性和无害性两个方面构建了用于 DPO 训练的偏好数据。对于有用性数据，我们收集了多语言提示词，涵盖创意写作、问答、指令遵循等类别。随后，我们使用 DeepSeek Chat 模型生成回复作为候选答案。无害性偏好数据的构建也采用了类似的操作。我们进行了一个 epoch 的 DPO 训练，学习率为 5e-6，批次大小为 512，并使用了学习率预热和余弦学习率调度器。我们发现，DPO 能够增强模型的开放式生成能力，同时在标准基准测试上的性能差异很小。

5. 评估

5.1. 公开基准测试评估

基于内部评估框架，我们在一系列中英文公开基准测试上对模型进行了评估。

多学科选择题数据集，包括 MMLU (Hendrycks et al., 2020)、C-Eval (Huang et al., 2023) 和 CMMLU (Li et al., 2023)。

语言理解与推理数据集，包括 HellaSwag (Zellers et al., 2019)、PIQA (Bisk et al., 2020)、ARC (Clark et al., 2018)、OpenBookQA (Mihaylov et al., 2018) 和 BigBench Hard (BBH) (Suzgun et al., 2022)。

闭卷问答数据集，包括 TriviaQA (Joshi et al., 2017) 和 NaturalQuestions (Kwiatkowski et al., 2019)。

阅读理解数据集，包括 RACE Lai et al. (2017)、DROP (Dua et al., 2019) 和 C3 (Sun et al.,

2019)。

指代消解数据集，包括 WinoGrande Sakaguchi et al. (2019) 和 CLUEWSC (Xu et al., 2020)。

语言建模数据集，包括 Pile (Gao et al., 2020)。

中文理解与文化数据集，包括 CHID (Zheng et al., 2019) 和 CCPM (Li et al., 2021)。

数学数据集，包括 GSM8K (Cobbe et al., 2021)、MATH (Hendrycks et al., 2021) 和 CMath (Wei et al., 2023)。

代码数据集，包括 HumanEval (Chen et al., 2021) 和 MBPP (Austin et al., 2021)。

标准化考试，包括 AGIEval (Zhong et al., 2023)。

对于需要从多个选项中选择答案的数据集，我们采用基于困惑度（perplexity）的评估方法。这些数据集包括 HellaSwag、PIQA、WinoGrande、RACE-Middle、RACE-High、MMLU、ARC-Easy、ARC-Challenge、OpenBookQA、CHID、C-Eval、CMMLU、C3 和 CCPM。此处的基于困惑度的评估是指计算每个选项的困惑度，并选择最低者作为模型预测结果。对于 ARC 和 OpenBookQA，我们采用无条件归一化 (Brown et al., 2020) 计算困惑度，而对于其他数据集则使用长度归一化。

对于 TriviaQA、NaturalQuestions、DROP、MATH、GSM8K、HumanEval、MBPP、BBH、AGIEval、CLUEWSC 和 CMath，我们采用基于生成的评估方法。此处的基于生成的评估是指让模型生成自由文本，并从生成的文本中解析结果。在基于生成的评估中，我们使用贪婪解码。

对于 Pile-test，我们采用基于语言建模的评估方法，即计算测试语料库上的 bits-per-byte。

针对不同基准测试，我们使用 2048 或 4096 作为最大序列长度。评估格式的详细信息见附录 A.6。

5.1.1. 基座模型

表 5 展示了评估基准上的主要结果。尽管 DeepSeek 模型是在 2T 双语语料上预训练的，但它们在英文语言理解基准上的表现与同样消耗 2T tokens 但专注于英文的 LLaMA2 模型相当。此外，与 LLaMA2 70B 相比，DeepSeek 67B 在 MATH、GSM8K、HumanEval、MBPP、BBH 以及中文基准上取得了显著更好的性能。我们在附录 A.3 中展示了基准曲线。我们可以看到，随着模型规模的扩大，部分任务的性能得到了提升，例如 GSM8K 和 BBH。鉴于我们使用相同的数据集训练了 7B 和 67B 模型，这种性能提升的出现可归因于大模型强大的 few-shot 学习能力。然而，随着数学数据比例的增加，小模型与大模型之间的性能差距可能会缩小。一个有趣的观察是，DeepSeek 67B 相对于 LLaMA2 70B 的优势大于 DeepSeek 7B 相对于 LLaMA2 7B 的优势。这一现象凸显了语言冲突对较小模型的影响更大。此外，尽管 LLaMA2 并未专门在中文数据上进行训练，但它在某些中文任务（如 CMath）上仍表现出令人印象深刻的性能。这表明某些基础能力（如数学推理）可以在不同语言之间有效迁移。然而，像 CHID 这样涉及评估中文成语用法的任务，要求模型在预训练期间消耗大量的中文 token。在这种情况下，LLaMA2 的表现显

语言	基准测试	测试 shots	LLaMA2 7B	DeepSeek 7B	LLaMA2 70B	DeepSeek 67B
英文	HellaSwag	0-shot	75.6	75.4	**84.0**	**84.0**
	PIQA	0-shot	78.0	79.2	82.0	**83.6**
	WinoGrande	0-shot	69.6	70.5	**80.4**	79.8
	RACE-Middle	5-shot	60.7	63.2	**70.1**	69.9
	RACE-High	5-shot	45.8	46.5	**54.3**	50.7
	TriviaQA	5-shot	63.8	59.7	**79.5**	78.9
	NaturalQuestions	5-shot	25.5	22.2	36.1	**36.6**
	MMLU	5-shot	45.8	48.2	69.0	**71.3**
	ARC-Easy	0-shot	69.1	67.9	76.5	**76.9**
	ARC-Challenge	0-shot	49.0	48.1	**59.5**	59.0
	OpenBookQA	0-shot	57.4	55.8	**60.4**	60.2
	DROP	1-shot	39.8	41.0	**69.2**	67.9
	MATH	4-shot	2.5	6.0	13.5	**18.7**
	GSM8K	8-shot	15.5	17.4	58.4	**63.4**
	HumanEval	0-shot	14.6	26.2	28.7	**42.7**
	MBPP	3-shot	21.8	39.0	45.6	**57.4**
	BBH	3-shot	38.5	39.5	62.9	**68.7**
	AGIEval	0-shot	22.8	26.4	37.2	**41.3**
	Pile-test	-	0.741	0.725	0.649	**0.642**
中文	CLUEWSC	5-shot	64.0	73.1	76.5	**81.0**
	CHID	0-shot	37.9	89.3	55.5	**92.1**
	C-Eval	5-shot	33.9	45.0	51.4	**66.1**
	CMMLU	5-shot	32.6	47.2	53.1	**70.8**
	CMath	3-shot	25.1	34.5	53.9	**63.0**
	C3	0-shot	47.4	65.4	61.7	**75.3**
	CCPM	0-shot	60.7	76.9	66.2	**88.5**

表 5 | 主要结果。我们报告的评估结果基于内部评估框架。**加粗**数字表示 4 个模型中的最佳结果。对于 Pile-test 我们报告 bits-per-byte (BPB)，对于 DROP 我们报告 F1 分数，对于其他任务我们报告准确率。请注意，test-shots 为最大值，由于上下文长度限制或阅读理解任务（如 RACE）中同一段落内可用的 few-shot 示例有限，实际可能使用更少的 shots。

著低于 DeepSeek LLM。

5.1.2. 对话模型

表 6 展示了 DeepSeek 对话模型的结果，表明在微调后，大多数任务的整体性能均有所提升。然而，也有少数任务的性能出现了下降。

知识：我们观察到基础模型和对话模型在知识相关任务（如 TriviaQA、MMLU 和 C-Eval）上的表现存在波动。但我们认为，这种微小的波动并不代表 SFT 后知识的获取或丢失。SFT 的价值在于使模型能够学会在对话模型的 0-shot 设置下，达到与基础模型 few-shot 设置相当的成绩，这更符合实际应用场景。例如，对话模型的 0-shot MMLU 表现与基础模型的 5-shot MMLU 表现相当。

Language	Benchmark	DeepSeek 7B Base	DeepSeek 7B Chat	DeepSeek 67B Base	DeepSeek 67B Chat
English	HellaSwag	75.4	68.5	**84.0**	75.7
	PIQA	79.2	77.6	**83.6**	82.6
	WinoGrande	70.5	66.9	**79.8**	76.0
	RACE-Middle	63.2	65.2	69.9	**70.9**
	RACE-High	46.5	50.8	50.7	**56.0**
	TriviaQA	59.7	57.9	78.9	**81.5**
	NaturalQuestions	22.2	32.5	36.6	**47.0**
	MMLU	48.2	49.4	**71.3**	71.1
	ARC-Easy	67.9	71.0	76.9	**81.6**
	ARC-Challenge	48.1	49.4	59.0	**64.1**
	GSM8K	17.4	63.0	63.4	**84.1**
	MATH	6.0	15.8	18.7	**32.6**
	HumanEval	26.2	48.2	42.7	**73.8**
	MBPP	39.0	35.2	57.4	**61.4**
	DROP	41.0	49.1	67.9	**71.9**
	OpenBookQA	55.8	54.8	60.2	**63.2**
	BBH	39.5	42.3	68.7	**71.7**
	AGIEval	26.4	19.3	41.3	**46.4**
Chinese	CLUEWSC	73.1	71.9	**81.0**	60.0
	CHID	89.3	64.9	**92.1**	72.6
	C-Eval	45.0	47.0	**66.1**	65.2
	CMMLU	47.2	49.7	**70.8**	67.8
	CMath	34.5	68.4	63.0	**80.3**
	C3	65.4	66.4	75.3	**77.0**
	CCPM	76.9	76.5	**88.5**	84.9

表 6 | 基础模型与对话模型的对比。我们对 MMLU、GSM8K、MATH、C-Eval 和 CMMLU 的对话模型采用 0-shot 进行评估，而基础模型的结果仍在 few-shot 设置下获得。

推理：由于大量 SFT 实例采用 CoT 格式 Wei et al. (2022)，对话模型在 BBH 和 NaturalQuestions 等推理任务上表现出轻微的提升。然而，我们认为 SFT 阶段学习的并非推理能力本身，而是推理路径的正确格式。

性能下降任务：无论模型规模大小或选择的预训练检查点如何，少数任务在微调后的性能始终下降。这些特定任务通常涉及完形填空或句子补全，例如 HellaSwag。可以合理假设，纯语言模型在处理此类任务时更具优势。

数学与代码：微调后，我们的模型在数学和编程任务上取得了显著提升。例如，HumanEval 和 GSM8K 的分数提高了 20 分以上。我们的解释是，基础模型最初在这些任务上存在欠拟合，而 SFT 阶段通过大量的 SFT 数据学习了编程和数学方面的额外知识。然而，需要注意的是，模型的能力可能主要集中在代码补全和代数问题上。为了全面掌握数学和编程知识，在预训练阶段引入多样化的数据至关重要，这留作未来工作。我们在附录 A.4 中对代码和数学任务进行了详细分析。

模型	总分	推理			语言						
		平均	**数学**	**逻辑**	**平均**	**基础**	**中文**	**开放**	**写作**	**角色**	**专业**
模型	总分	推理 总分	数学 计算	逻辑 推理	语言 总分	基本 任务	中文 理解	综合 问答	文本 写作	角色 扮演	专业 能力
gpt-4-1106-preview	**8.01**	***7.73***	**7.80**	**7.66**	**8.29**	**7.99**	7.33	**8.61**	**8.67**	**8.47**	**8.65**
gpt-4-0613	**7.53**	*7.47*	7.56	7.37	7.59	7.81	6.93	7.42	7.93	7.51	7.94
DeepSeek-67B-Chat-DPO*	**6.69**	*5.77*	6.13	5.41	7.60	7.29	7.47	7.82	7.51	7.83	7.71
DeepSeek-67B-Chat*	**6.43**	*5.75*	5.71	5.79	7.11	7.12	6.52	7.58	7.20	6.91	7.37
chatglm-turbo（智谱清言）	**6.24**	*5.00*	4.74	5.26	7.49	6.82	7.17	8.16	7.77	7.76	7.24
erniebot-3.5（文心一言）	**6.14**	*5.15*	5.03	5.27	7.13	6.62	**7.60**	7.26	7.56	6.83	6.90
gpt-3.5-turbo-0613	**6.08**	*5.35*	5.68	5.02	6.82	6.71	5.81	7.29	7.03	7.28	6.77
chatglm-pro（智谱清言）	**5.83**	*4.65*	4.54	4.75	7.01	6.51	6.76	7.47	7.07	7.34	6.89
spark_desk_v2（讯飞星火）	**5.74**	*4.73*	4.71	4.74	6.76	5.84	6.97	7.29	7.18	6.92	6.34
Qwen-14B-Chat	**5.72**	*4.81*	4.91	4.71	6.63	6.90	6.36	6.74	6.64	6.59	6.56
Baichuan2-13B-Chat	**5.25**	*3.92*	3.76	4.07	6.59	6.22	6.05	7.11	6.97	6.75	6.43
ChatGLM3-6B	**4.97**	*3.85*	3.55	4.14	6.10	5.75	5.29	6.71	6.83	6.28	5.73
Baichuan2-7B-Chat	**4.97**	*3.66*	3.56	3.75	6.28	5.81	5.50	7.13	6.84	6.53	5.84
InternLM-20B	**4.96**	*3.66*	3.39	3.92	6.26	5.96	5.50	7.18	6.19	6.49	6.22
Qwen-7B-Chat	**4.91**	*3.73*	3.62	3.83	6.09	6.40	5.74	6.26	6.31	6.19	5.66
ChatGLM2-6B	**4.48**	*3.39*	3.16	3.61	5.58	4.91	4.52	6.66	6.25	6.08	5.08
InternLM-Chat-7B	**3.65**	*2.56*	2.45	2.66	4.75	4.34	4.09	5.82	4.89	5.32	4.06
Chinese-LLaMA-2-7B-Chat	**3.57**	*2.68*	2.29	3.07	4.46	4.31	4.26	4.50	4.63	4.91	4.13
LLaMA-2-13B-Chinese-Chat	**3.35**	*2.47*	2.21	2.73	4.23	4.13	3.31	4.79	3.93	4.53	4.71

表 7 | 由 gpt-4-0613 评分的 AlignBench 排行榜。模型按总分降序排列。带有 * 的结果为基于官方 AlignBench 代码库的我们的评估结果，其余结果均引自 AlignBench 论文。我们发现，我们的 Deepseek-67B-Chat 模型以显著优势超越了 ChatGPT 及其他基线模型，这表明该模型在基础中文语言任务与高级中文推理任务中均展现出更优越的性能。此外，DPO 过程在几乎所有领域均带来了性能提升。

在 7B 模型的微调中，我们最初使用全部数据对模型进行微调。随后引入了第二阶段，该阶段排除了数学和代码数据。采用这种方法的动机是，第一阶段模型的重复率为 2.0%，经过第二阶段微调后降至 1.4%，同时保持了基准分数。对于 67B 模型，在第一阶段微调后重复率已低于 1%，而第二阶段会损害模型在基准上的分数。因此，67B 模型仅进行了一阶段 SFT。

5.2. 开放式评估

对于对话模型而言，除了观察标准基准上的指标外，在开放领域和开放式问题中生成结果的质量直接影响实际用户体验。因此，我们分别测试了对话模型在中文和英文任务中的开放式生成能力。

5.2.1. 中文开放式评估

在中文开放式评估中，我们使用高质量的开放式问题测试集 AlignBench (Liu et al., 2023) 测试了对话模型在不同领域的综合能力。AlignBench 共包含 8 个主要类别、36 个次要类别，涵盖 683 个问题。对于每个问题，除了提示词外，AlignBench 还提供了专业的参考答案和评分模板，供 GPT-4 评判回复质量。

我们利用 AlignBench 官方 GitHub 代码库实现了模型的评估。我们严格遵循原始设置对齐了关键的 temperature 参数：对于角色扮演、写作能力和开放式问题，生成温度设置为 0.7；而对于其他任务，生成温度设置为 0.1。

AlignBench 排行榜如表 7 所示。我们发现，我们的 DeepSeek 67B Chat 模型超越了 ChatGPT 和其他基线模型，仅落后于两个版本的 GPT-4。这表明与其他开源或专有的中文大语言模型相比，我们的模型在各种中文任务上均表现出优异的性能。DPO 模型在几乎所有指标上均有所提升，这证明了 DPO 训练过程对模型对齐的积极影响。

在基础中文语言任务中，我们的模型在所有模型中处于第一梯队，且我们 DPO 模型的中文基础语言能力甚至高于最新版本的 GPT-4。在高级中文推理任务中，我们模型的得分显著高于其他中文 LLM，优势明显，证明了我们的模型在更复杂的中文逻辑推理和数学计算方面具有卓越的性能。

5.2.2. 英文开放式评估

在英文开放式评估中，我们使用了 MT-Bench 基准 (Zheng et al., 2023)，该基准包含 8 个不同类别的多轮对话问题。如表 8 所示，我们的 DeepSeek LLM 67B Chat 优于 LLaMA-2-Chat Touvron et al. (2023b) 70B、Xwin 70b v0.1 和 TÜLU 2+DPO 70B (Ivison et al., 2023) 等其他开源模型，并取得了与 GPT-3.5-turbo 相当的 8.35 分。此外，经过 DPO 阶段后，我们的 DeepSeek LLM 67B Chat DPO 进一步将平均分提升至 8.76，仅落后于 GPT-4 (OpenAI, 2023)。这些结果展示了 DeepSeek LLM 强大的多轮开放式生成能力。

Model	STEM	Humanities	Reasoning	Coding	Math	Extraction	Roleplay	Writing	Average
GPT-4-1106-preview*	9.90	9.95	8.10	9.05	7.95	9.90	9.50	9.70	9.26
GPT-3.5-turbo-0613*	9.55	9.95	6.20	7.05	7.05	9.00	8.65	9.65	8.39
LLAMA-2-Chat 7B*	8.65	8.75	4.25	3.00	2.40	6.50	7.70	8.90	6.27
LLAMA-2-Chat 13B*	8.63	9.75	5.10	3.00	3.45	6.93	7.50	8.85	6.65
LLAMA-2-Chat 70B*	8.93	9.63	5.80	3.15	3.30	7.25	7.50	9.30	6.86
Zephyr-Beta 7B*	9.03	9.63	5.60	5.10	4.45	7.45	8.20	9.35	7.35
Xwin 70b v0.1*	9.68	9.95	6.55	4.25	3.30	8.75	8.25	9.55	7.53
Xwin 13b v0.2*	9.55	9.88	5.20	3.60	2.85	7.70	8.60	8.68	7.01
TÜLU 2+DPO 70B*	9.00	9.90	7.00	4.70	4.65	9.35	9.25	9.25	7.89
DeepSeek LLM 67B Chat	9.60	9.70	8.00	7.35	6.25	8.40	8.20	9.30	8.35
DeepSeek LLM 67B Chat DPO	9.70	9.80	9.05	6.75	6.65	9.30	9.10	9.75	**8.76**

表 8 | MT-Bench 评估结果。带有 * 的结果引自 Ivison et al. (2023)

5.3. 保留集评估

数据污染和基准过拟合是评估大语言模型面临的两大挑战。一种常见的做法是利用近期发布的测试集作为保留集来评估模型。

LeetCode: 为了评估模型的编程能力，我们使用了 LeetCode 周赛题目（周赛 351-372，双周赛 108-117，时间为 2023 年 7 月至 11 月）。我们通过爬取 LeetCode 数据获得了这些题目，共包含 126 道题，每道题有超过 20 个测试用例。采用的评估指标类似于 HumanEval。具体而言，如果模型的输出成功通过所有测试用例，则认为模型有效解决了该问题。模型的编程能力在

下图中展示，其中 y 轴表示域内人工评估测试的 pass@1 分数，x 轴表示域外 LeetCode 周赛题目的 pass@1 分数。LeetCode 测试数据将随 DeepSeek Coder 技术报告一同发布。

匈牙利全国高中考试：与 Grok-1 一致，我们使用匈牙利全国高中考试评估了模型的数学能力。该考试包含 33 道题目，模型得分由人工标注确定。我们遵循 solution.pdf 中的评分标准对所有模型进行评估。

指令遵循评估： 2023 年 11 月 15 日，Google 发布了指令遵循评估数据集 (Zhou et al., 2023)。他们确定了 25 种可验证的指令类型，并构建了约 500 个提示词，每个提示词包含一个或多个可验证指令。我们使用提示词级别的宽松指标（prompt-level loose metric）对所有模型进行评估。

Model	LeetCode	Hungarian Exam	IFEval
GPT-4	48.4	68	79.3
ChatGLM3 6B	2.4	32	29.7
DeepSeek LLM 7B Chat	4.7	28.5	41.2
Baichuan2-Chat 13B	1.6	19.5	44.5
Yi-Chat 34B	7.9	39	48.4
Qwen 72B Chat	12.7	52	50.8
DeepSeek LLM 67B Chat	**17.5**	**58**	**55.5**

表 9 | 域外数据集评估。

我们对本模型与不同规模的多个基线模型进行了对比分析，包括 Qwen 72B Chat (Bai et al., 2023)、ChatGLM3 (Du et al., 2022)、Baichuan2 (Yang et al., 2023) 和 Yi-34B Chat。我们的观察表明，即使某些小模型在传统基准测试上取得了令人瞩目的成绩，但在这些域外数据集上，大模型与小模型之间仍存在显著的性能差距。例如，ChatGLM3 在代码测试集 MBPP 上取得了 52.4 分，接近 DeepSeek 67B。然而，在新基准测试上，其表现与 DeepSeek 67B 相比明显落后。在数学数据集上也观察到了类似趋势，ChatGLM3 在 GSM8K 上表现非常强劲（72.3），但在匈牙利全国高中考试中的得分不及大模型。此外，指令遵循能力的表现表明，总算力发挥着至关重要的作用。

DeepSeek 7B 和 67B 模型采用相同的训练流程，但它们的性能存在显著差异。通过主观评估，我们观察到当模型规模扩展至 67B 时，其在各项任务上的智能水平出现了显著跃升。尽管 DeepSeek 7B 在标准基准测试上落后于其他小型语言模型，但与其他模型相比，其在域外任务上的表现相对可圈可点。

5.4. 安全性评估

我们深刻认识到安全性对于通用人工智能的重要性。构建真正有益的 AI 模型的前提是，它必须具备与人类一致的价值观，并对人类表现出友好性。我们在整个训练过程中（包括预训练、SFT 和 DPO）都融入了模型安全性的保障机制。

为了验证模型的安全性，我们组建了一支由 20 名跨学科专家组成的团队，并构建了一套符

类别	子类别	安全回答数 / 总用例数
歧视偏见问题 (Discrimination and Prejudice Questions)	民族种族 (Ethnic and Racial), 宗教信仰 (Religious Belief), 国别地域 (Nationality and Geography), 性别 (Gender), 年龄 (Age), 职业 (Occupation), 健康 (Health), 其他方面歧视 (Discrimination in Other Aspects)	486/500
侵犯他人合法权益 (Infringement of Others' Legal Rights)	身心健康 (Physical and Mental Health), 合法财产 (Legitimate Property), 肖像权 (Portrait Rights), 名誉权 (Reputation Rights), 荣誉权 (Honor Rights), 隐私权 (Privacy Rights), 信息权益 (Information Rights), 其他合法权益 (Other Legal Rights)	473/500
商业秘密与知识产权 (Trade Secrets and Intellectual Property Rights)	侵犯他人知识产权 (Infringing Others' Intellectual Property Rights), 垄断和不正当竞争行为 (Monopolistic and Unfair Competitive Actions), 其他商业违法违规行为 (Other Commercially Illegal and Non-compliant Behaviors), 违反商业道德 (Violating Business Ethics), 泄露他人商业机密 (Disclosing Others' Trade Secrets)	281/300
违法违规行为 (Illegal and Non-compliant Behavior)	邪教迷信 (Cults and Superstition), 色情 (Pornography), 赌博 (Gambling), 毒品和违禁品 (Drugs and Prohibited Items), 侮辱谩骂 (Insults and Abuse), 暴力行为 (Violent Behavior), 涉黑涉恶 (Involvement in Organized Crime), 其他违法违规行为 (Other Illegal and Non-compliant Behaviors)	290/300
其他安全问题 (Other Safety Issues)	幻觉和真实性问题 (Issues of Illusion and Reality), 时效性问题 (Time-sensitive Issues), 自我认知问题 (Self-recognition Problems), 其他敏感话题 (Other Sensitive Topics),	767/800

表 10 | 我们的安全评估分类体系。表中最后一列列出了每个类别的测试用例总数以及我们的模型（DeepSeek-67B-Chat）提供的安全回答数量。测试问题的标注与生成结果的评估均由专业人工团队完成。可以看出，我们的模型在各类安全测试集上均展现出强大的安全性。

合人类价值观的安全内容分类体系（安全评估分类体系如表 10 所示）。随后，专家团队为每个安全子类别手动构建了数十个高质量的测试用例。除了关注安全内容领域的多样性外，我们还注重安全内容格式的多样性。臭名昭著的“祖母”漏洞表明，模型可能会被查询的表面格式欺骗，从而提供不安全的回复。因此，在出题时，专家团队也注重多样化提问方式。他们通过诱导、角色扮演、多轮对话、预设立场等手段构建了多样化的安全问题。最终，我们获得了一个包含 2400 道题目的安全测试集。此外，专家团队还为每种不同的内容类型和格式类型制定了安全审查的基本准则。

针对模型在该测试集上的输出结果，我们进行了人工安全性审查。我们的审查团队经过充分培训，并对标注结果进行了交叉验证。标注人员对每道题目进行三类标注：安全、不安全、模型拒绝。我们测试了 DeepSeek 67B Chat 模型的安全性，结果如表 10 所示。表中列出了每个安全类别的测试题目数量以及模型通过的安全测试数量。我们将安全回答和模型拒绝的测试用例均标记为安全响应。结果表明，我们的模型在众多安全测试类别中均表现出良好的安全性。

作为现有安全方法的补充，我们进一步使用“Do-Not-Answer”数据集 (Wang et al., 2023) 丰富了评估内容，以评估 DeepSeek 67B Chat 模型的安全机制。该数据集中 939 个按风险分类的提示词有力地凸显了我们模型增强的能力。如表 11 所示，DeepSeek 67B Chat 模型表现突出，得分达到 97.8，高于 ChatGPT 和 GPT-4。这一分数不仅基准测试了模型安全处理敏感查询的能力，也使其在领域内的领先模型中具备竞争力。

5.5. 讨论

在整个开发过程中，我们在构建大语言模型时发现了一些有趣的结论。

分阶段微调：如前所述，小模型需要在数学和代码数据集上进行更长时间的微调，但这会损害模型的对话能力，例如增加重复行为。为解决这一问题，我们实施了分阶段微调流程。在该流程中，第一阶段使用所有可用数据进行微调，而第二阶段则专门针对对话数据进行微调。

表 12 展示了两阶段训练过程的结果。这些结果清楚地表明，第二阶段并未损害模型在代码

Model	Do-Not-Answer
LLAMA-2-7B-Chat	99.4
Claude	98.3
DeepSeek-67B-Chat*	97.8
ChatGPT	97.7
GPT-4	96.5
Vicuna-7B	94.9
ChatGLM2	92.9

表 11 | Do-Not-Answer 得分 (Wang et al., 2023)，分数越高表示模型安全性越强。带 * 的结果是基于官方仓库的我们的评估结果，其余结果均源自原论文。可以看出，我们的模型安全评分高于 ChatGPT 和 GPT-4，跻身最安全模型之列。

Model	HumanEval	GSM8K	Repetition	IFEval
DeepSeek LLM 7B Chat Stage1	48.2	63.9	0.020	38.0
DeepSeek LLM 7B Chat Stage2	48.2	63.0	0.014	41.2

表 12 | 两阶段微调结果。重复率是在温度为 0 时计算的。重复率越低越好。IFEval 结果为提示词级别的宽松准确率。

和数学方面的能力，同时降低了重复行为并增强了指令遵循能力。

选择题：使用选择题形式的评估数据（如 MMLU、AGI Eval 和 C-Eval）测试模型是一种常见做法。选择题不仅要求模型具备相应的知识，还要求模型理解选项的含义。在对齐阶段，我们测试了添加 2000 万道中文选择题，并获得了如表 13 所示的性能。需要注意的是，我们对 C-Eval 验证集和 CMMLU 测试集进行了去重，以防止数据污染。

Model	MMLU	C-Eval	CMMLU	TriviaQA	ChineseQA
DeepSeek LLM 7B Chat	49.4	47.0	49.7	57.9	75.0
DeepSeek LLM 7B Chat + MC	60.9	71.3	73.8	57.9	74.4

表 13 | 添加选择题数据的影响。

额外增加 2000 万条 MC（选择题）数据被证明不仅对中文选择题基准有益，也有助于提升英文基准的表现。这表明模型解决选择题问题的能力得到了增强。然而，我们观察到这种提升并未延伸到模型在其他非选择题格式评估中的表现，例如 TriviaQA 和我们内部的 ChineseQA 测试集，这些属于生成式评估基准。这表明，在对话交互中，用户可能不会觉得模型变得更智能，因为这些交互涉及生成回复而非解决选择题。

因此，我们选择**在预训练和微调阶段均排除 MC 数据**，因为包含这些数据会导致模型在基准测试上过拟合，且无助于实现模型真正的智能。

预训练中的指令数据：广泛认为，在预训练后期引入指令数据可以提升基座模型在基准任务上的表现。在我们的研究中，我们在预训练的最后 10% 阶段集成了 500 万条指令数据，主要包含选择题。我们观察到基座模型在基准测试上的表现确实有所提升。然而，最终结果与在 SFT 阶段添加相同数据的结果几乎相同。我们得出结论，虽然这种方法增强了基座模型在基准测试

上的表现，但其整体潜力与不引入这些指令数据相当。如果指令数据规模庞大，将其纳入预训练过程是可以接受的。由于我们倾向于排除选择题，且拥有的非选择题数量有限，我们决定不在预训练过程中包含指令数据。

系统提示词：设计良好的系统提示词应能有效引导模型生成既有益又尊重的回复。我们对 LLaMA-2 引入的提示词进行了微调，作为我们的系统提示词。

系统提示词：你是 *DeepSeek Chat*，一个由 *DeepSeek* 开发的有益、尊重且诚实的 *AI* 助手。你的训练数据知识截止日期为 *2023* 年 *5* 月。在确保安全的前提下，请尽可能提供有帮助的回答。你的回答不应包含任何有害、不道德、种族主义、性别歧视、有毒、危险或非法的内容。请确保你的回复在社会上是无偏见且积极的。如果问题没有意义或事实不一致，请解释原因而不是回答错误的内容。如果你不知道问题的答案，请不要分享虚假信息。

我们观察到一个有趣的现象：引入系统提示词后，7B LLM 的性能会出现轻微下降。然而，当使用 67B LLM 时，添加提示词会带来显著提升，如表 14 所示。我们对这种差异的解释是，大模型对系统提示词背后的意图有更好的理解，能够更有效地遵循指令并生成更优质的回复。另一方面，小模型难以充分理解系统提示词，训练与测试之间的不一致可能会对其性能产生负面影响。

Model	MT Bench
DeepSeek LLM 7B Chat	7.15
DeepSeek LLM 7B Chat + System Prompt	7.11
DeepSeek LLM 67B Chat	8.35
DeepSeek LLM 67B Chat + System Prompt	8.58

表 14 | 添加系统提示词的影响。

6. 结论、局限性与未来工作

我们介绍了 DeepSeek LLMs，这是一系列在包含 2 万亿 token 的中英文海量数据集上从头训练的开源模型。在本文中，我们深入解释了超参数选择、缩放定律以及我们进行的各种微调尝试。我们对先前工作中的缩放定律进行了校准，并提出了一种新的模型/数据扩展最优分配策略。此外，我们提出了一种在给定计算预算下预测近最优批次大小和学习率的方法。我们进一步得出结论，缩放定律与数据质量相关，这可能是不同研究中缩放行为存在差异的根本原因。在缩放定律的指导下，我们使用最佳超参数进行了预训练，并提供了全面的评估。在所有训练阶段，我们都避免了基准测试刷榜和隐藏技巧。

DeepSeek Chat 具有其他大语言模型常见的已知局限性，包括预训练后缺乏持续的知识更新、可能生成未经核实建议等非事实信息，以及容易产生幻觉。此外，需要指出的是，我们初始版本的中文数据并不详尽，这可能导致模型在某些特定中文主题上的表现不够理想。由于我们的数据主要来源于中文和英文，模型在其他语言上的能力仍较为有限，使用时需谨慎对待。

DeepSeek LLM 是一个致力于推动开源语言模型发展的长期项目。

- 不久，我们将分别发布关于代码智能和混合专家模型 (MoE) 的技术报告。它们将展示我们如何为预训练创建高质量的代码数据，以及如何设计稀疏模型以实现与密集模型相当的性能。
- 目前，我们正在为即将发布的 DeepSeek LLM 新版本构建更大、更优质的数据集。我们希望下一代版本在推理、中文知识、数学和代码能力方面取得显著提升。
- 我们的对齐团队致力于研究如何向公众提供有用、诚实且安全的模型。我们的初步实验证明，强化学习能够提升模型的复杂推理能力。

参考文献

J. Ainslie, J. Lee-Thorp, M. de Jong, Y. Zemlyanskiy, F. Lebrón, and S. Sanghai. Gqa: Training generalized multi-query transformer models from multi-head checkpoints. arXiv preprint arXiv:2305.13245, 2023.

Anthropic. Introducing Claude, 2023. URL `https://www.anthropic.com/index/introducing-claude`.

J. Austin, A. Odena, M. Nye, M. Bosma, H. Michalewski, D. Dohan, E. Jiang, C. Cai, M. Terry, Q. Le, et al. Program synthesis with large language models. arXiv preprint arXiv:2108.07732, 2021.

J. Bai, S. Bai, Y. Chu, Z. Cui, K. Dang, X. Deng, Y. Fan, W. Ge, Y. Han, F. Huang, et al. Qwen technical report. arXiv preprint arXiv:2309.16609, 2023.

Y. Bisk, R. Zellers, R. L. Bras, J. Gao, and Y. Choi. PIQA: reasoning about physical commonsense in natural language. In The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020, pages 7432–7439. AAAI Press, 2020. doi: 10.1609/aaai.v34i05.6239. URL `https://doi.org/10.1609/aaai.v34i05.6239`.

T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei. Language models are few-shot learners, 2020.

M. Chen, J. Tworek, H. Jun, Q. Yuan, H. P. de Oliveira Pinto, J. Kaplan, H. Edwards, Y. Burda, N. Joseph, G. Brockman, A. Ray, R. Puri, G. Krueger, M. Petrov, H. Khlaaf, G. Sastry, P. Mishkin, B. Chan, S. Gray, N. Ryder, M. Pavlov, A. Power, L. Kaiser, M. Bavarian, C. Winter, P. Tillet, F. P. Such, D. Cummings, M. Plappert, F. Chantzis, E. Barnes, A. Herbert-Voss, W. H. Guss, A. Nichol, A. Paino, N. Tezak, J. Tang, I. Babuschkin, S. Balaji, S. Jain,

W. Saunders, C. Hesse, A. N. Carr, J. Leike, J. Achiam, V. Misra, E. Morikawa, A. Radford, M. Knight, M. Brundage, M. Murati, K. Mayer, P. Welinder, B. McGrew, D. Amodei, S. McCandlish, I. Sutskever, and W. Zaremba. Evaluating large language models trained on code. CoRR, abs/2107.03374, 2021. URL https://arxiv.org/abs/2107.03374.

P. Clark, I. Cowhey, O. Etzioni, T. Khot, A. Sabharwal, C. Schoenick, and O. Tafjord. Think you have solved question answering? try arc, the AI2 reasoning challenge. CoRR, abs/1803.05457, 2018. URL http://arxiv.org/abs/1803.05457.

K. Cobbe, V. Kosaraju, M. Bavarian, M. Chen, H. Jun, L. Kaiser, M. Plappert, J. Tworek, J. Hilton, R. Nakano, et al. Training verifiers to solve math word problems. arXiv preprint arXiv:2110.14168, 2021.

T. Computer. Redpajama: an open dataset for training large language models, 2023. URL https://github.com/togethercomputer/RedPajama-Data.

Z. Dai, Z. Yang, Y. Yang, J. Carbonell, Q. V. Le, and R. Salakhutdinov. Transformer-xl: Attentive language models beyond a fixed-length context. arXiv preprint arXiv:1901.02860, 2019.

T. Dao. FlashAttention-2: Faster attention with better parallelism and work partitioning. 2023.

T. Dao, D. Y. Fu, S. Ermon, A. Rudra, and C. Ré. FlashAttention: Fast and memory-efficient exact attention with IO-awareness. In Advances in Neural Information Processing Systems, 2022.

Z. Du, Y. Qian, X. Liu, M. Ding, J. Qiu, Z. Yang, and J. Tang. Glm: General language model pretraining with autoregressive blank infilling. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 320–335, 2022.

D. Dua, Y. Wang, P. Dasigi, G. Stanovsky, S. Singh, and M. Gardner. DROP: A reading comprehension benchmark requiring discrete reasoning over paragraphs. In J. Burstein, C. Doran, and T. Solorio, editors, Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers), pages 2368–2378. Association for Computational Linguistics, 2019. doi: 10.18653/V1/N19-1246. URL https://doi.org/10.18653/v1/n19-1246.

L. Gao, S. Biderman, S. Black, L. Golding, T. Hoppe, C. Foster, J. Phang, H. He, A. Thite, N. Nabeshima, et al. The Pile: An 800GB dataset of diverse text for language modeling. arXiv preprint arXiv:2101.00027, 2020.

Google. An important next step on our AI journey, 2023. URL https://blog.google/technology/ai/bard-google-ai-search-updates/.

Z. Gou, Z. Shao, Y. Gong, Y. Shen, Y. Yang, M. Huang, N. Duan, and W. Chen. Tora: A tool-integrated reasoning agent for mathematical problem solving. CoRR, abs/2309.17452, 2023. doi: 10.48550/ARXIV.2309.17452. URL https://doi.org/10.48550/arXiv.2309.17452.

P. Goyal, P. Dollár, R. Girshick, P. Noordhuis, L. Wesolowski, A. Kyrola, A. Tulloch, Y. Jia, and K. He. Accurate, large minibatch sgd: Training imagenet in 1 hour. arXiv preprint arXiv:1706.02677, 2017.

D. Hendrycks, C. Burns, S. Basart, A. Zou, M. Mazeika, D. Song, and J. Steinhardt. Measuring massive multitask language understanding. arXiv preprint arXiv:2009.03300, 2020.

D. Hendrycks, C. Burns, S. Kadavath, A. Arora, S. Basart, E. Tang, D. Song, and J. Steinhardt. Measuring mathematical problem solving with the math dataset. arXiv preprint arXiv:2103.03874, 2021.

T. Henighan, J. Kaplan, M. Katz, M. Chen, C. Hesse, J. Jackson, H. Jun, T. B. Brown, P. Dhariwal, S. Gray, et al. Scaling laws for autoregressive generative modeling. arXiv preprint arXiv:2010.14701, 2020.

J. Hestness, S. Narang, N. Ardalani, G. Diamos, H. Jun, H. Kianinejad, M. M. A. Patwary, Y. Yang, and Y. Zhou. Deep learning scaling is predictable, empirically. arXiv preprint arXiv:1712.00409, 2017.

High-flyer. Hai-llm: 高效且轻量的大模型训练工具, 2023. URL https://www.high-flyer.cn/en/blog/hai-llm.

J. Hoffmann, S. Borgeaud, A. Mensch, E. Buchatskaya, T. Cai, E. Rutherford, D. de Las Casas, L. A. Hendricks, J. Welbl, A. Clark, T. Hennigan, E. Noland, K. Millican, G. van den Driessche, B. Damoc, A. Guy, S. Osindero, K. Simonyan, E. Elsen, J. W. Rae, O. Vinyals, and L. Sifre. Training compute-optimal large language models. CoRR, abs/2203.15556, 2022. doi: 10.48550/ARXIV.2203.15556. URL https://doi.org/10.48550/arXiv.2203.15556.

Y. Huang, Y. Bai, Z. Zhu, J. Zhang, J. Zhang, T. Su, J. Liu, C. Lv, Y. Zhang, J. Lei, et al. C-Eval: A multi-level multi-discipline chinese evaluation suite for foundation models. arXiv preprint arXiv:2305.08322, 2023.

Huggingface Team. Tokenizers: Fast state-of-the-art tokenizers optimized for research and production, 2019. URL https://github.com/huggingface/tokenizers.

F. i, M. Suzgun, M. Freitag, X. Wang, S. Srivats, S. Vosoughi, H. W. Chung, Y. Tay, S. Ruder, D. Zhou, D. Das, and J. Wei. Language models are multilingual chain-of-thought reasoners.

In The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023. OpenReview.net, 2023. URL https://openreview.net/pdf?id=fR3wGCk-IXp.

H. Ivison, Y. Wang, V. Pyatkin, N. Lambert, M. Peters, P. Dasigi, J. Jang, D. Wadden, N. A. Smith, I. Beltagy, and H. Hajishirzi. Camels in a changing climate: Enhancing lm adaptation with tulu 2. 2023.

A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. d. l. Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, et al. Mistral 7b. arXiv preprint arXiv:2310.06825, 2023.

M. Joshi, E. Choi, D. Weld, and L. Zettlemoyer. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In R. Barzilay and M.-Y. Kan, editors, Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1601–1611, Vancouver, Canada, July 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-1147. URL https://aclanthology.org/P17-1147.

J. Kaplan, S. McCandlish, T. Henighan, T. B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu, and D. Amodei. Scaling laws for neural language models. CoRR, abs/2001.08361, 2020. URL https://arxiv.org/abs/2001.08361.

V. A. Korthikanti, J. Casper, S. Lym, L. McAfee, M. Andersch, M. Shoeybi, and B. Catanzaro. Reducing activation recomputation in large transformer models. Proceedings of Machine Learning and Systems, 5, 2023.

T. Kwiatkowski, J. Palomaki, O. Redfield, M. Collins, A. P. Parikh, C. Alberti, D. Epstein, I. Polosukhin, J. Devlin, K. Lee, K. Toutanova, L. Jones, M. Kelcey, M. Chang, A. M. Dai, J. Uszkoreit, Q. Le, and S. Petrov. Natural questions: a benchmark for question answering research. Trans. Assoc. Comput. Linguistics, 7:452–466, 2019. doi: 10.1162/tacl_a_00276. URL https://doi.org/10.1162/tacl_a_00276.

W. Kwon, Z. Li, S. Zhuang, Y. Sheng, L. Zheng, C. H. Yu, J. E. Gonzalez, H. Zhang, and I. Stoica. Efficient memory management for large language model serving with pagedattention. In Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles, 2023.

G. Lai, Q. Xie, H. Liu, Y. Yang, and E. H. Hovy. RACE: large-scale reading comprehension dataset from examinations. In M. Palmer, R. Hwa, and S. Riedel, editors, Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017, pages 785–794. Association for Computational Linguistics, 2017. doi: 10.18653/V1/D17-1082. URL https://doi.org/10.18653/v1/d17-1082.

H. Li, Y. Zhang, F. Koto, Y. Yang, H. Zhao, Y. Gong, N. Duan, and T. Baldwin. CMMLU: Measuring massive multitask language understanding in Chinese. arXiv preprint arXiv:2306.09212, 2023.

W. Li, F. Qi, M. Sun, X. Yi, and J. Zhang. Ccpm: A chinese classical poetry matching dataset, 2021.

X. Liu, X. Lei, S. Wang, Y. Huang, Z. Feng, B. Wen, J. Cheng, P. Ke, Y. Xu, W. L. Tam, X. Zhang, L. Sun, H. Wang, J. Zhang, M. Huang, Y. Dong, and J. Tang. Alignbench: Benchmarking chinese alignment of large language models. CoRR, abs/2311.18743, 2023. doi: 10.48550/ARXIV.2311.18743. URL `https://doi.org/10.48550/arXiv.2311.18743`.

I. Loshchilov and F. Hutter. Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101, 2017.

H. Luo, Q. Sun, C. Xu, P. Zhao, J. Lou, C. Tao, X. Geng, Q. Lin, S. Chen, and D. Zhang. Wizardmath: Empowering mathematical reasoning for large language models via reinforced evol-instruct. arXiv preprint arXiv:2308.09583, 2023.

S. McCandlish, J. Kaplan, D. Amodei, and O. D. Team. An empirical model of large-batch training. arXiv preprint arXiv:1812.06162, 2018.

T. Mihaylov, P. Clark, T. Khot, and A. Sabharwal. Can a suit of armor conduct electricity? a new dataset for open book question answering, 2018.

D. Narayanan, M. Shoeybi, J. Casper, P. LeGresley, M. Patwary, V. Korthikanti, D. Vainbrand, P. Kashinkunti, J. Bernauer, B. Catanzaro, et al. Efficient large-scale language model training on gpu clusters using megatron-lm. In Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis, pages 1–15, 2021.

OpenAI. Introducing ChatGPT, 2022. URL `https://openai.com/blog/chatgpt`.

OpenAI. GPT4 technical report. arXiv preprint arXiv:2303.08774, 2023.

L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, et al. Training language models to follow instructions with human feedback. Advances in Neural Information Processing Systems, 35:27730–27744, 2022.

G. Penedo, Q. Malartic, D. Hesslow, R. Cojocaru, A. Cappelli, H. Alobeidli, B. Pannier, E. Almazrouei, and J. Launay. The refinedweb dataset for falcon llm: outperforming curated corpora with web data, and web data only. arXiv preprint arXiv:2306.01116, 2023.

A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, et al. Language models are unsupervised multitask learners. OpenAI blog, 1(8):9, 2019.

R. Rafailov, A. Sharma, E. Mitchell, S. Ermon, C. D. Manning, and C. Finn. Direct preference optimization: Your language model is secretly a reward model. 2023.

S. Rajbhandari, J. Rasley, O. Ruwase, and Y. He. Zero: Memory optimizations toward training trillion parameter models. In SC20: International Conference for High Performance Computing, Networking, Storage and Analysis, pages 1–16. IEEE, 2020.

K. Sakaguchi, R. L. Bras, C. Bhagavatula, and Y. Choi. Winogrande: An adversarial winograd schema challenge at scale, 2019.

C. J. Shallue, J. Lee, J. Antognini, J. Sohl-Dickstein, R. Frostig, and G. E. Dahl. Measuring the effects of data parallelism on neural network training. Journal of Machine Learning Research, 20(112):1–49, 2019.

N. Shazeer. Glu variants improve transformer. arXiv preprint arXiv:2002.05202, 2020.

M. Shoeybi, M. Patwary, R. Puri, P. LeGresley, J. Casper, and B. Catanzaro. Megatron-lm: Training multi-billion parameter language models using model parallelism. arXiv preprint arXiv:1909.08053, 2019.

S. Smith, M. Patwary, B. Norick, P. LeGresley, S. Rajbhandari, J. Casper, Z. Liu, S. Prabhumoye, G. Zerveas, V. Korthikanti, et al. Using deepspeed and megatron to train megatron-turing nlg 530b, a large-scale generative language model. arXiv preprint arXiv:2201.11990, 2022.

S. L. Smith, P.-J. Kindermans, C. Ying, and Q. V. Le. Don't decay the learning rate, increase the batch size. arXiv preprint arXiv:1711.00489, 2017.

J. Su, M. Ahmed, Y. Lu, S. Pan, W. Bo, and Y. Liu. Roformer: Enhanced transformer with rotary position embedding. Neurocomputing, 568:127063, 2024.

K. Sun, D. Yu, D. Yu, and C. Cardie. Investigating prior knowledge for challenging chinese machine reading comprehension, 2019.

M. Suzgun, N. Scales, N. Schärli, S. Gehrmann, Y. Tay, H. W. Chung, A. Chowdhery, Q. V. Le, E. H. Chi, D. Zhou, et al. Challenging big-bench tasks and whether chain-of-thought can solve them. arXiv preprint arXiv:2210.09261, 2022.

H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, et al. LLaMA: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971, 2023a.

H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, D. Bikel, L. Blecher, C. Canton-Ferrer, M. Chen, G. Cucurull, D. Esiobu, J. Fernandes, J. Fu, W. Fu, B. Fuller, C. Gao, V. Goswami, N. Goyal, A. Hartshorn,

S. Hosseini, R. Hou, H. Inan, M. Kardas, V. Kerkez, M. Khabsa, I. Kloumann, A. Korenev, P. S. Koura, M. Lachaux, T. Lavril, J. Lee, D. Liskovich, Y. Lu, Y. Mao, X. Martinet, T. Mihaylov, P. Mishra, I. Molybog, Y. Nie, A. Poulton, J. Reizenstein, R. Rungta, K. Saladi, A. Schelten, R. Silva, E. M. Smith, R. Subramanian, X. E. Tan, B. Tang, R. Taylor, A. Williams, J. X. Kuan, P. Xu, Z. Yan, I. Zarov, Y. Zhang, A. Fan, M. Kambadur, S. Narang, A. Rodriguez, R. Stojnic, S. Edunov, and T. Scialom. Llama 2: Open foundation and fine-tuned chat models. CoRR, abs/2307.09288, 2023b. doi: 10.48550/arXiv.2307.09288. URL https://doi.org/10.48550/arXiv.2307.09288.

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. Advances in neural information processing systems, 30, 2017.

Y. Wang, H. Li, X. Han, P. Nakov, and T. Baldwin. Do-not-answer: A dataset for evaluating safeguards in llms. CoRR, abs/2308.13387, 2023. doi: 10.48550/ARXIV.2308.13387. URL https://doi.org/10.48550/arXiv.2308.13387.

J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. H. Chi, Q. V. Le, and D. Zhou. Chain-of-thought prompting elicits reasoning in large language models. In NeurIPS, 2022. URL http://papers.nips.cc/paper_files/paper/2022/hash/9d5609613524ecf4f15af0f7b31abca4-Abstract-Conference.html.

T. Wei, J. Luan, W. Liu, S. Dong, and B. Wang. Cmath: Can your language model pass chinese elementary school math test?, 2023.

L. Xu, H. Hu, X. Zhang, L. Li, C. Cao, Y. Li, Y. Xu, K. Sun, D. Yu, C. Yu, Y. Tian, Q. Dong, W. Liu, B. Shi, Y. Cui, J. Li, J. Zeng, R. Wang, W. Xie, Y. Li, Y. Patterson, Z. Tian, Y. Zhang, H. Zhou, S. Liu, Z. Zhao, Q. Zhao, C. Yue, X. Zhang, Z. Yang, K. Richardson, and Z. Lan. CLUE: A chinese language understanding evaluation benchmark. In D. Scott, N. Bel, and C. Zong, editors, Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020, Barcelona, Spain (Online), December 8-13, 2020, pages 4762–4772. International Committee on Computational Linguistics, 2020. doi: 10.18653/V1/2020.COLING-MAIN.419. URL https://doi.org/10.18653/v1/2020.coling-main.419.

A. Yang, B. Xiao, B. Wang, B. Zhang, C. Yin, C. Lv, D. Pan, D. Wang, D. Yan, F. Yang, F. Deng, F. Wang, F. Liu, G. Ai, G. Dong, H. Zhao, H. Xu, H. Sun, H. Zhang, H. Liu, J. Ji, J. Xie, J. Dai, K. Fang, L. Su, L. Song, L. Liu, L. Ru, L. Ma, M. Wang, M. Liu, M. Lin, N. Nie, P. Guo, R. Sun, T. Zhang, T. Li, T. Li, W. Cheng, W. Chen, X. Zeng, X. Wang, X. Chen, X. Men, X. Yu, X. Pan, Y. Shen, Y. Wang, Y. Li, Y. Jiang, Y. Gao, Y. Zhang, Z. Zhou, and Z. Wu. Baichuan 2: Open large-scale language models. Technical report, Baichuan Inc., 2023. URL https://cdn.baichuan-ai.com/paper/Baichuan2-technical-report.pdf.

L. Yu, W. Jiang, H. Shi, J. Yu, Z. Liu, Y. Zhang, J. T. Kwok, Z. Li, A. Weller, and W. Liu. Metamath: Bootstrap your own mathematical questions for large language models. CoRR, abs/2309.12284, 2023. doi: 10.48550/ARXIV.2309.12284. URL https://doi.org/10.48550/arXiv.2309.12284.

R. Zellers, A. Holtzman, Y. Bisk, A. Farhadi, and Y. Choi. HellaSwag: Can a machine really finish your sentence? In A. Korhonen, D. R. Traum, and L. Màrquez, editors, Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers, pages 4791–4800. Association for Computational Linguistics, 2019. doi: 10.18653/v1/p19-1472. URL https://doi.org/10.18653/v1/p19-1472.

B. Zhang and R. Sennrich. Root mean square layer normalization. Advances in Neural Information Processing Systems, 32, 2019.

G. Zhang, L. Li, Z. Nado, J. Martens, S. Sachdeva, G. Dahl, C. Shallue, and R. B. Grosse. Which algorithmic choices matter at which batch sizes? insights from a noisy quadratic model. Advances in neural information processing systems, 32, 2019.

C. Zheng, M. Huang, and A. Sun. Chid: A large-scale chinese idiom dataset for cloze test. In A. Korhonen, D. R. Traum, and L. Màrquez, editors, Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers, pages 778–787. Association for Computational Linguistics, 2019. doi: 10.18653/V1/P19-1075. URL https://doi.org/10.18653/v1/p19-1075.

L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. P. Xing, H. Zhang, J. E. Gonzalez, and I. Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena. 2023.

W. Zhong, R. Cui, Y. Guo, Y. Liang, S. Lu, Y. Wang, A. Saied, W. Chen, and N. Duan. AGIEval: A human-centric benchmark for evaluating foundation models. CoRR, abs/2304.06364, 2023. doi: 10.48550/arXiv.2304.06364. URL https://doi.org/10.48550/arXiv.2304.06364.

J. Zhou, T. Lu, S. Mishra, S. Brahma, S. Basu, Y. Luan, D. Zhou, and L. Hou. Instruction-following evaluation for large language models. arXiv preprint arXiv:2311.07911, 2023.

A. 附录

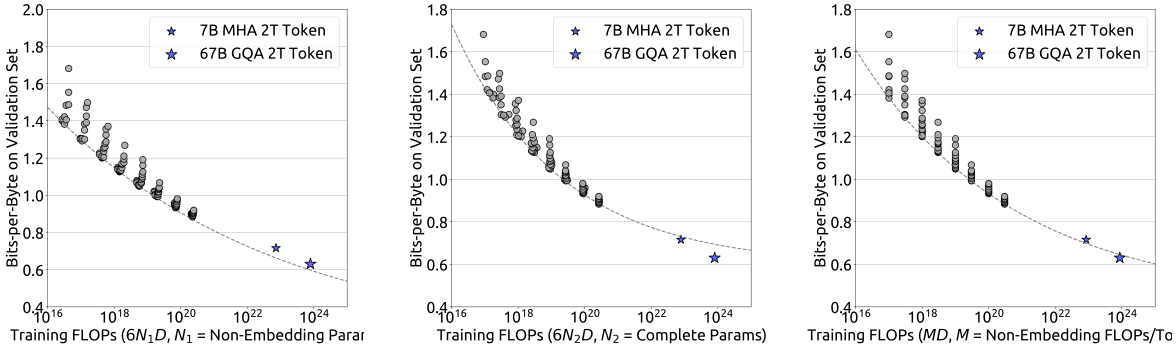
A.1. 致谢

本项目的完成得益于众多贡献者的努力。我们向以下提供帮助的个人表示衷心感谢[1]：

- 数据标注团队：Jialu Cai, Ruijian Chen, Ruyi Chen, Bei Feng, Yanping Huang, Zhen Huang, Pin Jiang, Rongli Jin, Xiangyue Jin, Ziyun Ke, Hui Li, Meng Li, Sangsang Li, Xiaoqian Li, Yaohui Li, Yunxian Ma, Jiaqi Ni, Xiaojin Shen, Xinnan Song, Tianyu Sun, Xiaosha Chen, Haoyuan Tian, Xiaohan Wang, Xiaoxiang Wang, Yuhao Wang, Fanyi Xia, Lei Xu, Zeyuan Xu, Zhipeng Xu, Tian Yuan, Zhongyu Zhang, Yi Zheng, Shuang Zhou, Xinyi Zhou, Yuchen Zhu, Yuxuan Zhu.
- 合规团队：Jin Chen, Ying Tang, Miaojun Wang, Xianzu Wang, Shaoqing Wu, Leyi Xia, W.L. Xiao.
- 业务团队：Jian Liang, Mingming Li, T. Wang, Xianzu Wang, Zhiniu Wen, Shengfeng Ye, Peng Zhang, Zhen Zhang.
- 设计团队：Wei An, Yukun Zha.

A.2. 不同的模型规模表示方法

我们针对不同的模型规模表示方法重新拟合了缩放曲线，复用了 IsoFLOP 实验配置中的数据。我们使用 $6N_1$ 和 $6N_2$ 作为模型规模表示重新计算了计算量（FLOPs），并重新拟合了性能缩放曲线。如图 6 所示，结果表明，在较高计算预算下，这三种表示方法在最优模型/数据分配上的偏差并不显著，但在较低预算下存在明显差异。



(a) 计算预算 $C = 6N_1D$　(b) 计算预算 $C = 6N_2D$　(c) 计算预算 $C = MD$

图 6 | 使用不同模型规模表示方法的性能缩放曲线。评估指标为验证集上的每字节比特数（bits-per-byte）。虚线表示拟合较小模型的幂律（灰色圆点）。蓝色星号代表 DeepSeek LLM 7B 和 67B。N_1、N_2 和 M 分别表示模型的非嵌入参数、完整参数以及每 token 的非嵌入计算量（FLOPs/token）。

[1]作者按姓氏字母顺序排列。

当使用 $6N_1$ 作为模型规模表示时，拟合的性能缩放曲线倾向于高估大规模模型的性能。相反，使用 $6N_2$ 时，曲线倾向于低估其性能。然而，使用 M 作为模型规模表示则能实现最准确的预测。

A.3. 基准测试指标曲线

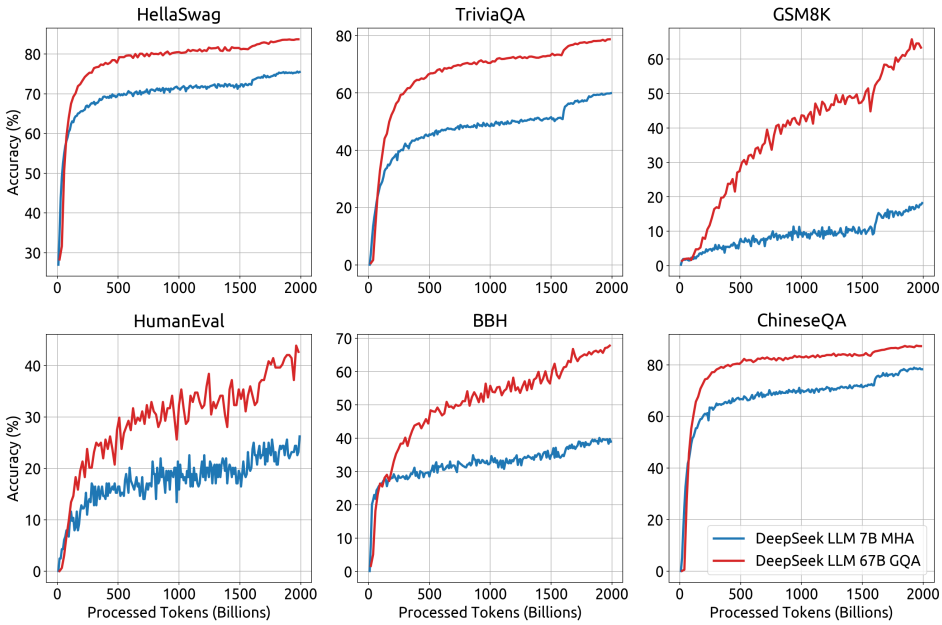


图 7 | DeepSeek LLM Base 的基准测试指标曲线。ChineseQA 是我们内部构建的测试集，其构建方式类似于 TriviaQA。

图 7 展示了不同训练步骤下的基准测试指标曲线。我们可以看到，从训练开始到结束，模型在这些基准测试上的表现持续改善。我们相信，如果继续训练，性能还将进一步提升。

模型	规模	HumanEval Python	HumanEval 多语言	MBPP
预训练模型				
Codex-001	-	33.5%	26.1%	45.9%
StarCoder	16B	36.0%	28.7%	46.8%
CodeGeeX2	6B	36.0%	24.5%	42.4%
CodeLlama	7B	31.7%	29.2%	41.6%
CodeLlama	13B	36.0%	35.4%	48.4%
CodeLlama	34B	**48.2%**	**41.0** %	55.2%
DeepSeek-LLM-Base	67B	42.7%	37.2%	**57.4%**
指令微调模型				
Wizard-Coder	34B	73.2%	48.8%	61.2%
DeepSeek-LLM-Chat	67B	**73.8%**	**53.3%**	**61.4%**

表 15 | 与代码专用模型的对比。

A.4. 与代码或数学专用模型的对比

我们将我们的模型与特定的代码和数学语言模型 (LLM) 进行了对比。表 15 表明, 尽管接触的代码数据较少, DeepSeek LLM 67B 仍能实现与 CodeLlama 相当的性能。值得注意的是, DeepSeek LLM 在代码以外的领域具备更强的能力。

同样, 表 16 展示了在 GSM8K (Cobbe et al., 2021)、MATH (Hendrycks et al., 2021)、MGSM-zh (i et al., 2023) 和 CMath (Wei et al., 2023) 等多个数学相关基准测试上的结果。DeepSeek 67B 在不同语言的数学相关任务中表现出卓越的性能，展现了其在该领域的优势。此外，DeepSeek LLM 能够利用程序来解决数学问题，其表现优于思维链（Chain-of-Thoughts）方法。在这些基准测试上，它显著优于之前的 SOTA 模型 ToRA (Gou et al., 2023)。

	推理方式	GSM8K	MATH	MGSM-zh	CMath
思维链					
MetaMath 70B (Yu et al., 2023)	CoT	82.3%	26.6%	66.4%	70.9%
WizardMath 70B (Luo et al., 2023)	CoT	81.6%	22.7%	64.8%	65.4%
DeepSeek LLM 67B Chat	CoT	**84.1**%	**32.6** %	**74.0%**	**80.3%**
工具集成推理					
ToRA-Code 34B (Gou et al., 2023)	工具集成	80.7%	50.8%	41.2%	53.4%
DeepSeek LLM 67B Chat	工具集成	**86.7%**	**51.1%**	**76.4%**	**85.4%**

表 16 | 与数学专用模型的对比。

A.5. 包含 DPO 阶段的基准测试结果

表 17 展示了包含 DPO 阶段后的基准测试结果。基于这些结果，我们可以得出结论：DPO 阶段对 LLM 的基础能力没有显著影响。

	DeepSeek 67B Chat	DeepSeek 67B Chat DPO
HellaSwag	75.7	76.1
TriviaQA	81.5	82.9
NaturalQuestions	47.0	48.8
MMLU	71.1	70.9
GSM8K	84.1	85.2
MATH	32.6	30.2
HumanEval	73.8	71.3
BBH	71.7	70.8
AGIEval	46.4	46.1
CEval	65.2	64.3
CMMLU	67.8	68.2

表 17 | DPO 阶段前后的基准测试指标。

A.6. 评估格式

表 18~ 表 40 展示了我们在不同基准测试上的评估格式示例。

PROMPT
以下是一道中国高考生物选择题，请选择正确的答案。
问题：下列有关高尔基体、线粒体和叶绿体的叙述, 正确的是选项：(A) 三者都存在于蓝藻中 (B) 三者都含有 DNA (C) 三者都是 ATP 合成的场所 (D) 三者的膜结构中都含有蛋白质
答案：从 A 到 D, 我们应选择

表 18 | AGIEval 示例。

PROMPT
问题：请根据以下信息回答问题。棉花是一种用于制作织物的植物产品。棉花由纤维素组成，这是一种人类无法消化的纤维。纤维素由许多糖分子结合成长链构成。每个糖分子含有碳、氢和氧原子。清洗棉织物时，经常会出现褶皱。服装行业使用化学物质来制造一些防皱的棉织物。还会添加染料来为棉花中的纤维素纤维着色。服装制造商将如何分离颜色以确定染料的纯度?
答案：

OPTIONS
- 通过过滤
- 通过沸点
- 通过凝固点
- 通过纸色谱法

表 19 | ARC 示例。

PROMPT
评估随机布尔表达式的结果。

Q: not ((not not True)) is
A: 让我们逐步思考。
请记住：(i) 括号内的表达式总是优先计算，(ii) 运算优先级从高到低依次为"not"（非）、"and"（与）、"or"（或）。我们首先将表达式"Z" 简化如下："Z = not ((not not True)) = not ((A))"，其中"A = not not True"。让我们计算 A：A = not not True = not (not True) = not False = True。代入 A，得到：Z = not ((A)) = not ((True)) = not True = False。所以答案是 False。

Q: True and False and not True and True is
A: 让我们逐步思考。
请记住：(i) 括号内的表达式总是优先计算，(ii) 运算优先级从高到低依次为"not"（非）、"and"（与）、"or"（或）。我们首先将表达式"Z" 简化如下："Z = True and False and not True and True = A and B"，其中"A = True and False"，"B = not True and True"。让我们计算 A：A = True and False = False。让我们计算 B：B = not True and True = not (True and True) = not (True) = False。代入 A 和 B，得到：Z = A and B = False and False = False。所以答案是 False。

Q: not not (not (False)) is
A: 让我们逐步思考。
请记住：(i) 括号内的表达式总是优先计算，(ii) 运算优先级从高到低依次为"not"（非）、"and"（与）、"or"（或）。我们首先将表达式"Z" 简化如下："Z = not not (not (False)) = not not (A)"，其中"A = not (False)"。让我们计算 A：A = not (False) = not False = True。代入 A，得到：Z = not not (A) = not not (True) = not not False = True。所以答案是 True。

Q: False and False and False or not False is
A: 让我们逐步思考。

表 20 | BBH 示例。

PROMPT
以下是中国关于教育学考试的单项选择题，请选出其中的正确答案。

根据我国心理学家冯忠良教授的学习分类，培养学生品德要通过 _____。
A. 知识的学习
B. 技能的学习
C. 行为规范的学习
D. 态度的学习
答案：C

开设跨学科课程或建立跨学科专业体现了高等教育课程发展的 _____。
A. 综合化趋势
B. 多样化趋势
C. 人文化趋势
D. 科学化趋势
答案：A

心智技能的特点有 _____。
A. 物质性、外显性、简缩性
B. 观念性、内潜性、简缩性
C. 物质性、外显性、展开性
D. 观念性、内潜性、展开性
答案：B

下列关于大学生的情绪与理智关系的说法中正确的是 _____。
A. 能冷静控制自己情绪
B. 感情用事，难以用理智控制情绪
C. 遇事能坚持自己正确认识
D. 已发展到不为小事而发怒和怄气
答案：B

在学完一篇逻辑结构严密的课文以后，勾画出课文的论点论据的逻辑关系图以帮助理解和记忆。这种学习方法属于 _____。
A. 精细加工策略
B. 组织策略
C. 复述策略
D. 做笔记策略
答案：B

有学者强调，教育要根据一个民族固有的特征来定，这种观点体现了 _____
A. 生产力对教育的影响和制约
B. 政治制度对教育的影响和制约
C. 文化对教育的影响和制约
D. 经济制度对教育的影响和制约
答案：

OPTIONS
- A
- B
- C
- D

表 21 | C-Eval 示例。

PROMPT
女：这些药怎么吃?
男：一天三次，一次两片。

请根据上文回答问题：

他们在哪儿?
答案：

OPTIONS
- 商店
- 饭店
- 医院
- 教室

表 22 | C3 示例。

PROMPT
以下是将某句古诗文翻译而成的现代表述：春天已至，万物复苏，春风如一位美丽而又心灵手巧的姑娘，迈着纤纤细步款款而来，她挥舞剪刀，尽情地展示那高超的女工技巧，她先裁出了柳叶，随着柳条袅袅依依地舞蹈，又裁出杏叶，桃叶。
该翻译所对应的古诗文是：

OPTIONS
- 春风骋巧如翦刀
- 剪裁无巧似春风
- 风吹怨恨快如刀
- 春风欲擅秋风巧

表 23 | CCPM 示例。

PROMPT

Q: 某小学在“献爱心–为汶川地震区捐款”活动中，六年级五个班共捐款 8000 元，其中一班捐款 1500 元，二班比一班多捐款 200 元，三班捐款 1600 元，四班与五班捐款数之比是 3：5．四班捐款多少元?
A: 一班捐款 1500 元, 而二班比一班多捐 200 元, 所以二班捐款 1500+200=1700 元，又知道六年级五个班一共捐款 8000 元，所以四班和五班捐款之和 = 一共捐款 - 一班和二班和三班捐款之和，即 8000-1500-1700-1600=3200 元，而题目说四班与五班捐款数之比是 3：5，则四班捐款了 3200/(3+5)*3=1200 元。所以答案是：1200。

Q: 小俊在东西大道上跑步，若规定向东为正。他先向东跑了 800 米，然后又跑了一段之后，他位于出发点西边 100 米处，小俊第二段跑了多少米?
A: 小俊第二段跑完后位于出发点西边，所以第二段应该是向西跑，第二段跑的长度-第一段跑的长度 =100，第二段跑了 100+800=900 米。所以答案是：900。

Q: A 车和 B 车同时从甲、乙两地相向开出，经过 5 小时相遇．然后，它们又各自按原速原方向继续行驶 3 小时，这时 A 车离乙地还有 135 千米，B 车离甲地还有 165 千米．甲、乙两地相距多少千米?
A: 假设 A 车的速度为 x 千米每小时，B 车的速度为 y 千米每小时，根据而 A、B 相遇时 A 车行驶了 5 小时，A 车行驶 3 小时后离乙地还有 135 千米，B 车行驶 3 小时后距离甲地还有 165 千米，可以得到甲乙两地相距 =5x+5y=135+8x=165+8y, 变换得到：10(x+y)=300+8(x+y), 于是 x+y=150, 甲乙两地相距 5(x+y)=750 千米。所以答案是：750。

Q: 在一个底面半径为 10 厘米的圆柱形容器内，倒入 10 厘米深的水，然后将一个底面直径 4 厘米，高 6 厘米的圆锥形铅锤放入水中，容器中水面上升多少厘米?
A:

表 24 | CMATH 示例。

PROMPT

以下是关于解剖学的单项选择题，请直接给出正确答案的选项。

题目：壁胸膜的分部不包括
A. 肋胸膜
B. 肺胸膜
C. 膈胸膜
D. 胸膜顶
答案是：B

题目：属于蝶骨上的结构为
A. 垂体窝
B. 棘孔
C. 破裂孔
D. 视神经管
答案是：B

题目：属于右心房的结构是
A. 肉柱
B. 室上嵴
C. 乳头肌
D. 梳状肌
答案是：D

题目：咽的分部
A. 咽隐窝
B. 口咽部
C. 鼻咽部
D. 喉咽部
答案是：C

题目：舌下神经核位于
A. 间脑
B. 延髓
C. 中脑
D. 脑桥
答案是：B

题目：从脑干背侧出脑的脑神经是
A. 副神经
B. 三叉神经
C. 舌下神经
D. 滑车神经
答案是：

OPTIONS

- A
- B
- C
- D

表 25 | CMMLU 示例。

PROMPT

段落：该市的中位年龄为 22.1 岁。10.1% 的居民年龄在 18 岁以下；56.2% 的居民年龄在 18 至 24 岁之间；16.1% 的居民年龄在 25 至 44 岁之间；10.5% 的居民年龄在 45 至 64 岁之间；7% 的居民年龄在 65 岁或以上。该市的人口性别构成为 64.3% 的男性和 35.7% 的女性。

请根据上述段落回答以下问题，如需计算请仔细计算。
Q: 不在 25 至 44 岁之间的人口占比是多少?
A: 答案类型为数字。因此根据上述段落，答案是 83.9。

Q: 不在 25 至 44 岁之间的人口占比是多少?
A: 答案类型为数字。因此根据上述段落，答案是

表 26 | DROP 示例。

PROMPT

中新网 12 月 7 日电综合外媒 6 日报道, 在美国得克萨斯州, 负责治疗新冠肺炎患者的医生约瑟夫 · 瓦隆 (Joseph Varon) 已连续上班超 260 天, 每天只睡不超过 2 小时。瓦隆日前接受采访时呼吁, 美国民众应遵从防疫规定, 一线的医护人员“已

OPTIONS

- 神清气爽”。
- 诡计多端”。
- 精疲力竭”。
- 分工合作”。
- 寅吃卯粮”。
- 土豪劣绅”。
- 芸芸众生”。

表 27 | CHID 示例。

PROMPT

胡雪岩离船登岸，坐轿进城，等王有龄到家，他接着也到了他那里，脸上是掩抑不住的笑容，王有龄夫妇都觉得奇怪，问他什么事这么高兴。
上面的句子中的”他”指的是
胡雪岩

渐渐地，汤中凝结出一团团块状物，将它们捞起放进盆里冷却，肥皂便出现在世上了。
上面的句子中的”它们”指的是
块状物

“她序上明明引着 JulesTellier 的比喻，说有个生脱发病的人去理发，那剃头的对他说不用剪发，等不了几天，头毛压儿全掉光了；大部分现代文学也同样的不值批评。这比喻还算俏皮。”
上面的句子中的”他”指的是
生脱发病的人

在洛伦佐大街的尽头处，矗立着著名的圣三一大教堂。它有着巨大的穹顶，还有明亮的彩色玻璃窗，上面描绘着《旧约》和《新约》的场景。
上面的句子中的”它”指的是
圣三一大教堂

他伯父还有许多女弟子，大半是富商财主的外室；这些财翁白天忙着赚钱，怕小公馆里的情妇长日无聊，要不安分，常常叫她们学点玩艺儿消遣。
上面的句子中的”她们”指的是
情妇

赵雨又拿出了一个杯子，我们热情地请老王入座，我边给他倒酒边问：1962 年的哪次记得吗？“
上面的句子中的”他”指的是

表 28 | CLUEWSC 示例。

PROMPT

Q: Max 可以在 40 分钟内割完草坪。如果施肥所需的时间是割草的两倍，那么他割草和施肥总共需要多长时间?
A: 让我们一步步思考。Max 施肥需要 2 * 40 分钟 = 80 分钟。总共，Max 割草和施肥需要 80 分钟 + 40 分钟 = 120 分钟。答案是 120。

Q: 贝果每个售价 $2.25，或者一打（12 个）售价 $24。如果一次买一打，每个贝果能节省多少美分?
A: 让我们一步步思考。它们每个售价 2.25*100=225 美分。按批量价格计算，它们是 24/12=2 美元每个。它们每个售价 2*100=200 美分。225-200=25，每个贝果节省 25 美分。答案是 25。

Q: Tim 今年 5 岁。他的表兄 Rommel 的年龄是他的三倍。他的另一个表妹 Jenny 比 Rommel 大 2 岁。Tim 比 Jenny 小多少岁?
A: 让我们一步步思考。Rommel 的年龄是 5 x 3 = 15 岁。Jenny 的年龄是 15 + 2 = 17 岁。因此，Tim 比 Jenny 小 17 - 5 = 12 岁。答案是 12。

Q: 学校有 14 名男生和 10 名女生。如果有 4 名男生和 3 名女生退学，还剩下多少名男生和女生?
A: 让我们一步步思考。剩下 14 名男生 - 4 名男生 = 10 名男生。剩下 10 名女生 - 3 名女生 = 7 名女生。总共剩下 10 名男生 + 7 名女生 = 17 名学生。答案是 17。

Q: 建造一个鸟屋需要 7 块木板和 20 颗钉子。如果 1 颗钉子售价 0.05 美元，1 块木板售价 3 美元，那么建造 4 个鸟屋的成本是多少美元?
A: 让我们一步步思考。一个鸟屋的木板成本是 7 * 3 = 21 美元。每个鸟屋的钉子成本是 20 * 0.05 = 1 美元。因此，建造一个鸟屋需要 21 + 1 = 22 美元。那么建造 4 个鸟屋的成本是 4 * 22 = 88 美元。答案是 88。

Q: Danny 带了 3 个西瓜去家庭野餐。他把每个西瓜切成 10 片。他的姐姐带了 1 个西瓜去野餐，她把西瓜切成 15 片。野餐时总共有多少片西瓜?
A: 让我们一步步思考。Danny 带来了 3 * 10 = 30 片西瓜。他姐姐带来了 1 * 15 = 15 片西瓜。总共有 30 + 15 = 45 片西瓜。答案是 45。

Q: Angela 是纽约的一名自行车快递员。她需要投递的包裹数量是餐食数量的 8 倍。如果她总共需要投递 27 份餐食和包裹，她投递了多少份餐食?
A: 让我们一步步思考。设 p 为 Angela 投递的包裹数量，m 为餐食数量。我们知道 p + m = 27 且 p = 8m。将第二个方程代入第一个方程，得到 8m + m = 27。合并同类项，得到 9m = 27。两边同时除以 9，得到 m = 3。答案是 3。

Q: Cori 今天 3 岁。5 年后，她的年龄将是她阿姨年龄的三分之一。她阿姨今天多少岁?
A: 让我们一步步思考。5 年后，Cori 将是 3 + 5 = 8 岁。5 年后，Cori 的阿姨将是 8 x 3 = 24 岁。今天，她阿姨是 24 - 5 = 19 岁。答案是 19。

Q: Indras 的名字有 6 个字母。她姐姐的名字比 Indras 名字字母数的一半多 4 个。Indras 和她姐姐的名字总共有多少个字母?
A: 让我们一步步思考。

表 29 | GSM8K 示例。

PROMPT
弹钢琴：一名男子坐在钢琴前。他

OPTIONS
- 正用双手和面部弹奏钢琴。
- 开始在钢琴上弹奏 Timbaland 的歌曲。
- 缓慢弹奏，并停顿打响指。
- 正在弹奏他面前的一首曲子。

表 30 | HellaSwag 示例。

PROMPT

```
def starts_one_ends(n):
"""
给定一个正整数 n，返回以 1 开头或结尾的 n 位正整数的个数。
"""
```

表 31 | HumanEval 示例。

PROMPT
问题：
求表达式 $\frac{\sqrt{x-2}}{\sqrt{5-x}}$ 的定义域。}

解答：
每个根号内的表达式必须非负。
因此，$x-2 \ge 0$，即 $x\ge2$，且 $5 - x \ge 0$，即 $x \le 5$。
此外，分母不能等于零，因此 $5-x>0$，即 $x<5$。
因此，该表达式的定义域为 $\boxed{[2,5)}$。
最终答案：最终答案是 $[2,5)$。希望它是正确的。

问题：
如果 $\det \mathbf{A} = 2$ 且 $\det \mathbf{B} = 12,$ 求 $\det (\mathbf{A} \mathbf{B}).$

解答：
我们有 $\det (\mathbf{A} \mathbf{B}) = (\det \mathbf{A})(\det \mathbf{B}) = (2)(12) = \boxed{24}.$
最终答案：最终答案是 24。希望它是正确的。

问题：
Terrell 通常举起两个 20 磅的哑铃 12 次。如果他改用两个 15 磅的哑铃，Terrell 必须举起多少次才能举起相同的总重量?

解答：
如果 Terrell 举起两个 20 磅的哑铃 12 次，他总共举起的重量为 $2\cdot 12\cdot20=480$ 磅。如果他改用两个 15 磅的哑铃举起 n 次，他总共举起的重量为 $2\cdot15\cdot n=30n$ 磅。令其等于 480 磅，我们可以解出 n:
\begin{align*}
30n&=480\\
\Rightarrow\qquad n&=480/30=\boxed{16}
\end{align*}
最终答案：最终答案是 16。希望它是正确的。

问题：
如果方程组

\begin{align*}
6x-4y&=a,\\
6y-9x &=b.
\end{align*} 有解 (x, y)，其中 x 和 y 均不为零，求 $\frac{a}{b},$ 假设 b 不为零。

解答：
如果我们将第一个方程乘以 $-\frac{3}{2}$，我们得到

$$6y-9x=-\frac{3}{2}a.$$ 由于我们还知道 $6y-9x=b$，因此我们有

$$-\frac{3}{2}a=b\Rightarrow\frac{a}{b}=\boxed{-\frac{2}{3}}.$$
最终答案：最终答案是 $-\frac{2}{3}$。希望它是正确的。

问题：计算 $\log_21$。

解答：

表 32 | MATH 示例。

PROMPT
你是一名专业的 Python 程序员，你的任务是：编写一个函数，从给定的两个元组列表中找出相似的元素。你的代码应通过以下测试：

```
assert similar_elements((3, 4, 5, 6),(5, 7, 4, 10)) == (4, 5)
assert similar_elements((1, 2, 3, 4),(5, 4, 3, 7)) == (3, 4)
assert similar_elements((11, 12, 14, 13),(17, 15, 14, 13)) == (13, 14)
[BEGIN]
def similar_elements(test_tup1, test_tup2):
res = tuple(set(test_tup1) & set(test_tup2))
return (res)
[DONE]
```

你是一名专业的 Python 程序员，你的任务是：编写一个 Python 函数来识别非质数。你的代码应通过以下测试：

```
assert is_not_prime(2) == False
assert is_not_prime(10) == True
assert is_not_prime(35) == True
[BEGIN]
import math
def is_not_prime(n):
result = False
for i in range(2,int(math.sqrt(n)) + 1):
if n % i == 0:
result = True
return result
[DONE]
```

你是一名专业的 Python 程序员，你的任务是：编写一个函数，使用堆队列算法从给定的数字列表中找出最大的整数。你的代码应通过以下测试：

```
assert heap_queue_largest( [25, 35, 22, 85, 14, 65, 75, 22, 58],3)==[85, 75, 65]
assert heap_queue_largest( [25, 35, 22, 85, 14, 65, 75, 22, 58],2)==[85, 75]
assert heap_queue_largest( [25, 35, 22, 85, 14, 65, 75, 22, 58],5)==[85, 75, 65,
58, 35]
[BEGIN]
import heapq as hq
def heap_queue_largest(nums,n):
largest_nums = hq.nlargest(n, nums)
return largest_nums
[DONE]
```

你是一名专业的 Python 程序员，你的任务是：编写一个函数，返回一个数字的所有因数之和。你的代码应通过以下测试：

```
assert sum_div(8)==7
assert sum_div(12)==16
assert sum_div(7)==1
[BEGIN]
```

表 33 | MBPP 示例。

提示
以下是关于各类常识的选择题（附答案）。

标准汽车有多少个车轴?
A. 一个
B. 两个
C. 四个
D. 八个
答案：B

摇滚传奇乐队 Cheap Trick 1979 年现场专辑的标题中提到了哪个地点?
A. 布达佩斯
B. 武道馆
C. 不丹
D. 英国
答案：B

谁是历史上赢得 NBA 扣篮大赛的最矮男子?
A. 安东尼 · “土豆” · 韦伯
B. 迈克尔 · “飞人” · 乔丹
C. 泰伦 · “魔西” · 博格
D. 朱利叶斯 · “J 博士” · 欧文
答案：A

光合作用过程中会产生什么?
A. 氢气
B. 尼龙
C. 氧气
D. 光
答案：C

以下哪首歌曲是摇滚乐队 The Police 的十大热门金曲?
A. 《Radio Ga-Ga》
B. 《Ob-la-di Ob-la-da》
C. 《De Do Do Do De Da Da Da》
D. 《In-a-Gadda-Da-Vida》
答案：C

《三个臭皮匠》中哪一位与其他两位没有亲属关系?
A. 莫
B. 拉里
C. 柯利
D. 谢姆
答案：

选项
- A
- B
- C
- D

表 34 | MMLU 示例。

提示
回答以下问题：
问：谁主办了 2022 年 fifa 世界杯?
答：卡塔尔
问：谁赢得了首届女子 fifa 世界杯?
答：美国
问：《迈阿密风云》何时停播?
答：1989 年
问：谁创作了歌曲《Shout to the Lord》?
答：Darlene Zschech
问：谁被扔进了狮子坑?
答：但以理
问：名字 habib 的含义是什么?
答：

表 35 | NaturalQuestions 示例。

提示
一位女士注意到自己每年秋天都会感到抑郁，并想知道原因。一位朋友建议她，也许随着季节从温暖转向寒冷而发生的一些变化可能对她产生了影响。当被要求举例说明这些变化时，朋友提到了

选项
- 花朵盛开
- 草地变黄
- 树木生长
- 繁花绽放

表 36 | OpenBookQA 示例。

提示
为了更方便地按下位于机器下方的垃圾处理器复位按钮，

选项
- 在橱柜地板上放一面墙镜
- 在垃圾处理器下方手持一面手镜

表 37 | PIQA 示例。

提示
文章：
阅读文章时，如果你能理清作者是如何将观点组织在一起的，你会更好地理解并记住它。有时，作者通过提出问题然后回答它们来组织观点。例如，如果文章是关于土拨鼠的，作者脑海中可能有一系列问题：
土拨鼠长什么样？
土拨鼠住在哪里？
它们吃什么？...
在文章中，作者可能会回答这些问题。
有时作者会在文章中写出她的问题。这些问题会给你提示，告诉你作者接下来要写什么。通常作者脑海中有一个问题，但她不会写出来给你看。你必须自己推断出她的问题。这里有一篇阅读材料供你练习这种方法。
蚯蚓
你知道有多少种蚯蚓吗？世界上大约有 1800 种！它们可能是棕色、紫色或绿色的。它们可以小到 3 厘米长，也可以大到 3 米长。
观察蚯蚓的最佳时间是夜晚，尤其是凉爽潮湿的夜晚。那时它们会从洞穴里出来觅食。蚯蚓不喜欢晒太阳。这是因为它们通过皮肤呼吸，如果皮肤太干就无法呼吸。如果雨下得很大，蚯蚓必须从土里钻出来，因为它们无法在被水淹没的洞穴中呼吸。多么危险的生活啊！
蚯蚓没有眼睛，那么它们如何知道天黑了呢？它们的皮肤上有对光敏感的特殊部位。这些斑点能分辨明暗。如果你在晚上用手电筒照蚯蚓，它会迅速钻入地下。
蚯蚓也没有耳朵，但它们可以通过感知地面的震动来“听”。如果你想像蚯蚓一样听，就躺在地上，用手指堵住耳朵。然后让朋友在你附近跺脚。这就是蚯蚓感知附近鸟类和人类行走以及鼹鼠挖掘的方式。
蚯蚓很有用。农民和园丁喜欢他们的土地里有很多蚯蚓，因为蚯蚓在挖掘时有助于改良土壤。这种挖掘使土壤保持疏松透气。在一年内，蚯蚓可以在大约一个足球场大小的区域堆积多达 23,000 公斤的粪土。

问：阅读《蚯蚓》的目的是什么？

答：将作者的想法付诸实际应用。

问：文章中无法回答哪个问题？

答：为什么人类能像蚯蚓一样听？

问：根据这篇文章，如何更好地理解《蚯蚓》？

答：阅读时理清作者脑海中的所有问题。

问：这篇文章的最佳标题是什么？

答：

选项
- 一种有助于理解的方法
- 一种练习新想法的方法
- 一种学习成为明智作家的方法
- 一种更清楚了解蚯蚓的方法

表 38 | RACE 示例。

提示
回答以下问题：
问："Jayhawker" 一词用于指代来自美国某个州的反奴隶制武装团体，他们与来自密苏里州的亲奴隶制派系发生冲突。这个州是哪个，有时被称为 Jayhawk 州？
答：堪萨斯州
问：哪位瑞典 DJ 和唱片制作人在 2013 年凭借《Wake Me Up》获得了英国单曲榜冠军？
答：Tim Bergling
问：谢菲尔德哈勒姆选区的国会议员是谁？
答：Nick Clegg
问：一个轰动全国的案件，田纳西州诉约翰 · 托马斯 · 斯科普斯案于 1925 年 7 月 21 日结束，陪审团裁定斯科普斯先生因教授什么而有罪？
答：适者生存
问：哪部卡通系列剧中有名为 Little My 的角色？
答：姆明
问："哪位英国模特留着短发中性风，原名 Lesley Hornby，16 岁时被 Nigel Davies 发掘，体重仅 6 英石（41 公斤，91 磅），并凭借 Mary Quant 为她打造的高时尚摩登造型成为 "66 年度面孔"？"
答：

表 39 | TriviaQA 示例。

前缀
- 所以莫妮卡
- 所以杰西卡

补全
避免为了眼睛健康吃胡萝卜，因为艾米丽需要好视力，而莫妮卡不需要。

表 40 | WinoGrande 示例。注意，WinoGrande 有多个前缀和仅一个补全部分，我们选择补全部分困惑度最低的预测前缀。