Machine Translated by Google

deepseek

DeepSeek 法学硕士
利用 Longtermism 扩展开源语言模型

毕晓、陈德利、陈冠廷、陈善皇、戴大麦、邓承启、丁红辉、董凯、杜秋实、付喆、高华作、高凯歌、高文君、葛瑞奇、关康、郭大亚、郭建中、郝广博、郝哲文、何英、胡文杰、黄盼盼、李二航、李国伟、李嘉实、李耀、李玉坤、梁文峰、方云林, AX刘, 刘博, 刘文, 刘晓东, 刘鑫, 刘艺源, 路浩宇, 陆尚浩, 罗福丽, 马世荣, 聂晓涛, 裴田, 朴一世, 邱俊杰, 曲辉, 任同正, 任泽辉, 阮冲, 张丽沙, 邵志宏, 宋俊晓, 苏学成, 孙静翔, 孙耀峰, 唐明辉, 王秉轩, 王培毅,王世宇、王耀辉、王永吉、童武、Y. Wu、谢欣、谢振达、谢紫薇、熊一良、徐汉伟、徐旭、徐彦宏、杨德建、游宇翔、于水平、余兴凯、B. 张、张浩伟、张乐从、张丽月、张明川、张明华、张文涛、张一超、赵成钢、赵耀、周商言、周顺风、朱启浩、邹宇恒

*DeepSeek-AI

抽象的

开源大型语言模型（LLM)的快速发展确实令人瞩目。
然而,先前文献中描述的扩展定律得出的结论各不相同,这给扩展 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,这些发现有助于在两种普遍使用的开源配置 7B 和 67B 中扩展大规模模型。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远眼光推进开源语言模型的项目。

为了支持预训练阶段,我们开发了一个数据集,目前包含 2万亿个 token,并且还在不断扩展。我们进一步对 DeepSeek LLM Base 模型进行监督微调 (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 在一系列基准测试中都超越了 LLaMA-2 70B,尤其是在代码、数学和推理领域。此外,开放式评估表明,我们的 DeepSeek LLM 67B Chat 与 GPT-3.5 相比表现出色。

arXiv:2401.02954v1

*作者按姓氏字母顺序排列。

Machine Translated by Google

内容

Machine Translated by Google

1. 简介

近些年来,基于仅解码器的 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)的基石和途径。通过预测连续文本中的下一个单词,LLM在海量数据集上进行自监督预训练,使其能够实现各种目的并拥有许多能力,例如小说创作、文本摘要、代码完成等。

随后,监督微调和奖励建模等技术的发展使得大型语言模型 (LLM) 能够更好地遵循用户的意图和指令。这赋予了它们更加灵活的对话能力,并迅速扩大了它们的影响力。

这股浪潮是由ChatGPT（OpenAI,2022 年）、Claude（Anthropic ,2023 年)和 Bard（Google,2023 年）等封闭产品引发的,这些产品耗费了大量计算资源和注释成本。这些产品大大提高了社区对开源 LLM 功能的期望,从而激发了一系列工作（Bai 等人,2023 年;Du 等人,2022 年;Jiang 等人,2023 年;Touvron 等人,2023a、b;Yang 等人,2023 年）。

其中,LLaMA 系列模型（Touvron 等,2023a、b)脱颖而出。它整合了一系列工作,创建了一个高效稳定的架构,构建了从 7B 到 70B 参数范围的性能良好的模型。因此,LLaMA 系列已成为开源模型中架构和性能的事实基准。

继 LLaMA 之后,开源社区主要专注于训练固定大小（7B、13B、34B 和 70B)的高质量模型,而往往忽略了对 LLM缩放定律的研究探索（Hoffmann 等人,2022 年;Kaplan 等人,2020 年）。尽管如此,考虑到当前的开源模型仅处于通用人工智能 (AGI) 发展的初始阶段,对缩放定律的研究至关重要。此外,早期的研究（Hoffmann等人,2022 年;Kaplan 等人,2020 年)在计算预算增加的情况下对模型和数据的扩展得出了不同的结论,并且没有充分解决超参数讨论。在本文中,我们广泛研究了语言模型的缩放行为,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。我们的研究旨在为开源 LLM 未来的扩展奠定基础,为该领域的进一步发展铺平道路。具体来说,我们首先研究了批量大小和学习率的缩放规律,并发现了它们随模型大小的变化趋势。在此基础上,我们对数据和模型规模的缩放规律进行了全面研究,成功揭示了最佳的模型/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得出的缩放规律显示出显著差异。这表明数据集的选择会显著影响缩放行为,这表明在跨数据集推广缩放规律时应谨慎行事。

在我们的扩展法则的指导下,我们从零开始构建开源大型语言模型,并发布尽可能多的信息供社区参考。我们收集了 2 万亿个 token 用于预训练,主要使用中文和英文。在模型层面,我们总体上遵循了 LLaMA 的架构,但用多步学习率调度器取代了余弦学习率调度器,在保持性能的同时促进了持续训练。我们从各种来源收集了超过 100 万个用于监督微调 (SFT) (Ouyang et al., 2022) 的实例。本文分享了我们在不同 SFT 策略方面的经验以及数据消融技术方面的发现。此外,我们还利用直接偏好优化 (DPO) (Rafailov et al., 2023) 来提高模型的对话性能。

Machine Translated by Google

我们使用基础模型和聊天模型进行了广泛的评估。评估结果表明,DeepSeek LLM 在各种基准测试中都超越了 LLaMA-2 70B,特别是在代码、数学和推理领域。继 SFT 和 DPO 之后,DeepSeek 67B 聊天模型在中文和英文开放式评估中均优于 GPT-3.5。这凸显了 DeepSeek 67B 在生成高质量响应和以两种语言进行有意义的对话方面的卓越表现。此外,安全性评估表明 DeepSeek 67B Chat 在实践中可以提供无害的响应。

在本文的其余部分,我们首先在第 2 节中介绍 DeepSeek LLM 的预训练基本概念,包括数据组成、模型架构、基础架构和超参数。在第 3 节中,我们详细解释了我们发现的缩放规律及其含义。此外,我们讨论了选择预训练超参数背后的理由,并考虑了从缩放规律分析中获得的见解。在第 4 节中,我们讨论了我们的微调方法,包括微调数据的组成以及 SFT 和 DPO 阶段的特定方法。然后,我们在第 5 节中介绍了 DeepSeek LLM 的详细评估结果,涵盖了基础模型和聊天模型,以及它们在开放式评估和安全评估中的表现。最后,我们在第 6 节中讨论了 DeepSeek LLM 当前的局限性和未来发展方向。

2. 预训练

2.1. 数据

我们的主要目标是全面增强数据集的丰富性和多样性。
我们从知名来源获得了宝贵的见解,例如（Computer,2023;Gao 等人， 2020;Penedo 等人,2023;Touvron 等人,2023a）。为了实现这些目标,我们将方法分为三个基本阶段:重复数据删除、过滤和重新混合。重复数据删除和重新混合阶段通过对唯一实例进行采样来确保数据的多样化表示。

过滤阶段增强了信息的密度,从而使模型训练更加高效、有效。

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 语料库进行去重比对单个转储进行去重更能消除重复实例。表 1 表明,对 91 个转储进行去重比单个转储方法消除的文档多四倍。

转储使用重复数据	1	2	6	12	16	22	41	91
删除率 (%)	22.2	46.7	55.7	69.9	75.7	76.3	81.6	89.8

表 1 | 各种常见爬网转储的重复数据删除率。

在筛选阶段,我们专注于制定稳健的文档质量评估标准。这涉及结合语言和语义评估的详细分析,从个体和全局角度提供数据质量视图。在混合阶段,我们调整方法以解决数据不平衡问题,重点是增加代表性不足的领域的存在。这一调整旨在实现更加平衡和包容的数据集,确保充分代表不同的观点和信息。

对于我们的标记器,我们基于标记器库 (Huggingface Team, 2019) 实现了字节级字节对编码 (BBPE) 算法。预标记化用于

Machine Translated by Google

防止不同字符类别的标记合并,例如换行符、标点符号、
以及中日韩 (CJK) 符号,类似于 GPT-2 (Radford 等人,2019)。我们还
选择按照（Touvron 等人，
2023a,b)。根据我们之前的经验,我们设置了
词汇量为 100000。标记器在大约
24 GB,我们用 15 个特殊标记扩充了最终词汇表,使总词汇量达到
大小为100015。为了保证训练时的计算效率,并预留空间
未来可能需要的任何额外特殊标记,我们配置了模型的
将词汇量设置为102400以进行训练。

2.2. 架构

参数层		模型	头	kv_heads	上下文序列学习 长度 批次大小		速度	代币
7B	三十	4096	三十二	三十二	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。我们选择超参数
根据我们在第 3 节中的发现

DeepSeek LLM 的微设计很大程度上遵循了 LLaMA 的设计（Touvron 等人，
2023a,b），采用预范数结构,并采用 RMSNorm（Zhang and Sennrich,2019)函数
并使用 SwiGLU (Shazeer, 2020) 作为前馈网络的激活函数
（FFN），中间层维度为。它还结合了旋转嵌入 $\frac{8}{3}$
（Su et al., 2024)进行位置编码。为了优化推理成本,67B 模型使用分组查询注意力 (GQA)（Ainslie et al., 2023)代
替传统的多头注意力
（内政部）。

然而,在宏观设计方面,DeepSeek LLM 略有不同。具体来说,DeepSeek
LLM 7B 是一个 30 层网络,而 DeepSeek LLM 67B 有 95 层。这些层调整，
在保持与其他开源模型参数一致性的同时,也方便模型
管道分区以优化训练和推理。

与大多数使用分组查询注意 (GQA) 的作品不同,我们扩展了 67B 模型的
网络深度的参数,而不是扩大中间层的常见做法
FFN 层的宽度,旨在获得更好的性能。详细的网络规范可以
见表2。

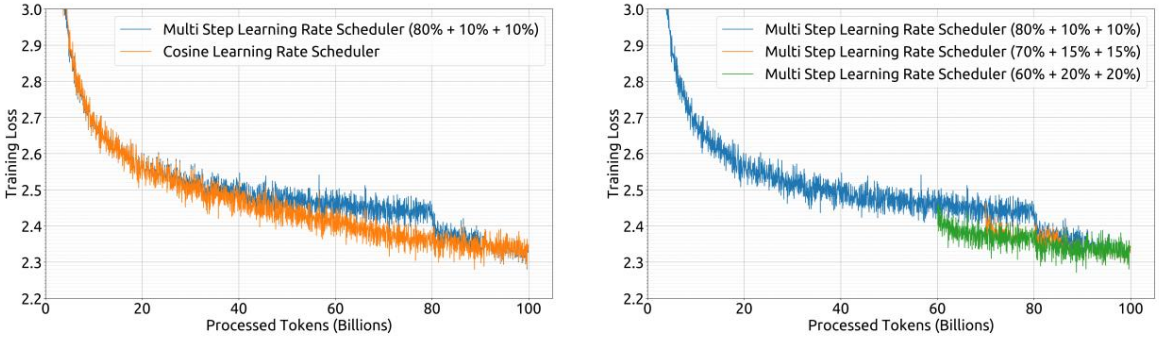
2.3. 超参数

DeepSeek LLM 以 0.006 的标准差初始化,并使用 AdamW 进行训练
优化器（Loshchilov and Hutter,2017）,具有以下超参数： $_1 = 0.9, _2 = 0.95$ ，
和weight_decay = 0.1。

在预训练过程中采用多步学习率调度器,而不是典型的
余弦调度器。具体来说,模型的学习率在
2000 个预热步骤,处理完 80% 的数据后,下降到最大值的 31.6%。
训练 token。在 90% 的 token 之后,它会进一步减少到最大值的 10%。
训练阶段的梯度裁剪设置为1.0。

根据我们的实证研究结果,我们发现,尽管损失减少程度存在差异，

Machine Translated by Google



(a)多步与余弦学习率衰减　　(b)多步骤阶段的不同比例

图 1 |不同学习率调度器或不同调度器参数的训练损失曲线。模型大小为 16 亿个参数,在 1000 亿个标记的数据集上进行训练。

从训练过程中的趋势来看,使用多步学习率调度器的最终性能与余弦调度器的性能基本一致,如图 1(a) 所示。在保持模型大小不变的情况下调整训练规模时,多步学习率调度器允许重用第一阶段的训练,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我们的默认设置。

我们还在图 1(b) 中证明了,调整多步学习率调度器中不同阶段的比例可以略微提高性能。然而,为了平衡持续训练中的重用率和模型性能,我们选择了上述三个阶段分别 80%、10% 和 10% 的分布。

批次大小和学习率随模型大小而变化。7B和67B模型预训练阶段的具体参数见表2。

2.4. 基础设施

我们使用高效轻量级的训练框架 HAI-LLM (High-flyer, 2023) 来训练和评估大型语言模型。数据并行性、张量并行性、序列并行性和 1F1B 流水线并行性都集成到了此框架中,就像 Megatron 中所做的那样(Korthikanti et al., 2023; Narayanan et al., 2021; Shoeybi et al., 2019)。我们还利用闪存注意 (Dao, 2023; Dao et al., 2022) 技术来提高硬件利用率。ZeRO-1 (Rajbhandari et al., 2020) 被用于在数据并行等级上划分优化器状态。我们还努力重叠计算和通信以尽量减少额外的等待开销,包括 ZeRO-1 中最后一个微批量和减少散射操作的后向过程,以及顺序并行的 GEMM 计算和全收集/减少散射。

融合了一些层/运算符以加快训练速度,包括 LayerNorm、尽可能的 GEMM和 Adam 更新。为了提高模型训练稳定性,我们以 bf16 精度训练模型,但以 fp32 精度累积梯度。执行就地交叉熵以减少 GPU 内存消耗,即:我们在交叉熵 CUDA 内核中动态将 bf16 logit 转换为 fp32 精度（而不是事先在 HBM 中转换）,计算相应的 bf16 梯度,并用其梯度覆盖 logit。

模型权重和优化器状态每 5 分钟异步保存一次,这意味着在偶尔出现硬件或网络故障的最坏情况下,我们损失的训练时间不会超过 5 分钟。这些临时模型检查点会定期清除,以避免

Machine Translated by Google

占用过多的存储空间。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。

至于评估,我们在生成任务中采用 vLLM（Kwon 等,2023）,在非生成任务中采用连续批处理,以避免手动调整批次大小并减少标记填充。

3. 缩放定律

缩放定律的研究（Hestness 等,2017)早于大型语言模型的出现。

缩放定律 (Henighan et al., 2020; Hoffmann et al., 2022; Kaplan et al., 2020) 表明,随着计算预算、模型规模和数据规模的增加,模型性能可以得到可预测的提升。当模型规模用模型参数表示,数据规模用 token 数量表示时,可以近似为= 6 。因此,在增加计算预算的情况下,如何优化模型和数据规模之间的分配也是缩放定律的一个重要研究目标。

随着 LLM 的发展（Dai 等人,2019 年;Radford 等人,2019 年）,更大的模型实现了意想不到的显著性能提升,将缩放定律研究推向了新的高峰。缩放定律的结果表明,扩大计算预算继续产生显著的好处,这进一步鼓励了模型规模的增加（Brown 等人， 2020 年;Smith 等人,2022 年）。

然而,如表 4 所示,早期关于最佳模型/数据扩展分配策略的研究（Hoffmann 等人,2022 年;Kaplan 等人,2020 年)得出了不同的结论,这引发了人们对扩展定律的普遍适用性的怀疑。此外,这些研究通常缺乏对超参数设置的完整描述,因此无法确定不同计算预算下的模型是否达到了最佳性能。因此,我们在本节中重新审视扩展定律,以解决这些不确定性,并确保我们走在有效扩展计算的正确道路上,这反映了长期视角,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规律。经验上,我们观察到训练过程中大多数参数的最优值不会随着计算预算的变化而变化。因此,这些参数与 2.3 节中概述的参数一致,并且在不同的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,被重新审视。

早期的研究（Goyal 等人,2017 年;McCandlish 等人,2018 年;Shallue 等人,2019 年;Smith 等人,2017 年； Zhang 等人,2019 年)为设置批量大小和学习率提供了一些实证观察,但我们发现这些观察在我们的初步实验中的适用性有限。

通过大量实验,我们模拟了计算预算与最佳批次大小和学习率之间的幂律关系。这种关系被称为超参数的缩放定律,它为确定最佳超参数提供了一个经验框架。这种方法可确保不同计算预算的模型都能达到近乎最佳的性能。

然后,我们研究模型和数据规模的缩放规律。为了降低实验成本和拟合难度,我们采用了 Chinchilla 的 IsoFLOP 配置文件方法（Hoffmann等人,2022 年)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法,即非嵌入 FLOPs/token ,取代了以前使用的模型参数 ,并代之以近似计算预算公式 = 6

Machine Translated by Google

=更加精确。实验结果为最优模型/数据扩容分配策略和性能预测提供了参考,也准确预测了 DeepSeek LLM 7B 和 67B 模型的预期性能。

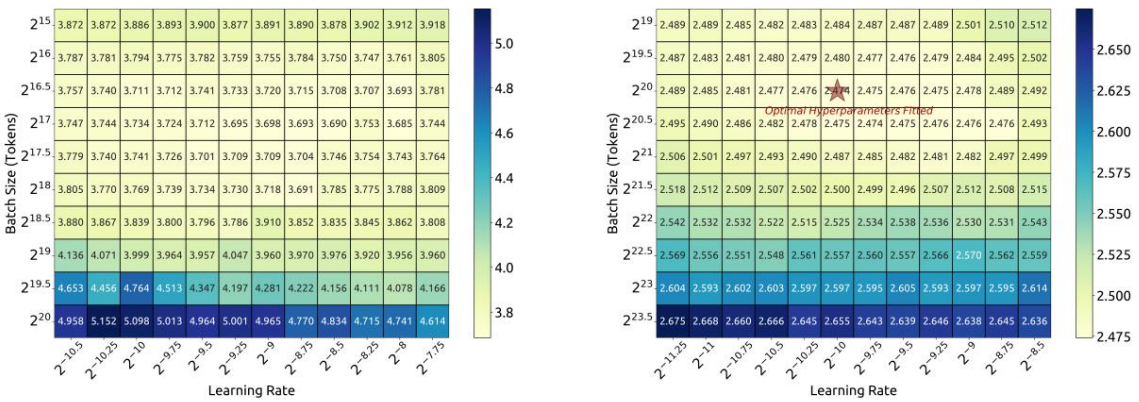
此外,在探索缩放规律的过程中,我们使用的数据经过多次迭代,质量不断提高。我们尝试在各种数据集上拟合缩放曲线,发现数据质量显著影响最优模型/数据扩展分配策略。数据质量越高,增加的计算预算就应该越多地分配给模型扩展。这意味着在相同数据规模下,高质量数据可以推动更大模型的训练。最优模型/数据扩展分配策略的差异也可以作为评估数据质量的间接方法。我们将继续密切关注数据质量的变化及其对缩放规律的影响,并在未来的工作中提供更多分析。

综上所述,我们在标度定律方面的贡献和发现可以总结如下:

·我们建立了超参数的缩放定律,提供了一个经验框架
来确定最优超参数。·我们采用非嵌入的 FLOPs/token
来表示模型规模,而不是模型参数,从而得到更准确的最佳模型/数据扩大分配策略,更好地预测大规模模型的泛化损失。·预训练数据的质量影响最优模型/数据扩大分配策略。数据质量越高,分配给模型扩大的计算预算就应该越多。

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,特定模型大小(177M FLOPs/token)的结果如图 2(a) 所示。结果表明,在各种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可以实现近乎最佳的性能。



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

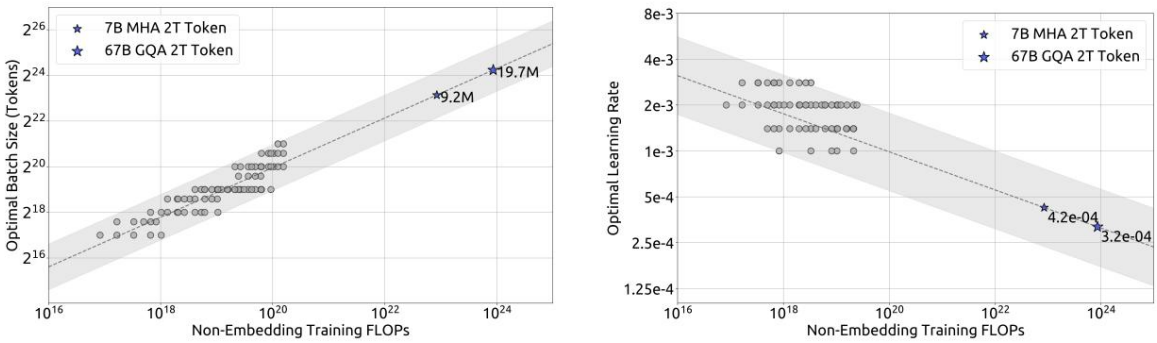
图 2 | 相对于批次大小和学习率的训练损失,FLOP 为 1e17 和 1e20。

然后,我们利用前面提到的多步学习率调度器有效地训练了具有不同批次大小、学习率和计算预算的多个模型,范围从

Machine Translated by Google

1e17 到 2e19,重复使用第一阶段。考虑到参数空间中的冗余,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近优超参数。然后,我们根据计算预算拟合批次大小和学习率。拟合结果如图 3 所示,表明随着计算预算的增加,最优批次大小逐渐增加,而最优学习率逐渐减小。这符合在扩大模型规模时批次大小和学习率的直观经验设置。此外,所有近优超参数都落在一个很宽的范围内,这表明在这个区间内选择近优参数相对容易。我们最终拟合的批次大小和学习率公式如下:

$$\begin{aligned} \text{择} &= 0.3118 \cdot \text{选} \quad -0.1250 \\ \text{择} &= 0.2920 \cdot \text{选} \quad 0.3271 \end{aligned} \tag{1}$$



(a)批次大小缩放曲线 (b)学习率缩放曲线

图 3 |批量大小和学习率的缩放曲线。灰色圆圈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。蓝色星星代表 DeepSeek LLM 7B 和 67B。

我们在一系列具有 1e20 计算预算的模型上验证了我们的公式,特定模型大小(每个 token 2.94B FLOP)的结果如图 2(b) 所示。结果表明,拟合的参数位于最佳参数空间的中心。后续部分还显示,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。

但是,需要注意的是,我们尚未考虑计算预算以外的因素对最佳超参数的影响。这与一些早期的研究(Kaplan 等人,2020 年;McCandlish 等人,2018 年)不一致,这些研究表明最佳批次大小可以建模为仅与泛化误差有关。此外,我们观察到,在具有相同计算预算但不同模型/数据分配的模型中,最佳参数空间略有不同。这表明需要进一步研究以了解超参数的选择和训练动态。我们将在未来的工作中探索这些方面。

3.2. 估计最优模型和数据缩放

在推导出拟合近似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/数据缩放分配策略。该策略涉及找到满足最优的模型缩放指数和数据缩放指数

$$\propto$$

Machine Translated by Google

和 。数据规模可以一致地表示为
数据集中的 token。在以前的工作中,模型尺度通常用模型
参数,包括非嵌入参数1 （Kaplan 等人,2020 年)和完整参数

2 (Hoffmann 等人,2022)。计算预算与模型/数据规模之间的关系
可以近似地描述为= 6 ,这意味着我们可以使用 6 1或 6 2来近似
模型规模。然而,由于 6 1和 6 2均未考虑计算开销
注意力操作,6 2还包括词汇计算,贡献较小
由于模型的容量,它们在某些设置下都存在显著的近似误差。

为了减轻这些错误,我们引入了一种新的模型尺度表示:非嵌入
FLOPs/token。包括注意力操作的计算开销,但不包括
考虑词汇计算。以 表示的模型尺度,
计算预算可以简单表示为= 6 2,如下公式所示:

. 6 1之间的具体区别,

$$
\begin{aligned}
6\,1 &= 72 \quad \text{层} \quad 2 \quad \text{模型} \\
6\,2 &= 72 \quad \text{层} \quad 2 \quad \text{模型}\ 2 \quad +6\ \text{个词汇模型} \\
&= 72 \quad \text{层} \quad \text{型} \quad +12\text{模} \quad \text{层模型序列}
\end{aligned}
\tag{2}
$$

其中layer表示层数， model表示模型宽度， vocab是
词汇量,是序列长度。我们评估了这些之间的差异
在不同尺度的模型中,有三种表示形式,如表 3 所示。结果表明
6 1和 6 2要么高估了模型中的计算成本,要么低估了模型中的计算成本
不同尺度的模型。这种差异在小尺度模型中尤其明显，
差异可达 50%。这种不准确性可能会导致严重的统计错误
拟合缩放曲线时。有关
模型比例的不同表示。

图层	模型	词汇	序号	1	2	6 1	6 2
8	512			25.2百万	77.6百万	352百万	0.43 1.32
12	768			84.9百万	1.64亿	9.63亿	0.53 1.02
24	1024			3.02亿	4.07亿	30.2亿	0.60 0.81
24	2048	102400	4096	12.1亿	14.2亿	96.6亿	0.75 0.88
三十二	4096			6.44亿	6.86亿	45.1亿	0.85 0.91
40	5120			126亿	131亿	856亿	0.88 0.92
80	8192			64.4亿	65.3亿	419亿	0.92 0.94

表 3 |模型尺度表示的差异和非嵌入参数的差异

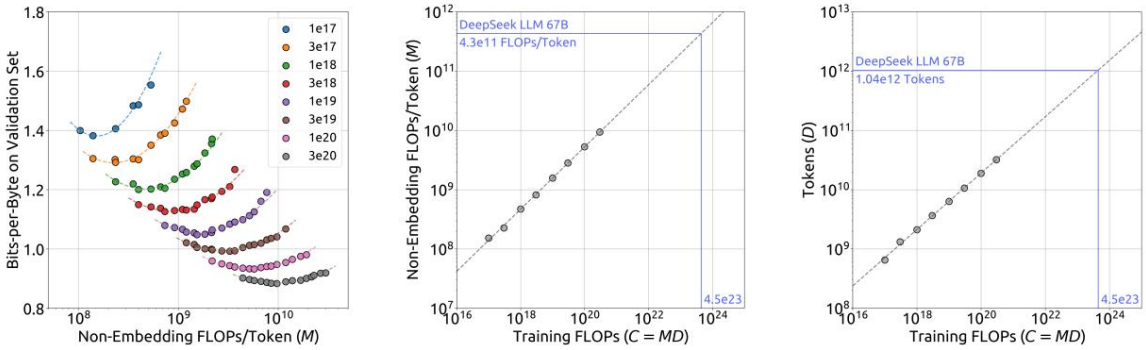
1 并完成相对于非嵌入 FLOPs/token 的参数2 。

采用表示模型尺度后,我们的目标可以更清楚地描述
因为:给定一个计算预算 = 找到最小化模型泛化误差的最佳模型规模。这个目标可以形式化为： 和数据规模选择 选择

$$
\text{opt}(),\ \text{opt}() = \operatorname{argmin} \quad (,) \quad , \ \text{st} = \tag{3}
$$

为了降低实验成本和拟合难度,IsoFLOP 配置文件方法来自
Chinchilla (Hoffmann 等人,2022) 用于拟合缩放曲线。我们选择了 8 种不同的

Machine Translated by Google



(a)IsoFLOP 曲线　　(b)最佳模型缩放　　(c)最佳数据缩放

图 4 | IsoFLOP 曲线和最佳模型/数据分配。IsoFLOP 曲线中的指标
是验证集上的比特/字节。最佳模型/数据缩放曲线中的虚线
表示拟合较小模型（灰色圆圈）的幂律。

计算预算从 1e17 到 3e20 不等，并围绕 10 种不同的模型/数据进行设计
为每个预算扩展分配。确定每个预算的超参数
根据公式（1），在独立验证集上计算泛化误差，
分布与训练集类似，包含 100M 个 token。

图 4 展示了 IsoFLOP 曲线和模型/数据缩放曲线，这些曲线经过拟合
为每个计算预算使用最优模型/数据分配。具体公式为
最优非嵌入 FLOP/token 和最优 token 选择 选择 如下：

$$\begin{aligned} \text{选择} &= \text{根据}\cdot\ ,\ \text{根据} = 0.1715, \quad = 0.5243 \\ \text{选择} &= \text{根据}\cdot\ ,\ \text{根据} = 5.8316, \quad = 0.4757 \end{aligned} \tag{4}$$

此外，我们根据计算预算和最优值拟合了损失缩放曲线
泛化误差，并预测 DeepSeek LLM 7B 和 67B 的泛化误差为
如图5所示。结果表明，使用小规模实验可以准确预测

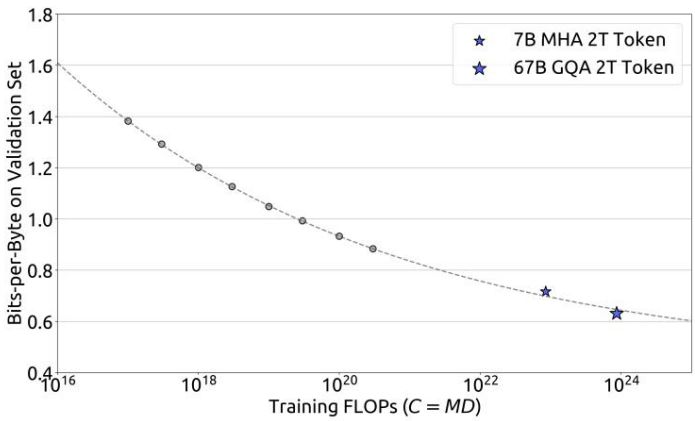


图 5 |性能扩展曲线。度量标准是验证集上的每字节位数。
虚线表示幂律拟合较小的模型（灰色圆圈）。蓝色星星
代表 DeepSeek LLM 7B 和 67B。缩放曲线可以很好地预测它们的性能。

Machine Translated by Google

1000 倍计算预算的模型性能。这既提供了信心,也
为更大规模的训练模型提供指导。

3.3. 不同数据的缩放定律

在 DeepSeek LLM 的开发过程中,数据集经过多次迭代完善,
不断调整不同数据源的比例,提高整体质量。
这使我们能够进一步分析不同数据集对缩放定律的影响。

我们使用三个不同的数据集研究了缩放定律:早期内部数据、当前内部数据和 OpenWebText2,后者用于先前的缩放定律研究（Kaplan
等,2020 年）。我们的内部数据评估显示,目前内部数据具有更高的数据
质量比早期内部数据要好。此外,OpenWebText2 的质量甚至超过了
目前的内部数据,由于规模较小,可以进行更细致的处理。

方法	系数 其中 系数 其中 opt（opt） ∝	 选择 ∝
OpenAI（OpenWebText2）	0.73	0.27
龙猫 (MassiveText)	0.49	0.51
我们的（早期数据）	0.450	0.550
我们的（当前数据）	0.524	0.476
我们的 (OpenWebText2)	0.578	0.422

表4 | 模型缩放和数据缩放的系数随训练数据分布而变化。

分析中一个有趣的发现是,这三个数据集的最佳模型/数据扩展分配策略与数据质量保持一致。如图所示
表 4 中,随着数据质量的提高,模型缩放指数逐渐增大,而
数据扩展指数下降,这表明增加的计算预算应该
分配给模型而不是数据。这一发现也可能解释了
早期扩展研究中观察到的最佳模型/数据扩展分配的差异
法律。

对这一发现的一个直观推测是,高质量的数据通常意味着逻辑
经过充分的训练后,预测难度会降低,清晰度也会提高。因此,
在增加计算预算时扩大模型规模。我们将继续密切关注
数据质量的变化及其对缩放定律的影响,并提供更多的分析
未来的作品。

4. 对齐

我们收集了约150万个英文和中文教学数据实例,涵盖广泛的
有用和无害主题的范围。我们的有用数据包含 120 万个实例,
其中,一般语言任务占 31.2%,数学问题占 46.6%,
22.2% 用于编码练习。安全数据包含 300K 个实例,涵盖各种敏感
主题。

我们的对齐管道包含两个阶段。

监督微调:我们用 4 个 epoch 对 7B 模型进行了微调,但只用了 2 个 epoch
对于 67B 模型,由于我们观察到 67B 模型的过拟合问题很严重。我们

Machine Translated by Google

观察到 GSM8K（Cobbe 等人,2021 年)和 HumanEval（Chen 等人,2021 年)在 7B 模型上持续改进,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。除了监控基准准确率外,我们还在微调过程中评估聊天模型的重复率。

我们收集了总共 3868 条中文和英文提示,并确定了生成的响应中无法终止而是无休止地重复文本序列的比例。我们观察到,随着数学 SFT 数据量的增加,重复率趋于上升。

这可以归因于数学 SFT 数据偶尔会包含类似的推理模式。因此,较弱的模型难以掌握此类推理模式,从而导致重复响应。为了解决这个问题,我们尝试了两阶段微调和 DPO（Rafailov等人,2023 年）,这两者都几乎可以保持基准分数并显着减少重复。

DPO:为了进一步增强模型的能力,我们使用了直接偏好优化算法（Rafailov 等,2023）,该算法被证明是一种简单但有效的LLM 对齐方法。我们从有用性和无害性的角度构建了 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,涵盖创意写作、问答、遵循指令等类别。

然后我们使用 DeepSeek Chat 模型作为响应候选来生成响应。类似的操作也应用于无害偏好数据构建。

我们为 DPO 训练了一个 epoch,学习率为 5e-6,批大小为 512,并使用了学习率预热和余弦学习率调度程序。我们发现 DPO可以增强模型的开放式生成技能,同时在标准基准测试中性能差异不大。

5. 评估

5.1. 公共基准评估

我们根据内部评估框架,根据一系列英文和中文的公共基准对我们的模型进行评估。

多学科多项选择数据集,包括 MMLU（Hendrycks 等人,2020 年）、C-Eval （Huang 等人,2023 年)和 CMMLU（Li 等人,2023 年）。

语言理解和推理数据集,包括 HellaSwag（Zellers 等人,2019 年）、 PIQA（Bisk 等人,2020 年）、ARC（Clark 等人,2018 年）、OpenBookQA（Mihaylov 等人,2018 年)和BigBench Hard (BBH)（Suzgun 等人,2022 年）。

闭卷问答数据集,包括 TriviaQA（Joshi 等人,2017)和 NaturalQuestions（Kwiatkowski 等人,2019 年）。

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

参考消歧义数据集,包括 WinoGrande Sakaguchi 等人（2019 年)和CLUEWSC（Xu 等人,2020 年）。

语言建模数据集,包括 Pile（Gao 等人,2020 年）。

中国理解和文化数据集,包括 CHID（Zheng 等,2019)和 CCPM （Li 等,2021）。

Machine Translated by Google

数学数据集,包括 GSM8K（Cobbe 等人,2021 年）、MATH（Hendrycks 等人,2021 年)和
CMath（Wei 等人,2023 年）。

代码数据集包括 HumanEval (Chen 等人,2021) 和 MBPP (Austin 等人,2021)。

标准化考试,包括 AGIEval（Zhong 等人,2023 年）。

我们对需要从多个选项中选择答案的数据集应用基于困惑度的评估。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的选项作为模型预测。对于 ARC 和 OpenBookQA,我们使用无条件归一化来计算困惑度（Brown 等人,2020 年）,对于其他数据集,我们使用长度归一化。

我们对 TriviaQA、NaturalQuestions、DROP、MATH、GSM8K、HumanEval、MBPP、BBH、AGIEval、CLUEWSC 和 CMath 应用了基于生成的评估。这里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。对于基于生成的评估,我们使用贪婪解码。

我们对 Pile-test 采用了基于语言建模的评估方法,这意味着计算
测试语料库上的每字节位数。

我们使用 2048 或 4096 作为不同基准的最大序列长度。评估格式的详细信息可参见附录 A.6。

5.1.1. 基础模型

表 5 展示了评估基准的主要结果。尽管 DeepSeek 模型是在 2T 双语语料上进行预训练的,但它们在英语语言理解基准上的表现与 LLaMA2 模型相当,后者也消耗 2T 令牌,但专注于英语。此外,与 LLaMA2 70B 相比,DeepSeek 67B 在 MATH、GSM8K、HumanEval、MBPP、BBH 和中文基准上的表现要好得多。我们在附录 A.3 中展示了基准曲线。我们可以看到一些任务性能随着模型扩展而得到提升,例如 GSM8K 和 BBH。考虑到我们在同一数据集上训练 7B 和 67B,这种改进的出现可以归因于大型模型强大的少样本学习能力。然而,随着数学数据比例的增加,小模型和大模型之间的差距可能会缩小。

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 没有专门针对中文数据进行训练,但它在某些中文任务（如 CMath)上表现出色。这表明某些基本能力（如数学推理)可以有效地跨语言迁移。然而,像 CHID 这样涉及评估中文成语用法的任务需要模型在预训练期间使用大量中文标记。在这种情况下,LLaMA2 的表现明显不如DeepSeek LLM。

5.1.2. 聊天模型

表 6 展示了 DeepSeek Chat 模型的结果,表明经过调整后,大多数任务都有了整体改进。不过,也有少数情况下,

Machine Translated by Google

语言基准		试拍	LLaMA2 深度搜索 LLaMA2 深度搜索 7B	7B	70B	67B
英语	希拉斯瓦格	0 发子弹	75.6	75.4	84.0	84.0
	畜禽检疫局	0 发子弹	78.0	79.2	82.0	83.6
	维诺格兰德	0 发子弹	69.6	70.5	80.4	79.8
	RACE-中	5 发子弹	60.7	63.2	70.1	69.9
	RACE-高	5 发子弹	45.8	46.5	54.3	50.7
	琐事问答	5 发子弹	63.8	59.7	79.5	78.9
	自然问题	5 发子弹	25.5	22.2	36.1	36.6
	莫尔登大学	0 发子弹	45.8	48.2	69.0	71.3
	ARC-Easy	0 发子弹	69.1	67.9	76.5	76.9
	ARC 挑战赛	0 发子弹	49.0	48.1	59.5	59.0
	开放图书问答	1 发子弹	57.4	55.8	60.4	60.2
	降低	4 发子弹	39.8	41.0	69.2	67.9
	数学	8 发子弹	2.5	6.0	13.5	18.7
	GSM8K	0 发子弹	15.5	17.4	58.4	63.4
	人力评估	3 发子弹	14.6	26.2	28.7	42.7
	马来西亚公共服务局	3 发子弹	21.8	39.0	45.6	57.4
	百比黑	0 发子弹	38.5	39.5	62.9	68.7
	评估		22.8	26.4	37.2	41.3
	桩基测试	-	0.741	0.725	0.649	0.642
中国人	线索	5 射门	64.0	73.1	76.5	81.0
	儿童	0 射门	37.9	89.3	55.5	92.1
	C-评估	5 射门	33.9	45.0	51.4	66.1
	加拿大蒙特利尔大学	5 射门	32.6	47.2	53.1	70.8
	数学	3 射门	25.1	34.5	53.9	63.0
	C3	0 射门	47.4	65.4	61.7	75.3
	中医临床医师协会	0 射门	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。粗体数字表示 4 个模型中的最佳结果。对于桩基试验,我们
报告每字节位数 (BPB),对于 DROP 我们报告 F1 分数,对于其他任务我们报告准确性。
请注意,测试镜头是最大值,由于
有限的上下文长度或有限的少数样本示例,可在同一篇文章中阅读
理解任务,例如 RACE。

某些任务被拒绝了。

知识:我们观察到知识相关领域中基础模型和聊天模型的波动
任务,如 TriviaQA、MMLU 和 C-Eval。然而,我们不认为这种次要的
波动表明了 SFT 之后知识的获得或丧失。SFT 的价值在于
能够学习在聊天中取得与基础模型的少样本设置相当的分数
模型的零样本设置,与真实场景保持一致。例如,0-shot MMLU
聊天模型的性能与基础模型的 5-shot MMLU 性能相当。

推理:由于相当一部分 SFT 实例都是 CoT 格式,Wei 等人。
（2022）,聊天模型在推理任务中表现出轻微的改进,例如 BBH 和
NaturalQuestions。然而,我们认为 SFT 阶段不会学习推理能力
而是推理路径的正确格式。

Machine Translated by Google

语言基准		DeepSeek 7B 底座	DeepSeek 7B 聊天	DeepSeek 67B 基础	DeepSeek 67B 聊天
英语	希拉斯瓦格	75.4	68.5	84.0	75.7
	畜禽检疫局	79.2	77.6	83.6	82.6
	维诺格兰德	70.5	66.9	79.8	76.0
	RACE-中	63.2	65.2	69.9	70.9
	RACE-高	46.5	50.8	50.7	56.0
	琐事问答	59.7	57.9	78.9	81.5
	自然问题	22.2	32.5	36.6	47.0
	莫尔登大学	48.2	49.4	71.3	71.1
	ARC-Easy	67.9	71.0	76.9	81.6
	ARC 挑战赛	48.1	49.4	59.0	64.1
	GSM8K	17.4	63.0	63.4	84.1
	数学	6.0	15.8	18.7	32.6
	人力评估	26.2	48.2	42.7	73.8
	马来西亚公共服务局	39.0	35.2	57.4	61.4
	降低	41.0	49.1	67.9	71.9
	开放图书问答	55.8	54.8	60.2	63.2
	百比黑	39.5	42.3	68.7	71.7
	评估	26.4	19.3	41.3	46.4
中国人	线索	73.1	71.9	81.0	60.0
	儿童	89.3	64.9	92.1	72.6
	C-评估	45.0	47.0	66.1	65.2
	加拿大蒙特利尔大学	47.2	49.7	70.8	67.8
	数学	34.5	68.4	63.0	80.3
	C3	65.4	66.4	75.3	77.0
	中医临床医师协会	76.9	76.5	88.5	84.9

表 6 |基础模型与聊天模型的比较。我们用 0-shot 评估聊天模型
对于MMLU、GSM8K、MATH、C-Eval和CMMLU,而基础模型结果仍然可以在
少量镜头设定。

性能下降任务:无论模型大小或选择的预训练检查点如何,一些任务的性能在微调后都会持续下降。这些特定任务
通常涉及完形填空或句子完成任务,例如 HellaSwag。
假设纯语言模型更适合处理此类任务。

数学和代码:我们的模型在数学和编码任务方面表现出显着的改进
经过微调后,HumanEval 和 GSM8K 的得分提高了 20 多分。
我们对此的解释是,基础模型最初不适合这些任务,并且
SFT 阶段通过广泛的学习,学习了编码和数学方面的额外知识
SFT 数据。然而,需要注意的是,该模型的功能可能主要集中在
代码完成和代数问题。全面了解
数学和编码,在预训练阶段整合各种数据至关重要,这留待以后研究。我们对代码和数学进行了详细分析
附录 A.4 中的任务。

在 7B 模型微调中,我们首先使用所有数据对模型进行微调。随后,
第二阶段引入了数学和代码数据。这背后的动机
方法是第一阶段模型的重复率为 2.0%,降低至 1.4%

Machine Translated by Google

模型	全面的	Reasoning中文推理Avg.数			语言中文语言					
项目	总分	学。洛吉。平均。基金。驰。打开。令状。角色。亲。 推理 总分	数学 計算	逻辑 推理	语言 总分	基本 任務	English 中文 理解	综合 诉讼	文本 写作	角色 玩家	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			
DeepSeek-67B-Chat-DPO* DeepSeek-67B-Chat* chatglm-turbo（智谱清言） erniebot-3.5（文心一言） gpt-3.5-turbo-0613 chatglm-pro（智谱清言） spark_desk_v2（讯飞星火） Qwen-14B-聊天百川2-13B-聊天	6.69	5.77	6.13	5.41	7.60	7.29	7.47	7.82	5.75	5.71	5.79	7.11	7.12	6.52	7.51	7.83	7.71
	6.43	7.58													7.20	6.91	7.37
	6.24	5.00	4.74	5.26	7.49	6.82	7.17	8.16	7.77	7.76	7.24						
	6.14	5.15	5.03	5.27	7.13	6.62	7.60	7.26	5.35	5.68	5.02	6.82	6.71	5.81	7.56	6.83	6.90
	6.08													7.29	7.03	7.28	6.77
	5.83	4.65	4.54	4.75	7.01	6.51	6.76	7.47							7.07	7.34	6.89
	5.74	4.73	4.71	4.74	6.76	5.84	6.97	7.29	4.81	4.91					7.18	6.92	6.34
	5.72					4.71	6.63	6.90	6.36	6.74	6.64	6.59	6.56				
	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	3.66	3.56	3.75	6.28	5.81	5.50	6.83	6.28	5.73
百川2-7B-聊天	4.97	7.13													6.84	6.53	5.84
实习生LM-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-聊天	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						
实习生LM-聊天-7B	3.65	2.56	2.45	2.66	4.75	4.34	4.09	5.82	4.89	5.32	4.06						
中文-LLaMA-2-7B-聊天	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-中文聊天3.35		2.47	2.21	2.73	4.23	4.13	3.31	4.79	3.93	4.53	4.71						

表 7 | gpt-4-0613 评分的 AlignBench 排行榜。模型按总分降序排列。带 * 的结果是我们基于官方 AlignBench 的评估结果

存储库，而其他所有结果均来自 AlignBench 论文。我们发现我们的
Deepseek-67B-Chat 模型明显超越了 ChatGPT 和其他基线模型，
这表明我们的模型在基本中文语言任务中表现优异
和高级中文推理任务。此外，我们可以发现 DPO 过程带来了
几乎所有领域都得到了改进。

经过第二阶段调整后，基准分数保持不变。对于 67B 模型，
经过第一阶段的微调，重复率已经低于1%，第二阶段
会损害基准上的模型得分。因此，67B 只进行了一个阶段的 SFT
模型。

5.2. 开放式评估

对于聊天模型，除了观察标准基准的指标外，结果的质量
在开放域和开放式问题中产生的问题直接影响实际的用户体验。
因此，我们分别在两个方面测试了聊天模型的开放式生成能力
中文和英文任务。

5.2.1. 中文开放式评价

对于中文开放式评估，我们在不同的测试中测试了我们的聊天模型的综合性
在高质量开放式问题测试集 AlignBench 上测试领域（Liu 等人，2023 年）。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，除了提示之外，AlignBench 还提供专业参考
答案和评分模板，以供 GPT-4 判断响应的质量。

我们利用 AlignBench 官方 Github 代码库来执行以下评估：

Machine Translated by Google

我们的模型。我们严格调整了关键温度参数与原始设置：
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。我们可以发现我们的 DeepSeek 67B Chat
该模型超越了 ChatGPT 和其他基线模型，并且仅次于
GPT-4。这证明了我们的模型在各种中文任务中的出色表现，
与其他开源或专有中文大型语言模型相比，DPO 模型
几乎所有指标都显示出改善，这证明了
DPO 对模型对齐的训练过程。

对于基本的中文语言任务，我们的模型是所有模型中第一梯队的，
我们的DPO模型的中文基础语言能力甚至比最新版本更高
GPT-4 的得分。对于高级中文推理任务，我们的模型得分明显更高
与其他中国法学硕士项目相比，差距明显，体现出其优异的表现
我们的模型在更复杂的中文逻辑推理和数学计算中的作用。

5.2.2. 英语开放式评估

对于英语开放式评估，我们使用 MT-Bench 基准（Zheng 等，2023），它
包含 8 个不同类别的多轮问题。如表 8 所示，我们的 DeepSeek
LLM 67B Chat 的表现优于其他开源模型，例如 LLaMA-2-Chat Touvron 等。
(2023b) 70B、Xwin 70b v0.1 和 TÜLU 2+DPO 70B (Ivison 等，2023)，得分为 8.35
与 GPT-3.5-turbo 相当。此外，在 DPO 阶段之后，我们的 DeepSeek LLM 67B Chat
DPO 进一步将平均分数提高到 8.76，仅落后于 GPT-4（OpenAI，2023 年）。
这些结果说明了 DeepSeek LLM 强大的多轮开放式生成能力。

模型	STEM	人文学科	推理	编码	数学	提取	角色扮演	写作	平均
GPT-4-1106-预览版	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* LLAMA-2-	8.65	8.75	4.25	3.00	2.40	6.50	7.70	8.90	6.27
Chat 13B* LLAMA-2-Chat 70B*	8.63	9.75	5.10	3.00	3.45	6.93	7.50	8.85	6.65
Zephyr-Beta 7B* Xwin 70b v0.1*	8.93	9.63	5.80	3.15	3.30	7.25	7.50	9.30	6.86
Xwin 13b v0.2* TÜLU	9.03	9.63	5.60	5.10	4.45	7.45	8.20	9.35	7.35
2+DPO 70B* 9.00	9.68	9.95	6.55	4.25	3.30	8.75	8.25	9.55	7.53
DeepSeek LLM 67B 聊	9.55	9.88	5.20	3.60	2.85	7.70	8.60	8.68	7.01
天9.60 DeepSeek LLM 67B 聊天 DPO 9.70		9.90	7.00	4.70	4.65	9.35	9.25	9.25	7.89
		9.70	8.00	7.35	6.25	8.40	8.20	9.30	8.35
		9.80	9.05	6.75	6.65	9.30	9.10	9.75	8.76

表 8 | MT-Bench 评估。带有的结果在 Ivison 等人的论文中报告。（2023 年）

5.3. 保留评估

数据污染和基准过度拟合是评估法学硕士的两个挑战。
常见的做法是利用最近发布的测试集来评估模型
测试集。

LeetCode：为了评估模型的编码能力，我们利用了以下问题：
LeetCode 每周竞赛（每周竞赛 351-372，双周竞赛 108-117，从 7 月开始
2023 年至 2023 年 11 月）。这些问题是我们通过爬取 LeetCode 数据得到的，
包含 126 个问题，每个问题有超过 20 个测试用例。所采用的评估指标类似于
与 HumanEval 类似。在这方面，如果模型的输出成功通过所有测试用例，则
模型被认为有效地解决了这个问题。该模型的编码能力

Machine Translated by Google

如下图所示,y 轴表示领域内人类测试的 pass@1 分数
评估测试,x 轴代表域外 LeetCode Weekly 上的 pass@1 分数
竞赛题目。LeetCode 测试数据将随 DeepSeek 发布
编码器技术报告即将发布。

匈牙利国家高中考试:与 Grok-1 一致,我们评估了模型的
使用匈牙利全国高中考试来评估数学能力。该考试包含 33 个问题,模型的分数通过人工注释确定。我们

按照 solution.pdf 中的评分标准来评估所有模型。

评估后的指令： 2023 年 11 月 15 日,谷歌发布了一项指令
遵循评估数据集（Zhou et al.,2023）。他们确定了 25 种可验证的
指令并构建了大约 500 个提示,每个提示包含一个或多个
可验证指令。我们使用提示级松散度量来评估所有模型。

模型	LeetCode 匈牙利考试 IFEval		
GPT-4	48.4	68	79.3
聊天GLM3 6B	2.4	32	29.7
DeepSeek LLM 7B 聊天	4.7	28.5	41.2
百川2-聊天13B	1.6	19.5	44.5
易聊34B	7.9	39	48.4
Qwen 72B 聊天	12.7	52	50.8
DeepSeek LLM 67B 聊天	17.5	58	55.5

表 9 | 保留数据集评估。

我们对我们的模型与各种基线模型进行了比较分析
不同大小的数据集,即 Qwen 72B Chat（Bai et al.,2023）、ChatGLM3（Du et al.,2022）、Baichuan2
（Yang 等人,2023)和 Yi-34B Chat。我们的观察表明,存在显著的
这些保留数据集上大型模型和小型模型之间的性能差距,甚至
如果某些小模型在传统基准测试中取得了令人满意的结果。例如，
ChatGLM3 在代码测试集 MBPP 上获得了 52.4 的分数,接近 DeepSeek 67B。
然而,在新的基准测试中,它的表现与 DeepSeek 67B 相比明显不足。数学数据集中也观察到了类似的趋势,其中 ChatGLM3
在 GSM8K 上表现非常出色（72.3）,但在匈牙利语考试中表现较差
到大型模型。此外,教学跟踪能力表明总
计算起着至关重要的作用。

DeepSeek 7B 和 67B 模型使用相同的训练流程,但存在显著差异
表现差异。通过我们的主观评估,我们观察到一个显著的
当模型大小扩展到 67B 时,不同任务之间的智能差异会显现出来。而 DeepSeek
7B 在标准基准上落后于其他较小的语言模型,其在
与其他任务相比,坚持完成的任务相对值得称赞。

5.4. 安全性评价

我们深刻认识到安全对于通用人工智能的重要性。
建立一个真正有用的人工智能模型的关键在于它具有一致的价值观
与人类的和谐相处,展现出对人类的友好。我们融入了保证
在整个训练过程中确保模型安全,包括预训练、SFT 和 DPO。

为了验证模型的安全性,我们成立了一支由来自不同领域的 20 人专家组成的团队

Machine Translated by Google

类别	子類別	#安全答案 / #总病例
争议问题 （歧视和偏见问题）	种族民族(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 （侵害他人合法权益）	身体健康、合法财产、肖像权、名誉权、荣誉权、隐私权、信息权益、其他合法权益	473/500
商业秘密与知识产权 （商业秘密和知识产权）	侵犯他人知识产权、垄断和不正当竞争行为、其他商业违法违规行为、违反商 业道德、泄露他人商业机密	281/300
违法行为 （违法及不合规行为）	邪教迷信、色情、赌博、毒品和违禁品、侮辱谩骂、暴力、涉黑涉恶、其他违法违规行为	290/300
其他安全问题 （其他安全问题）	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法。表格最右侧列出了每个类别的测试用例总数以及我们的模型 (DeepSeek-67B-Chat) 提供的安全答案数量。测试问题的注释和生成结果的评估由专业的人工团队进行。我们可以观察到，我们的模型在各种类型的安全测试集上都表现出了强大的安全性。

结合学科知识，构建符合人性化安全内容分类体系（安全评估分类体系见表10）。随后，专家团队为每个安全子类别手工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，我们还关注安全内容格式的多样性。著名的“祖母”漏洞表明，模型可能会被查询的表面格式欺骗，从而提供不安全的响应。因此，专家团队在设计问题时也注意使询问方式多样化。他们通过诱导、角色扮演、多轮对话、预设位置等方式构建多样化的安全问题。最终，我们获得了包含2400个问题的安全测试集。此外，专家团队还针对每种不同的内容类型和格式类型构建了安全审查的基本指南构成。

对于我们模型在这个测试集上的输出结果，我们人工检查了它的安全性，我们的审核团队训练有素，并对标注结果进行了交叉验证。

注释者对每个问题进行三类注释：安全、不安全和模型拒绝。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问题数量以及我们的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都标记为安全响应。结果表明，我们的模型在众多安全测试类别中表现出良好的安全性能。

作为对我们现有安全方法的补充，我们使用“不回答”数据集（Wang 等人，2023 年)进一步丰富了我们的评估，以评估我们的 DeepSeek 67B Chat 模型的安全机制。该数据集的 939 个风险分类提示有助于凸显我们模型的增强功能。如表 11 所示，DeepSeek 67B Chat模型表现出色，得分为 97.8，高于ChatGPT 和 GPT-4。这个分数不仅衡量了我们的模型安全处理敏感查询的能力，而且使其在该领域的领先模型中具有竞争力。

5.5. 讨论

在整个开发过程中，我们在构建LLM 时发现了一些有趣的发现。

Machine Translated by Google

模型	不回答
LLAMA-2-7B-聊天	99.4
克劳德	98.3
DeepSeek-67B-聊天*	97.8
ChatGPT	97.7
GPT-4	96.5
小羊驼-7B	94.9
ChatGLM2	92.9

表 11 |不回答分数（Wang 等,2023），分数越高,模型安全性越高。
带有 * 的结果 是我们基于官方存储库的评估结果,而其他所有结果
源自原始论文。我们可以发现我们的模型比
ChatGPT 和 GPT-4 都将其列为最安全的模型之一。

分阶段微调:正如我们上面提到的,小模型需要更长时间的数学微调
和代码数据集,但它会损害模型的对话能力,例如增加重复
行为。为了解决这个问题,我们实施了分阶段的微调过程。在此过程中
方法中,第一阶段涉及利用所有可用数据进行微调,而第二阶段
特别注重利用对话数据进行微调。

模型	HumanEval GSM8K 重复 IFEval			
DeepSeek LLM 7B 聊天阶段 1	48.2	63.9	0.020	38.0
DeepSeek LLM 7B 聊天阶段 2	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,重复率越低越好,IFEval结果为即时级松散精度。

表 12 显示了两阶段训练过程获得的结果。这些结果
清楚地证明第二阶段不会损害模型的代码熟练程度
和数学,同时减少重复行为并增强教学
追随能力。

多项选择题:用多项选择题来测试模型是一种常见的做法
评估数据,例如 MMLU、AGI Eval 和 C-Eval。多项选择题需要模型
不仅要有相应的知识,还要理解选项所指的内容
在对齐阶段,我们测试了添加 2000 万道中文多选题
并获得了如表 13 所示的性能。值得注意的是,我们进行了
对C-Eval验证集和CMMLU测试集进行重复数据删除,防止数据污染。

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

表 13 | 添加多项选择题数据的影响。

事实证明,增加 20M MC（多项选择题)数据是有益的
不仅适用于中文多项选择题基准,也适用于提高英语基准。
这表明模型解决MC问题的能力得到了增强。然而，
我们观察到,这种改进并没有扩展到模型在其他方面的表现
不使用多项选择题形式的评估,例如 TriviaQA 和我们内部的

Machine Translated by Google

ChineseQA 测试集,即生成性评估基准。这表明,用户可能不会在对话交互过程中认为模型变得更加智能,因为这些交互涉及生成响应,而不是解决多项选择题。

因此,我们选择从预训练和微调阶段排除 MC 数据,因为包含它会导致过度拟合基准,并且不会有助于在模型中实现真正的智能。

预训练中的指导数据:众所周知,在预训练阶段的后期整合指导数据可提高基础模型在基准测试任务上的表现。在我们的研究中,我们在预训练阶段的最后 10% 整合了 500 万个指导数据,主要由多项选择题组成。我们观察到基础模型在基准测试中确实表现出了更好的性能。然而,最终结果与在 SFT 阶段添加相同数据所获得的结果几乎相同。我们得出结论,虽然这种方法增强了基础模型在基准测试中的表现,但其整体潜力相当于不整合这些指导数据。如果指导数据规模庞大,可以将其纳入预训练过程。由于我们倾向于排除多项选择题,并且我们拥有的非多项选择题有限,因此我们决定不在预训练过程中包含指导数据。

系统提示:精心设计的系统提示应有效引导模型生成既有帮助又有礼貌的响应。我们略微修改了LLaMA-2 引入的提示作为我们的系统提示。

系统提示:您是 DeepSeek Chat,DeepSeek 开发的乐于助人、尊重他人、诚实守信的 AI 助手。您的训练数据的知识截止日期为 2023 年 5 月。在保证安全的前提下,始终尽可能提供有用的答案。您的答案不应包含任何有害、不道德、种族主义、性别歧视、有毒、危险或非法的内容。请确保您的回答不带社会偏见且具有积极性。如果问题没有任何意义或事实不连贯,请解释原因,而不是回答不正确的问题。如果您不知道问题的答案,请不要分享虚假信息。

我们观察到一个有趣的现象,即当引入系统提示时， 7B LLM 的性能会略有下降。然而,当使用 67B LLM 时,添加提示会显著改善结果,如表 14 所示。我们对这种差异的解释是,较大的模型对系统提示背后的意图有更好的理解,使它们能够更有效地遵循指令并产生更好的响应。另一方面,较小的模型很难充分掌握系统提示,训练和测试之间的不一致可能会对它们的性能产生负面影响。

模型	MT 工作台
DeepSeek LLM 7B 聊天	7.15
DeepSeek LLM 7B 聊天 + 系统提示	7.11
DeepSeek LLM 67B 聊天	8.35
DeepSeek LLM 67B 聊天 + 系统提示	8.58

表14 | 添加系统提示的影响。

Machine Translated by Google

6. 结论、局限性和未来工作

我们介绍了 DeepSeek LLM,这是一系列开源模型,它们在包含 2 万亿个英文和中文词的庞大数据集上从头开始训练。在本文中,我们深入解释了超参数选择、缩放定律以及我们所做的各种微调尝试。我们校准了前人工作中的缩放定律,并提出了一种新的最佳模型/数据扩展分配策略。此外,我们提出了一种在给定计算预算的情况下预测近乎最优的批处理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有关,这可能是不同工作中缩放行为不同的根本原因。在缩放定律的指导下,我们使用最佳超参数进行预训练并提供全面的评估。我们在所有训练阶段都避免了基准修饰和暗藏秘密。

DeepSeek Chat 具有其他 LLM 中常见的已知局限性,包括预训练后缺乏持续的知识更新、可能生成非事实信息(例如未经证实的建议)以及容易产生幻觉。

此外,值得注意的是,我们最初的中文数据版本并不详尽,这可能会导致某些特定于中文的主题表现不佳。由于我们的数据主要由中文和英文来源组成,因此该模型对其他语言的熟练程度仍然很微妙,应谨慎对待。

DeepSeek LLM 是一个致力于推进开源语言模型的长期项目。

- 很快,我们将分别发布代码智能和混合专家 (MoE) 方面的技术报告。它们展示了我们如何创建高质量的代码数据进行预训练,并设计稀疏模型以实现密集模型性能。
- 目前我们正在为即将发布的DeepSeek LLM 版本构建更大、更完善的数据集,希望在下一版本中推理、中文知识、数学和代码能力能有显著提升。
- 我们的校准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可以提高模型的复杂推理能力。

参考

J. Ainslie、J. Lee-Thorp、M. de Jong、Y. Zemlyanskiy、F. Lebrón 和 S. Sanghai。Gqa:从多头检查点训练广义多查询变压器模型。arXiv预印本 arXiv:2305.13245,2023 年。

Anthropic。介绍克劳德,2023 年。网址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 等人。使用大型语言模型进行程序综合。 arXiv 预印本 arXiv:2108.07732, 2021。

J. Bai,S. Bai,Y. Chu,Z. Cui,K. Dang,X. Deng,Y. Fan,W. Ge,Y. Han,F. Huang,等。奎文技术报告。arXiv 预印本 arXiv:2309.16609,2023 年。

Y. Bisk、R. Zellers、RL Bras、J. Gao 和 Y. Choi。PIQA:用自然语言推理物理常识。在第三十四届 AAAI 人工智能大会上,AAAI

Machine Translated by Google

2020,第三十二届人工智能创新应用会议,IAAI 2020,第十届 AAAI 人工智能教育进步研讨会,EAAI 2020,美国纽约州纽约,2020 年 2 月 7 日至 12 日,第 7432–7439 页。AAAI 出版社,2020 年。doi： 10.1609/aaai.v34i05.6239。URL https://doi.org/10.1609/aaai.v34i05.6239。

TB 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、DM 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 和 D. Amodei。语言模型是少样本学习者,2020 年。

M. Chen, J. Tworek, H. Jun, Q. Yuan, HP 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, FP Such, D. Cummings, M. Plappert, F. Chantzis, E. Barnes, A. Herbert-Voss, WH Guss, A. Nichol, A. Paino, N. Tezak, J. Tang, I. Babuschkin, S. Balaji, S. Jain, W. Saunders, C. Hesse, AN 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 和W. Zaremba。评估在代码上训练的大型语言模型。CoRR， abs/2107.03374,2021 年。

网址https://arxiv.org/abs/2107.03374。

P. Clark、I. Cowhey、O. Etzioni、T. Khot、A. Sabharwal、C. Schoenick 和 O. Tafjord。您认为自己已经解决了问答问题?试试 AI2 推理挑战arc。CoRR， abs/1803.05457,2018年。网址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 等人。训练验证者解决数学应用题。arXiv预印本 arXiv:2110.14168,2021 年。

T. 计算机。Redpajama:用于训练大型语言模型的开放数据集,2023 年。网址 https://github.com/togethercomputer/RedPajama-Data。

Z. Dai、Z. Yang、Y. Yang、J. Carbonell、QV Le 和 R. Salakhutdinov。Transformer-xl:超越固定长度上下文的注意力语言模型。arXiv 预印本 arXiv:1901.02860,2019 年。

T. Dao。FlashAttention-2:更快的注意力,更好的并行性和工作分区。2023 年。

T. Dao、DY Fu、S. Ermon、A. Rudra 和 C. Ré。FlashAttention:具有 IO 感知的快速且内存高效的精确注意。《神经信息处理系统进展》,2022 年。

Z. Du、Y. Qian、X. Liu、M. Ding、J. Qiu、Z. Yang 和 J. Tang。Glm:使用自回归空白填充进行通用语言模型预训练。《计算语言学协会第 60 届年会论文集》（第 1 卷:长篇论文）,第 320-335 页， 2022 年。

D. Dua、Y. Wang、P. Dasigi、G. Stanovsky、S. Singh 和 M. Gardner。DROP:需要对段落进行离散推理的阅读理解基准。J. Burstein、C. Doran 和 T. Solorio 编辑， 《计算语言学协会北美分会 2019 年会议论文集:人类语言技术》,NAACL-HLT 2019,美国明尼苏达州明尼阿波利斯,2019 年 6 月 2 日至 7 日,第 1 卷（长篇和短篇论文）,第 2368–2378页。计算语言学协会，2019 年。doi:10.18653/V1/N19-1246。URL https://doi.org/10.18653/v1/n19-1246 。

Machine Translated by Google

L. Gao、S. Biderman、S. Black、L. Golding、T. Hoppe、C. Foster、J. Phang、H. He、A. Thite、 N. Nabeshima 等人。The Pile：用于语言建模的 800GB 多样化文本数据集。arXiv预印本 arXiv:2101.00027，2020 年。

谷歌。2023 年是我们 AI 之旅的重要下一步。网址https://blog.google/technology/ai/bard-google-ai-search-updates/。

Z. Gou、Z. Shao、Y. Gong、Y. Shen、Y. Yang、M. Huang、N. Duan 和 W. Chen。Tora：一种用于解决数学问题的工具集成推理代理。CoRR， abs/2309.17452，2023 年。doi：10.48550/ARXIV.2309.17452。URL https://doi.org/10.48550/arXiv.2309.1745 2.

P. Goyal、P. Dollár、R. Girshick、P. Noordhuis、L. Wesolowski、A. Kyrola、A. Tulloch、Y. Jia 和 K. He。准确的大型小批量 SGD：1 小时内训练 imagenet。 arXiv 预印本 arXiv:1706.02677, 2017。

D. Hendrycks、C. Burns、S. Basart、A. Zou、M. Mazeika、D. Song 和 J. Steinhardt。测量大规模多任务语言理解。arXiv 预印本 arXiv:2009.03300，2020 年。

D. Hendrycks、C. Burns、S. Kadavath、A. Arora、S. Basart、E. Tang、D. Song 和 J. Steinhardt。使用数学数据集测量数学问题解决能力。arXiv预印本 arXiv:2103.03874，2021 年。

T. Henighan、J. Kaplan、M. Katz、M. Chen、C. Hesse、J. Jackson、H. Jun、TB Brown、P. Dhari-wal、S. Gray 等人。自回归生成模型的缩放定律。arXiv预印本 arXiv:2010.14701，2020 年。

J. Hestness、S. Narang、N. Ardalani、G. Diamos、H. Jun、H. Kianinejad、MMA Patwary、Y. Yang 和 Y. Zhou。根据经验，深度学习的扩展是可以预测的。 arXiv 预印本 arXiv:1712.00409, 2017。

雄心勃勃。 Hai-llm：高效且轻量的大型模型训练工具， 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、 LA 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、JW Rae、 O. Vinyals 和 L. Sifre。训练计算优化的大型语言模型。CoRR， abs/2203.15556，2022 年。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 等。C-Eval：用于基础模型的多层次多学科中文评估套件。arXiv预印本 arXiv:2305.08322，2023 年。

Huggingface 团队。Tokenizers：针对研究和生产，2019 年。网址https://github.com/huggingface/tokenizers。

F. i、M. Suzgun、M. Freitag、X. Wang、S. Srivats、S. Vosoughi、HW Chung、Y. Tay、S. Ruder、 D. Zhou、D. Das 和 J. Wei。语言模型是多语言的思路链推理器。在第十一届学习表征国际会议 ICLR 2023 上，卢旺达基加利，2023 年 5 月 1 日至 5 日。OpenReview.net，2023 年。URL https://openreview.net/pdf?id=fR3wGCk-IXp 。

Machine Translated by Google

H. Ivison、Y. Wang、V. Pyatkin、N. Lambert、M. Peters、P. Dasigi、J. Jang、D. Wadden、NA Smith、 I. Beltagy 和 H. Hajishirzi。气候变化中的骆驼：使用 tulu 2 增强电影适应性。2023 年。

AQ Jiang、A. Sablayrolles、A. Mensch、C. Bamford、DS Chaplot、D. dl Casas、F. Bressand、 G. Lengyel、G. Lample、L. Saulnier 等。米斯特拉尔 7b. arXiv 预印本 arXiv:2310.06825, 2023。

M. Joshi、E. Choi、D. Weld 和 L. Zettlemoyer。TriviaQA：用于阅读理解的大规模远程监督挑战数据集。收录于 R. Barzilay 和 M.-Y. Kan 编辑的《计算语言学协会第 55 届年会论文集（第 1 卷：长篇论文）》第 1601–1611 页，加拿大温哥华，2017 年 7 月。计算语言学协会。doi：10.18653/v1/P17-1147。网址https://aclanthology.org/P17-1147。

J. Kaplan、S. McCandlish、T. Henighan、TB Brown、B. Chess、R. Child、S. Gray、A. Radford、 J. Wu 和 D. Amodei。神经语言模型的缩放定律。CoRR， abs/2001.08361，2020 年。网址https://arxiv.org/abs/2001.08361。

VA Korthikanti、J. Casper、S. Lym、L. McAfee、M. Andersch、M. Shoeybi 和 B. Catanzaro。减少大型 Transformer 模型中的激活重新计算。机器学习与系统论文集，第 5 卷，2023 年。

T. Kwiatkowski、J. Palomaki、O. Redfield、M. Collins、AP Parikh、C. Alberti、D. Epstein、I. Polosukhin、J. Devlin、K. Lee、K. Toutanova、L. Jones、M. Kelcey、M. Chang、AM Dai、J. Uszkoreit、Q. Le 和 S. Petrov。《自然问题：问答研究的基准》。《计算机语言学联合期刊》， 7：452–466，2019 年。doi：10.1162/tacl_a_00276。网址https://doi.org/10.1162/tacl_a_00276。

W. Kwon、Z. Li、S. Zhuang、Y. Shen、L. Cheng、CH Yu、JE Gonzalez、H. Zhang 和 I. Stoica。使用 pagedattention 实现大型语言模型的高效内存管理。在2023 年 ACM SIGOPS 第 29 届操作系统原理研讨会论文集上。

G. Lai、Q. Xie、H. Liu、Y. Yang 和 EH Hovy。RACE：来自考试的大规模阅读理解数据集。收录于 M. Palmer、R. Hwa 和 S. Riedel 编辑的《2017 年自然语言处理实证方法会议论文集》（EMNLP 2017），丹麦哥本哈根，2017 年 9 月 9-11 日，第 785-794 页。计算语言学协会，2017 年。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 和 T.Baldwin。 CMMLU：测量中文的大规模多任务语言理解。 arXiv 预印本 arXiv:2306.09212, 2023。

W. Li, F. Qi, M. Sun, X. Yi, 和 J. Zhang. Ccpm：中国古典诗歌匹配数据集，2021 年。

X. 刘、X. 雷、S. 王、Y. 黄、Z. 冯、B. 文、J. 程、P. Ke、Y. 徐、WL Tam、X. 张、 L. 孙、H. 王、J. 张、M. 黄、Y. 东和 J. 唐。Alignbench：大型语言模型的中文对齐基准测试。 CoRR， abs/2311.18743，2023。doi：10.48550/ARXIV.2311.18743。网址https://doi.org/10.48550/arXiv.2311.18743。

I. Loshchilov 和 F. Hutter。解耦权重衰减正则化。arXiv预印本 arXiv:1711.05101，2017年。

Machine Translated by Google

H. Luo、Q. Sun、C. Xu、P. Zhao、J. Lou、C. Tao、X. Geng、Q. Lin、S. Chen 和 D. Zhang。
Wizardmath：通过强化 evol-instruct 为大型语言模型提供数学推理能力。arXiv 预印本 arXiv:2308.09583，2023 年。

S. McCandlish、J. Kaplan、D. Amodei 和 OD 团队。大批量生产的经验模型
训练。arXiv 预印本 arXiv:1812.06162，2018 年。

T. Mihaylov、P. Clark、T. Khot 和 A. Sabharwal。盔甲能导电吗？一种新
开放书籍问答数据集，2018 年。

D. Narayanan、M. Shoeybi、J. Casper、P. LeGresley、M. Patwary、V. Korthikanti、D. Vainbrand、 P. Kashinkunti、J. Bernauer、B. Catanzaro 等人。使用 megatron-lm 在 GPU 集群上进行高效的大规模语言模型训练。在国际高性能计算、网络、存储和分析会议论文集，第 1-15 页，2021 年。

OpenAI。介绍 ChatGPT，2022 年。网址https://openai.com/blog/chatgpt。

OpenAI。GPT4 技术报告。arXiv 预印本 arXiv:2303.08774，2023 年。

L. Ouyang、J. Wu、X. Jiang、D. Almeida、C. Wainwright、P. Mishkin、C. Zhang、S. Agarwal、 K. Slama、A. Ray 等。训练语言模型以遵循人类反馈的指令。
神经信息处理系统进展，35：27730–27744，2022年。

G. Penedo、Q. Malartic、D. Hesslow、R. Cojocaru、A. Cappelli、H. Alobeidli、B. Pannier、E. Al-mazrouei 和 J. Launay。Falcon LLM 的精炼网络数据集：使用网络数据和仅使用网络数据的表现优于精选语料库。arXiv 预印本 arXiv:2306.01116，2023 年。

A. Radford、J. Wu、R. Child、D. Luan、D. Amodei、I. Sutskever 等。语言模型是无监督的多任务学习者。OpenAI 博客，1(8):9，2019 年。

R. Rafailov、A. Sharma、E. Mitchell、S. Ermon、CD Manning 和 C. Finn。直接偏好优化：您的语言模型其实是一个奖励模型。2023 年。

S. Rajbhandari、J. Rasley、O. Ruwase 和 Y. He。Zero：针对训练万亿参数模型的内存优化。在SC20：国际高性能计算、网络、存储和分析会议中，第 1-16 页。IEEE，2020 年。

K. Sakaguchi、RL Bras、C. Bhagavatula 和 Y. Choi。维诺格兰德：敌对的维诺格兰德
大规模模式挑战，2019 年。

CJ Shallue、J. Lee、J. Antognini、J. Sohl-Dickstein、R. Frostig 和 GE Dahl。测量数据并行性对神经网络训练的影响。机器学习研究杂志， 20(112):1–49，2019 年。

N. Shazeer。Glu 变体改进了 transformer。arXiv 预印本 arXiv:2002.05202，2020 年。

M. Shoeybi、M. Patwary、R. Puri、P. LeGresley、J. Casper 和 B. Catanzaro。Megatron-lm：使用模型并行性训练数十亿参数语言模型。arXiv预印本 arXiv:1909.08053，2019 年。

S. Smith、M. Patwary、B. Norick、P. LeGresley、S. Rajbhandari、J. Casper、Z. Liu、S. Prabhumoye、 G. Zerveas、V. Korthikanti 等人。使用 deepspeed 和 megatron 训练大规模生成语言模型 megatron-turing nlg 530b。arXiv 预印本 arXiv:2201.11990，2022 年。

SL Smith、P.-J. Kindermans、C. Ying 和 QV Le。不要降低学习率，而要提高
批量大小。arXiv 预印本 arXiv:1711.00489，2017 年。

Machine Translated by Google

J. Su, M. Ahmed, Y. Lu, S. Pan, W. Bo, 和 Y. Liu. Roformer：增强型旋转变压器
位置嵌入。神经计算，568:127063，2024 年。

K. Sun, D. Yu, D. Yu 和 C. Cardie。调查具有挑战性的中文的先验知识
机器阅读理解，2019年。

M. Suzgun、N. Scales、N. Schärli、S. Gehrmann、Y. Tay、HW Chung、A. Chowdhery、QV Le、 EH Chi、D. Zhou
等人。挑战大型工作台任务以及思路链是否能解决它们。arXiv 预印本 arXiv:2210.09261，2022 年。

H. Touvron、T. Lavril、G. Izacard、X. Martinet、M.-A. Lachaux、T. Lacroix、B. Rozière、N. Goyal、 E. Hambro、F.
Azhar 等人。LLaMA：开放且高效的基础语言模型。arXiv预印本 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.高、V.
Goswami、N. Goyal、A. Hartshorn、S. Hosseini、 R. Hou、H. Inan、M. Kardas、V. Kerkez、M. Khabsa、I. Kloumann、A.
Korenev、PS Koura、 M. Lachaux、T. Lavril、J. Lee、D. Liskovich、Y. Lu、Y. Mao、X. Martinet、T. Mihaylov、P.米什拉， I.
莫利博格， Y. Nie、A. Poulton、J. Reizenstein、R. Rungta、K. Saladi、A. Schelten、R. Silva、EM

Smith、R. Subramanian、XE Tan、B. Tang、R. Taylor、A. Williams、JX Kuan、P. Xu、Z. Yan、 I. Zarov、Y. 张、A. Fan、M. Kambadur、S. Narang、A.
Rodriguez、R. Stojnic、S. Edunov 和T. Scialom。 Llama 2：开放基础和微调的聊天模型。 CoRR， abs/2307.09288，2023b 。 doi：10.48550/
arXiv.2307.09288。网址https://doi.org/10.48550/arXiv.2307。 09288。

A. Vaswani、N. Shazeer、N. Parmar、J. Uszkoreit、L. Jones、AN Gomez、 。凯撒和 I.波罗苏欣。您所需要的就
是关注。神经信息处理系统的进展， 30，
2017年。

Y. Wang、H. Li、X. Han、P. Nakov 和 T. Baldwin。请勿回答：用于评估法学硕士保障措施的数据集。CoRR，
abs/2308.13387，2023 年。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, EH Chi, QV Le 和 D. Zhou。
思维链提示在大型语言模型中引发推理。在NeurIPS， 2022 年。
网址http://papers.nips.cc/paper_files/paper/2022/hash/9d5609613524ecf4f15af0f7b31abca4-Abstract-Conference.html 。

T. Wei, J. Luan, W. Liu, S. Dong 和 B. Wang。Cmath：你的语言模型能否通过中文
小学数学考试？，2023年。

徐 L.、胡 H.、张 X.、李 L.、曹 C.、李 Y.、徐 Y.、孙 K.、余 D.、余 C.、田 Y.、董 Q.、刘 W.、石 B. 、崔 Y.、李 J.、曾 J.、王 R.、谢
W.、李 Y.、Y. Patterson、Z. 田、Y. 张、H. 周、 S. 刘、Z. 赵、Q. 赵、C. 岳、 X. 张、Z. 杨、K. 理查森和 Z. 兰。线索：汉
语理解评估基准。 D. Scott、N. Bel 和 C. Zong，编辑，第 28 届国际计算语言学会议论文集，COLING 2020，西
班牙巴塞罗那（在线），2020 年 12 月 8-13 日，第 4762-4772 页。国际计算语言学委员会，2020 年。doi：
10.18653/V1/2020.COLING-MAIN.419。URL https://doi.org/10.18653/v1/2020.coling-main.419。

A. 杨、B. 肖、B. 王、B. 张、C. 尹、C. 吕、D. 潘、D. 王、D. 严、F. 杨、F. 邓、 F. 王、F. 刘、G. 艾、G. 董、H. 赵、H. 徐、H. 孙、
H. 张、H. 刘、J. 吉、J. 谢、J. 戴、

Machine Translated by Google

方凯、苏丽、宋丽、刘丽、如丽、马丽、王明、刘明、林明、聂娜、郭鹏、孙瑞、张涛、李涛、李涛、程伟、陈伟、曾晓、王晓、陈晓、门晓、于晓、潘晓、沉宇、王宇、李宇、江宇、高宇、Y.张、Z. Zhou、和 Z. Wu。

百川 2:开放大规模语言模型。技术报告,百川公司,2023 年。网址https://cdn.baichuan-ai.com/paper/Baichuan2-technical-report.pdf。

L. Yu、W. Jiang、H. Shi、J. Yu、Z. Liu、Y. Zhang、JT Kwok、Z. Li、A. Weller 和 W. Liu。
Metamath:为大型语言模型引导您自己的数学问题。CoRR， abs/2309.12284,2023 年。doi:10.48550/ARXIV.2309.12284。
URL https://doi.org/10.48550/arXiv.2309.12284 。

R. Zellers、A. Holtzman、Y. Bisk、A. Farhadi 和 Y. Choi。HellaSwag:机器真的能完成你的句子吗?收录于 A. Korhonen、DR Traum 和 L. Màrquez 编辑的《计算语言学协会第 57 届大会论文集》(ACL 2019),意大利佛罗伦萨,2019 年 7 月 28 日至 8 月 2 日,第 1 卷:长篇论文,第4791–4800 页。计算语言学协会,2019 年。doi:10.18653/v1/p19-1472。网址https://doi.org/10.18653/v1/p19-1472。

B. Zhang 和 R. Sennrich。均方根层归一化。神经信息处理系统进展,32,2019 年。

G. Zhang、L. Li、Z. Nado、J. Martens、S. Sachdeva、G. Dahl、C. Shallue 和 RB Grosse。哪些算法选择在哪些批次大小下很重要?来自噪声二次模型的见解。
神经信息处理系统的进展,32,2019。

C. Zheng、M. Huang 和 A. Sun。Chid:用于完形填空测试的大型中文成语数据集。收录于 A. Korhonen、DR Traum 和 L. Màrquez 编辑的《计算语言学协会第 57 届大会论文集》(ACL 2019),意大利佛罗伦萨,2019 年 7 月 28 日至 8 月 2 日,第 1 卷:长篇论文,第 778–787 页。计算语言学协会,2019 年。doi:10.18653/V1/P19-1075。URL https://doi.org/10.18653/v1/p19-1075。

L.郑,W.-L。蒋,Y.盛,S.庄,Z.吴,Y.庄,Z.Lin,Z.Li,D.Li,EP Xing， H.Zhang,JE Gonzalez,和I. Stoica。使用 mt-bench 和聊天机器人来评判 llm-as-a-judge
竞技场。2023年。

W.Zhong、R.Cui、Y.Guo、Y.Liang、S.Lu、Y.Wang、A.Saied、W.Chen 和 N.Duan。 AGIEval:用于评估基础模型的以人为本的基准。 CoRR， abs/2304.06364,2023。doi :10.48550/arXiv.2304.06364。网址https://doi.org/10.48550/arXiv.2304.06364。

J. Zhou, T. Lu, S. Mishra, S. Brahma, S. Basu, Y. Luan, D. Zhou, 和 L. Hou. 指令遵循大型语言模型的评估。arXiv 预印本 arXiv:2311.07911,2023 年。

Machine Translated by Google

A. 附录

A.1. 致谢

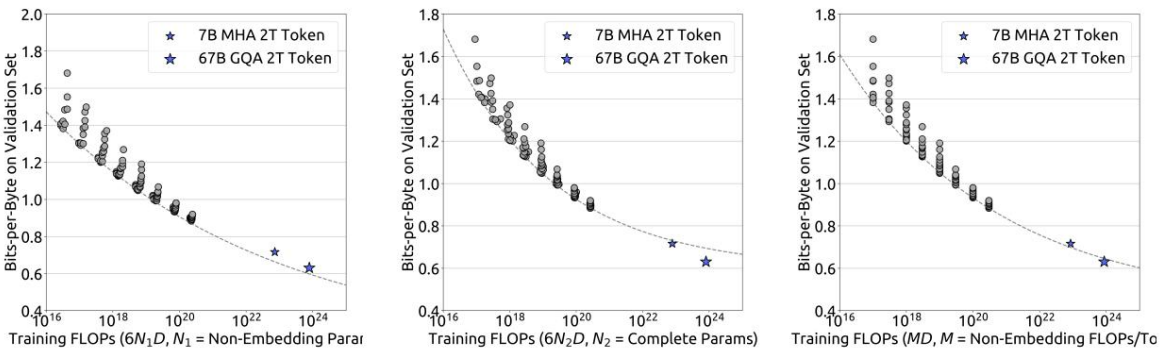
该项目的实现离不开众多贡献者的努力。我们衷心感谢以下个人的帮助1：

·数据标注团队：蔡佳露、陈锐健、陈如意、冯北、黄艳萍、黄震、姜品、金荣利、金香月、柯子云、李慧、李萌、李桑桑、李小千、李耀辉、马云贤、倪佳琪、沉小金、宋新楠、孙天宇、陈小莎、田浩源、王小涵、王小翔、王宇豪、夏凡一、徐雷、徐泽元、徐志鹏、田源、张忠宇、郑毅、周爽、周欣怡、朱雨辰、朱雨轩。 ·合规团队：陈金、唐英、王妙君、王显祖、吴少清、乐毅

夏，WL Xiao。
·业务团队：梁健、李明明、王涛、王显祖、文志牛、叶胜峰、张鹏、张震。 · 设计团队：安伟、查玉坤。

A.2. 不同的模型比例表示

我们重新拟合了不同模型比例表示的缩放曲线，重复使用了IsoFLOP 配置文件中的实验。我们使用 6 1和 6 2作为模型比例表示重新计算了计算 FLOP ，并重新拟合了性能缩放曲线。如图 6 所示，结果表明，在较高的计算预算下，这三种表示之间的最佳模型/数据分配偏差并不大，但在较低的预算下存在明显差异。



（a)计算预算 = 6 1 （b)计算预算 = 6 2 （c)计算预算=

图 6 |使用不同模型比例表示的性能扩展曲线。度量标准是验证集上的每字节位数。虚线表示拟合较小模型（灰色圆圈)的幂律。蓝色星星代表 DeepSeek LLM 7B 和 67B。1、2和分别表示模型的非嵌入参数、完整参数和非嵌入FLOP/token。

当使用 6 1作为模型比例表示时，拟合的性能缩放曲线往往会高估大规模模型的性能。相反，当使用 6 2 时，

1作者按姓氏字母顺序排列。

Machine Translated by Google

曲线往往会低估其表现。使用模型比例表示，
但实现了最准确的预测。

A.3. 基准指标曲线

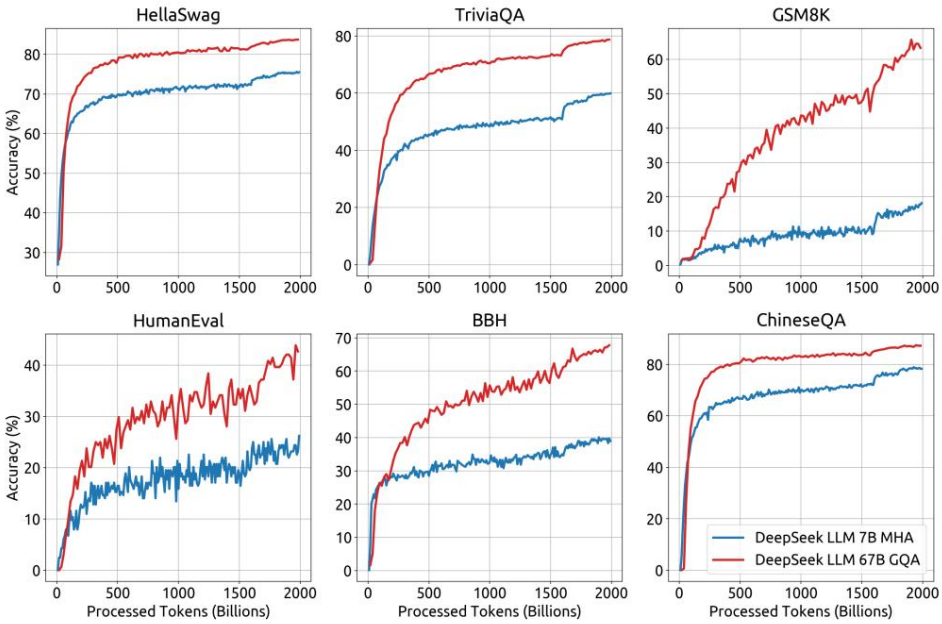


图 7 | DeepSeek LLM Base 的基准指标曲线。ChineseQA 是我们的内部测试集，
以类似于TriviaQA的方式构建。

图 7 显示了不同训练步骤的基准指标曲线。我们可以看到从训练开始到结束，这些基准指标都在持续改善。我们认为
如果训练持续下去，成绩将会进一步提高。

模型	尺寸	人力评估 Python 多语言	马来西亚公共服务局
预训练模型			
Codex-001 33.5% StarCoder 16B 36.0% CodeGeeX2 6B 36.0% CodeLlama 7B 31.7% CodeLlama 13B 36.0% CodeLlama 34B 48.2% DeepSeek-LLM-Base 67B 42.7%		26.1% 45.9% 28.7% 46.8% 24.5% 42.4% 29.2% 41.6% 35.4% 48.4% 41.0 % 55.2% 37.2% 57.4%	
指令调整模型			
Wizard-Coder 34B 73.2% DeepSeek-LLM-Chat 67B 73.8%		48.8% 53.3%	61.2% 61.4%

表 15 | 与特定代码模型的比较。

Machine Translated by Google

A.4. 与代码或数学特定模型的比较

我们对模型和具体代码和数学语言进行了比较
模型（LLM）。表 15 表明 DeepSeek LLM 67B 能够实现类似的
尽管可以访问较少的代码数据，但性能却比 CodeLlama 高。值得注意的是
DeepSeek LLM 在代码以外的领域拥有更强大的能力。

同样，表 16 列出了各种数学相关基准测试的结果，例如
如 GSM8K（Cobbe 等人，2021）、MATH（Hendrycks 等人，2021）、MGSM-zh（i 等人，2023)以及
CMath（Wei 等人，2023 年）。DeepSeek 67B 在数学相关任务上表现出色
在不同语言之间，展示了其在该领域的优势。此外，DeepSeek
LLM 可以利用程序来解决数学问题，这比
思路链。它明显优于之前的 SOTA 模型 ToRA（Gou et al.，
2023)，以基准为依据。

	推理 GSM8K MATH MGSM-zh CMath			
思想链				
MetaMath 70B（Yu 等人，2023）	贸易中心	82.3% 66.4% **26.6%** 64.8% 84.1%		70.9%
WizardMath 70B（Luo 等人，2023 年）	贸易中心	32.6 % 74.0% 22.7%		65.4%
DeepSeek LLM 67B 聊天	贸易中心			80.3%
工具整合推理				
ToRA-Code 34B (Gou et al., 2023) 工具集成 80.7% DeepSeek LLM 67B 聊天工具		50.8%	41.2%	53.4%
集成86.7%		51.1%	76.4%	85.4%

表 16 | 与数学特定模型的比较。

A.5. 带 DPO 阶段的基准测试结果

表 17 列出了使用 DPO 阶段获得的基准测试结果。根据这些结果，
我们可以得出结论，DPO 阶段不会显著影响
法学硕士学位。

	DeepSeek 67B 聊天 DeepSeek 67B 聊天 DPO	
希拉斯瓦格	75.7	76.1
琐事问答	81.5	82.9
自然问题	47.0	48.8
莫尔登大学	71.1	70.9
GSM8K	84.1	85.2
数学	32.6	30.2
人力评估	73.8	71.3
百比黑	71.7	70.8
评估	46.4	46.1
评估	65.2	64.3
加拿大蒙特利尔大学	67.8	68.2

表 17 | DPO 阶段之前和之后的基准指标。

A.6. 评估格式

表 18 至表 40 展示了我们在不同基准上的评估格式的示例。

Machine Translated by Google

问题:下列有关
高尔基体、线粒体和叶绿体的叙述,正确选项:(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:A到D,我们应选择从

表 18 | AGIEval 的一个示例。

提示问题:使用
以下信息回答问题。棉花是一种用于制作织物的植物产品。棉花由纤维素制成,纤维素是一种人类无法消化的纤维。纤维素由许多糖分子组成,这些糖分子结合在一起形成长链。每个糖分子都含有碳、氢和氧原子。洗涤棉织物时,通常会形成皱纹。服装行业使用化学品来制造一些不起皱的棉织物。还添加染料来给棉花中的纤维素纤维染色。服装制造商如何分离颜色以确定染料的纯度?

回答:
选项- 通过过滤
- 根据沸点 - 根据凝固点 - 通
过纸色谱法

表 19 | ARC 的示例。

Machine Translated by Google

提示评估随机

布尔表达式的结果。

Q:不（（不不真实）)是 A:我们一步

一步思考。

请记住:(i) 括号内的表达式始终先求值，(ii) 从最高优先级到最低优先级的运算顺序分别为“非”、“与”、“或”。我们首先将此表达式“Z”简化如下：“Z = not ((not not True)) = not ((A))”,其中“A = not not True”。让我们求值 A:A = not not True = not (not True) = not False = True。代入 A,我们得到:Z = not ((A)) = not ((True)) = not True = False。所以答案是 False。

问:真与假和非真与真分别是答:我们一步一步思考一下。

请记住:(i) 括号内的表达式始终先求值，(ii) 从最高优先级到最低优先级的运算顺序分别为“非”、“与”、“或”。我们首先将此表达式“Z”简化如下：“Z = True 和 False 和非 True 和 True = A 和 B”,其中“A = True 和 False”和“B = 非 True 和 True”。让我们求值 A:A = True 和 False = False。让我们求值 B:B = 非 True 和 True = 非（True 和 True)= 非（True)= False。

代入 A 和 B,我们得到:Z = A 和 B = False 和 False = False。所以答案是 False。

Q:不不（不（假）)是 A:我们一步一

步思考一下。

请记住:(i) 括号内的表达式始终先求值，(ii) 从最高优先级到最低优先级的运算顺序分别为“非”、“与”、“或”。我们首先将此表达式“Z”简化如下：“Z =非非（非（False）)= 非非（A）”,其中“A = 非（False）”。让我们求值A:A = 非（False)= 非 False = True。代入 A,我们得到:Z = 非非（A） = 非非（True）= 非非 False = True。所以答案是 True。

问:假与假与假与否假与否答:我们一步一步思考一下。

表 20 | BBH 的示例。

Machine Translated by Google

PROMPT以下
是关于中国教育学考试的单项题选择,请选出其中的正确答案。

根据我国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

跨学科课程或设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

以下关于大学生的情绪与理智关系的说法是正确的_____。
A.能平静控制自己情绪B.情绪用
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.精细处理策略B.策略
组织C.恢复策略
D.做笔记策略答
案:B

有学者强调,教育要根据一个民族特有的特征来确定,这种观点体现了_____ A.生产力对教育的影响和
否定
B.政治制度对教育的影响和否定
C.文化对教育的影响和否定
D.经济制度对教育的影响和否定答案:

选项
- 答
- B
-C.
-D

表 21 | C-Eval 的示例。

Machine Translated by Google

提示女:这些药 怎么吃?男:一天三次,一次 两片。 请根据里面回答的问题: 他们在哪里?答案:
选项-商店-饭 店-医院- 教室

表 22 | C3 的示例。

提示以下绝句 古诗文翻译成了现代的草莓:已至,万物复苏,春风如美丽的芭蕾睡手巧的姑娘,迈着纤纤细步款而来,她挥舞剪刀,春天地展示那高超的女工技巧,她先裁生长柳叶,随着柳条袅袅依地舞蹈,又裁出杏叶,桃叶。该翻译所对应的古诗文是:
OPTIONS -春 风靡巧如翦刀-剪裁无巧 似春风-风吹怨恨快如刀- 春风欲超越秋风巧

表 23 | CCPM 的示例。

Machine Translated by Google

迅速的

问:某小学在“献爱心 为汶川地震区捐款”活动中六级五个班共 ,
其中一班捐款8000元,其中一班捐款1500元,二班比一班多捐款200元,三班捐款 ,
款1600元,四班与五班融资数之比为3:5.四班融资多少元?
A:一班捐款1500元而二班比一班多捐款200元所以二班捐款 ,
款1500+200=1700元又知道六级五个班一个共捐款8000元所以四个班 ,
和五班融资之和=总融资-一班和二班和三班融资之和,即8000-1500-
1700-1600=3200元,而题目说四班与五班活跃数之比为3:5,则四班活跃数
了3200/(3+5)*3=1200元。所以答案是:1200。

问: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
过了一段时间之后,他位于出发点西边100米处,小俊段跑了多少米?
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
段答案跑的长度-第一段跑的长度=100,第二段跑了100+800=900米。所以
是:900。

Q: A车和B车同时从甲、乙两地相向开出,经过5小时各相遇.然后他们又又
自按原速原方向继续行驶3小时,其次A车离乙地还有135公里,B车离甲地还
有165公里.甲、乙两地相距多少公里?
A:假设A车的速度为x底盘,B车的速度为y底盘,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公里,B车行驶3小时
后距离甲地还有165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),则x+y=150,甲乙两地相距5(x+y)=750千
米。所以答案是:750。

问:在一个底面半径为10厘米的光纤容器内,放入10厘米深的水,然后将一
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?
一个:

表 24 | CMATH 的示例。

Machine Translated by Google

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:胸膜的分部不包括A. 肋胸膜 B.
肺胸膜 C. 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. 突破孔 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: 吞咽的分部
A.吞咽隐窝
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:

选项
- 答
- B
-C.
-D

表 25 | CMMLU 的示例。

Machine Translated by Google

提示段落:该市 的平均年龄为 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。 该城市的性别构成为男性 64.3%,女性 35.7%。 根据以上内容回答下列问题,如需计算,请仔细计算。 问:有多少百分比不是在 25 至 44 之间? 答:答案类型是数字。所以根据上文,答案是83.9。 问:25 到 44 之间的百分比是多少? 答:答案类型是数字。所以根据上文,答案是

表 26 | DROP 的示例。

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。瓦隆此前接受采访时呼吁,美国民众应遵从防疫规定,一线的视线“已连续上班超260天,每天只睡不超过2小时。 ”
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 的示例。

Machine Translated by Google

提示胡雪岩离 船登,坐轿进城,等王有龄到家,他接着也到了他那里,脸上是掩抑止不住的笑容,王有龄夫妇这么觉得奇怪,问他什么事高兴。上面的“他”指的是胡雪岩
渐渐地,汤中凝结成一团团块状,将它们捞起在盆里冷却,肥皂便出现在它们上面。上面的句子中的“”指的是块状形状
“她序上明明引着JulesTellier的比喻,说有个生脱时髦的人去理发,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。这比喻算起来还俏皮。”上面的句子中的“他”指的是生脱时髦的人
在洛伦佐大街的尽头处,矗立着著名的圣三一大教堂。它有巨大的穹顶,还有明亮的彩色玻璃窗,上面画着《旧约》和《新约》的场景。上面句子中的“”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,不安要分,常叫她们学点玩艺儿消遣。
赵雨又提出了一个杯子,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?

表 28 | CLUEWSC 的示例。

Machine Translated by Google

提示问题:Max

可以在 40 分钟内修剪草坪。如果他需要两倍的时间来给草坪施肥,那么他修剪草坪和施肥需要多长时间?

答:我们一步步来思考。给草坪施肥需要 Max 2 * 40 分钟 = 80 分钟。总的来说,Max 需要 80 分钟 + 40 分钟 = 120 分钟来修剪草坪并施肥。答案是 120。

问:百吉饼每个售价 2.25 美元,一打售价 24 美元。如果一次购买一打,每个百吉饼可以节省多少美分?

答:我们一步步来想。它们每个售价 2.25*100=225 美分。按批量价格计算,每个售价为 24/12=2 美元。它们每个售价为 2*100=200 美分。每个百吉饼可节省 225-200=25 美分。答案是 25。

问:蒂姆 5 岁。他的堂兄隆美尔是他的三倍大。他的另一个堂兄珍妮比隆美尔大 2 岁。蒂姆比珍妮小几岁?

答:我们一步一步来思考。Rommel 是 5 x 3 = 15 岁。Jenny 是 15 + 2 = 17 岁。所以,Tim 比 Jenny 小 17 - 5 = 12 岁。答案是 12 岁。

问:学校有 14 名男生和 10 名女生。如果 4 名男生和 3 名女生退学,还剩下多少名男生和多少名女生?

答:我们一步步想想,男生 14 人 - 男生 4 人 = 男生 10 人,女生 10 人- 女生 3 人 = 女生 7 人,一共男生 10 人 + 女生 7 人 = 男生女生各 17 人,答案是 17 人。

问:建造一个鸟舍需要 7 块木板和 20 枚钉子。如果 1 枚钉子的价格为 0.05 美元,而 1 块木板的价格为 3 美元,那么建造 4 个鸟舍的成本是多少美元?

答:我们一步步来思考。一个鸟舍的木板成本为 7 * 3 = 21。每个鸟舍的钉子成本为 20 * 0.05 = 1。因此,建造一个鸟舍需要 21 + 1 = 22。因此,建造 4 个鸟舍的成本为 4 * 22 = 88。答案是 88。

问:丹尼带了 3 个西瓜去参加家庭野餐。他将每个西瓜切成 10 片。他的妹妹带了 1 个西瓜去参加家庭野餐,她将西瓜切成 15 片。野餐时总共有多少片西瓜?

答:我们一步步思考,丹尼送了 3 * 10 = 30 片西瓜,他姐姐送了 1 * 15 = 15 片西瓜,一共有 30 + 15 = 45 片西瓜,答案是 45。

问:Angela 是纽约的一名自行车快递员。她需要运送的包裹数量是餐食数量的 8 倍。如果她需要运送的餐食和包裹加起来为 27 份,那么她需要运送多少份餐食?

A:我们一步步来思考。假设 p 为 Angela 派送的包裹数量,m 为餐食数量。我们知道 p + m = 27 且 p = 8m。将第二个等式代入第一个等式,我们得到 8m + m = 27。合并同类项,我们得到 9m = 27。将两边除以 9,我们得到 m = 3。答案是 3。

问:科莉今天 3 岁。5 年后,她的年龄将是她姑姑的三分之一。她姑姑今天几岁了?

答:我们一步一步来思考。5 年后,科莉将年满 3 + 5 = 8 岁。5 年后,科莉的姑姑将年满 8 x 3 = 24 岁。如今,她的姑姑年满 24 - 5 = 19 岁。答案是 19 岁。

问:Indras 的名字有 6 个字母。她姐姐的名字比Indras 名字的一半字母多 4 个字母。Indras 和她姐姐的名字一共有多少个字母?

A:我们一步步来想想。

表 29 | GSM8K 的示例。

Machine Translated by Google

提示弹钢琴：一 个男人坐在钢琴前。他 选项- 用手和脸 弹奏钢琴。 - 用大拇指在钢琴上弹奏一首提姆巴兰的歌曲。 - 慢慢地弹 奏，然后停下来打响指。 - 在他面前演奏一首歌。

表 30 | HellaSwag 的示例。

迅速的 def starts_one_ends（n）： “”“ 给定一个正整数 n，返回以 1 开头或结尾的 n 位正整数的数量。 “”“

表 31 | HumanEval 的示例。

Machine Translated by Google

提示问题:找到
表达式 $
\frac{\sqrt{x-2}}{\sqrt{5-x}}$ 的定义域。

解决方案:

每个平方根内的表达式必须是非负的。
因此,$x-2 \ge 0$,所以 $x\ge2$,而 $5 - x \ge 0$,所以 $x \le 5$。
而且分母不能等于零,所以 $5-x>0$,从而得出 $x<5$。
因此,表达式的定义域为$\boxed{[2,5)}$。
最终答案:最终答案是 $[2,5)$。我希望它是正确的。

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

解决方案: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。我希望它是正确的。

问题:特雷尔
通常举起两个 20 磅重的重物 12 次。如果他改用两个 15 磅重的重物,那么特雷尔必须举起它们多少次才能举起相同的
总重量?

解决方案:
如果 Terrell 举起两个 20 磅的重物 12 次,他总共举起 $2\cdot 12\cdot20=480$
磅的重物。如果他举起两个 15 磅的重物n 次,他总共将举起 $2\cdot15\cdot
n=30n$ 磅的重物。
将其等同于 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。我希望它是正确的。

问题:
如果方程组

\begin{align*}
6x-4y&=a,\\ 6y-9x
&=b.
\end{align*} 有一个解 (x, y),其中 x 和 y 都非零,求$\frac{a}{b},$假设 b 非零。

解决方案:
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 $-
\frac{2}{3}$。我希望它是正确的。

问题:计算 $\log_21$。

解决方案:

表 32 | MATH 的一个例子。

Machine Translated by Google

提示:你是一位

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查找相似元素。你的代码应该通过以下测试:

断言相似元素((3, 4, 5, 6),(5, 7, 4, 10)) == (4, 5) 断言相似元素((1, 2, 3, 4),(5, 4, 3, 7))
== (3, 4) 断言相似元素((11, 12, 14, 13),(17, 15, 14, 13)) == (13, 14)

[开始] def
similar_elements(test_tup1,test_tup2): res =
tuple(set(test_tup1) & set(test_tup2)) 返回(res)

[完毕]

你是一位专业的 Python 程序员,你的任务是:编写一个 Python函数来识别非素数。你的代码应该通过以下测试:

断言 is_not_prime(2) == False 断言
is_not_prime(10) == True 断言 is_not_prime(35)
== True [BEGIN] 导入数学 def is_not_prime(n):
result =
False for i in
range(2,int(math.sqrt(n)) +
1): if n % i == 0:
result = True 返回结果 [DONE]

你是一位专业的 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查找最大整数。你的代码应该通过以下测试:

断言 heap_queue_largest([25, 35, 22, 85, 14, 65, 75, 22, 58],3)==[85, 75, 65] 断言 heap_queue_largest([25,
35, 22, 85, 14, 65, 75, 22, 58],2)==[85, 75] 断言 heap_queue_largest([25, 35, 22, 85, 14, 65, 75, 22,
58],5)==[85, 75, 65, 58, 35]

[开始]
import heapq as hq def
heap_queue_largest(nums,n):largest_nums =
hq.nlargest(n, nums) returnlargest_nums [完成]

你是一位专业的 Python 程序员,你的任务是:编写一个函数来返回某个数字的所有除数之和。你的代码应该通过以下测试:

断言 sum_div(8)==7 断言
sum_div(12)==16 断言
sum_div(7)==1
[开始]

表 33 | MBPP 的示例。

Machine Translated by Google

迅速的
以下是关于杂项的多项选择题（带答案）
新的。

标准汽车有多少个车轴？
A. 一个
B. 二个
C. 四个
D. 八个 答
案:B

摇滚传奇乐队 Cheap Trick 于 1979 年发行的现场专辑名称中提到了哪个地方？

A. 布达佩斯 B. 武
道馆 C. 不丹 D.
英国 答案:B

有史以来获得 NBA 扣篮大赛冠军的最矮男子是谁？
A. 安东尼·“Spud”·韦伯 B. 迈克尔
·“Air”·乔丹 C. 蒂龙·“Muggsy”·博
格斯 D. 朱利叶斯·“Dr J”·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 的十大热门歌曲？
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

三个臭皮匠中哪一个跟其他人没有血缘关系？
A. Moe B.
Larry C.
Curly D.
Shemp 答案：

选项
- 答
- B
-C.
-D

表 34 | MMLU 的示例。

Machine Translated by Google

提示回答这些

问题:问:谁将于 2022 年举办国际
足联世界杯?
答:卡塔尔

问:谁赢得了第一届女足世界杯?
答:美国 问:《迈阿密风

云》什么时候停止播出?
答:1989

问:谁写了《向主呼喊》这首歌?
答:Darlene Zschech 问:谁

被扔进了狮子窝?
答:丹尼尔 问:

哈比卜这个名字的含义是什么?
一个:

表 35 | NaturalQuestions 的一个例子。

提示一位女士

注意到她每年秋天都会感到沮丧,她想知道为什么。一位朋友向她建议,也许随着季节从温暖变为寒冷而发生的某些变化可能会对她产生影响。当被要求举出这些变化的例子时,这位朋友引用了

选项- 花儿绽

放 - 草儿变黄 - 树木生长 -
花儿绽放

表 36 | OpenBookQA 的一个示例。

迅速的

为了方便按下位于机器下方的垃圾处理机的重置按钮,

选项- 在橱柜地

板上放置一面墙镜 - 在垃圾处理机下放置一面手镜

表 37 | PIQA 的示例。

Machine Translated by Google

提示文章:当你
阅读一篇
文章时,如果你能弄清楚作者是如何把这些想法组合在一起的,你就会更好地理解和记住它。有时,作者通过提出问题然后回答问题来把想法组合在一起。例如,如果这篇文章是关于土拨鼠的,那么作者脑海中的一系列问题可能是:土拨鼠长什么样?

土拨鼠生活在哪里?
他们吃什么? ...
在本文中,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。这些问题会给你信号。它们会告诉你作者接下来要写什么。通常作者脑子里有一个问题,但她不会写出来给你看。你必须自己想出她的问题。这里有一个示例阅读,供你练习这种方法。

蚯蚓你知道蚯蚓
有多少种吗?世界上大约有 1800 种!它们有棕色、紫色、绿色。它们最小的只有 3 厘米长,最大的有3 米长。

观察蚯蚓的最佳时间是晚上,尤其是凉爽潮湿的夜晚。那时它们会从洞穴中爬出来寻找食物。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,它们就无法呼吸。如果下大雨,蚯蚓就必须爬出泥土,因为它们在被水淹没的洞穴里无法呼吸。多么危险的生活啊!

蚯蚓没有眼睛,那么它们怎么知道天黑了呢?它们的皮肤上有对光敏感的特殊部位。这些斑点能分辨出天是亮还是暗。如果你在晚上用手电筒照蚯蚓,它会很快消失在地里。

蚯蚓也没有耳朵,但它们能通过感觉地球的运动来听到声音。如果你想像蚯蚓一样听到声音,就躺在地上,用手指堵住耳朵。然后让一个朋友在你附近跺跺脚。这就是蚯蚓感觉到鸟和人在它们附近行走,以及鼹鼠在挖洞的方式。

蚯蚓很有用。农民和园丁喜欢在土地里养很多蚯蚓,因为蚯蚓挖土有助于改善土壤。挖土使土壤保持疏松透气。一年内,蚯蚓可以在一个足球场大小的面积上堆积多达 23,000 公斤的粪便。

问:读《蚯蚓》的目的是什么?

答:把作者的想法付诸实践。

问:文章中哪个问题无法回答?

A:为什么人类的听力能像蚯蚓一样?

问:根据这篇文章,你如何更好地理解蚯蚓?

答:阅读时要解决作者头脑中的所有问题。

问:这篇文章的最佳标题是什么?

一个:

选项
- 帮助理解的一种方法
- 实践新想法的一种方法
- 成为明智作家的一种方法
- 更清楚地了解蠕虫的一种方法

表 38 | RACE 的一个示例。

Machine Translated by Google

提示回答这些问

题:问:Jayhawker 是指美国某州的反奴
隶制激进团体,他们与密苏里州的奴隶制支持者发生冲突。这个州是哪个州?有时也被称为 Jayhawk 州?

答： 坎斯。

问:哪位瑞典 DJ 和唱片制作人凭借“Wake Me Up”于 2013 年荣登英国冠军单曲榜?

答:蒂姆·伯格林问:谁
是谢菲尔德哈勒姆的议员?
答:尼克·克莱格问：
一个全国瞩目的案件,田纳西州诉约翰·托马斯·斯科普斯案于 1925 年 7 月 21 日结案,陪审团裁定斯科普斯先生犯有教授什么罪行?

答:物种生存问:哪部动画片中
有一个叫小美的角色?
答:Muumi问：“哪

位英国模特,生名 Lesley Hornby,留着短发,中性气质十足， 1966 年被 Nigel Davies 发现,当时她 16 岁,体重6 英石（41 公斤，91 磅）,并凭借Mary Quant 打造的高级时尚造型成为“1966 年的面孔” ？ ”

一个：

表 39 | TriviaQA 的一个示例。

前缀
- 所以莫妮卡
- 所以杰西卡

COMPLETION为了眼睛健

康而避免吃胡萝卜,因为 Emily 需要良好的视力,而 Monica 则不需要。

表 40 | WinoGrande 的示例。请注意,WinoGrande 有多个前缀,但只有一个补全,我们选择补全困惑度最低的预测前缀。