GLM-130B：开放式双语预训练模型

Aohan Zeng◦†*, Xiao Liu◦†*, Zhengxiao Du◦†, Zihan Wang◦, Hanyu Lai◇, Ming Ding◦, Zhuoyi Yang◇, Yifan Xu◦, Wendi Zheng◇、Xiao Xia◦、Weng Lam Tam◇§、Zixuan Ma◦、Yufei Xue§、Jidong Zhai◦、Wenguang Chen◇、Peng Zhang§、Yuxiao Dong◇‡、Jie Tang◦‡。

清华大学◦ Zhipu.AI§

摘要

我们介绍的 GLM-130B 是一个拥有 1300 亿个参数的双语（英语和中文）预训练语言模型。我们试图开源一个至少与 GPT-3 （davinci）一样好的千亿级模型，并揭示如何成功预训练如此大规模的模型。在这一努力过程中，我们面临着许多意想不到的技术和工程挑战，尤其是在损失峰值和发散方面。在本文中，我们将介绍 GLM-130B 的训练过程，包括其设计选择、提高效率和稳定性的 训练策略以及工程方面的努力。由此产生的 GLM-130B 模型在各种流行的英语基准测试中的性能明显优于 GPT-3 175B (davinci)，而在 OPT-175B 和 BLOOM-176B 中则没有表现出性能优势。此外，它的性能还一直明显优于 ERNIE TITAN 3.0 260B--最大的中文模型--跨相关基准。最后，我们利用 GLM-130B 的独特扩展特性，在几乎没有性能损失的情况下，无需后期训练即可实现 INT4 量化，使其成为首个 100B 级模型，更重要的是，它可以在 4×RTX 3090 （24G）或 8×RTX 2080 Ti（11G）GPU（使用 100B 级模型所需的最经济实惠的 GPU）上进行有效推理。GLM-130B 模型的权重可公开访问，其代码、训练日志、相关工具包和经验教训均已开源，网址为 https://github.com/THUDM/GLM-130B/。

1 导言

大型语言模型（LLMs），特别是那些参数超过1000亿（100B）的模型（Brown等人，2020年；Thoppilan等人，2022年；Rae等人，2021年；Chowdhery等人，2022年；Wang等人，2021年），呈现出极具吸引力的缩放规律（Wei等人，2022年b），其中突然出现了零镜头和少镜头能力。其中，具有 175B 参数的 GPT-3 模型（Brown 等人，2020 年）在各种基准测试中，用 32 个标注示例产生的性能明显优于全监督的 BERT-Large 模型，从而推动了对

100B 规模 LLM 的研究。然而，迄今为止，GPT-3（以及许多其他闭源的 100B 级模型）--模型本身--以及如何对其进行训练，对公众来说都是不透明的。在与所有人共享模型和训练过程的情况下，训练如此大规模的高质量 LLM 具有至关重要的价值。

因此，我们的*目标是*在考虑到伦理问题的前提下，*预训练出一个开放且高度精确的 100B 级模型*。在尝试的过程中，我们认识到，与训练 10B 尺度的模型相比，在预训练的效率、稳定性和收敛性方面，预训练如此尺度的高密度 LLM 会带来许多意想不到的技术和工程挑战。在训练 OPT-175B （Zhang 等人，2022 年）和 BLOOM- 176B（Scao 等人，2022 年）时也同时观察到了类似的困难，这进一步证明了 GPT-3 作为一项先驱研究的重要性。

*两位主要作者 AZ 和 XL 的贡献相同 ({zengaohan,shawliu9}@gmail.com)
†AZ、XL 和 ZD 在 Zhipu.AI 实习时完成的部分工作。
‡团队领导：YD 和 JT。通讯作者：JT ()JT (jietang@tsinghua.edu.cn) 作者贡献详见附录 E。

作为会议论文发表于 2023 年国际比较文学
和历史研究国际会议（ICLR 2023）。

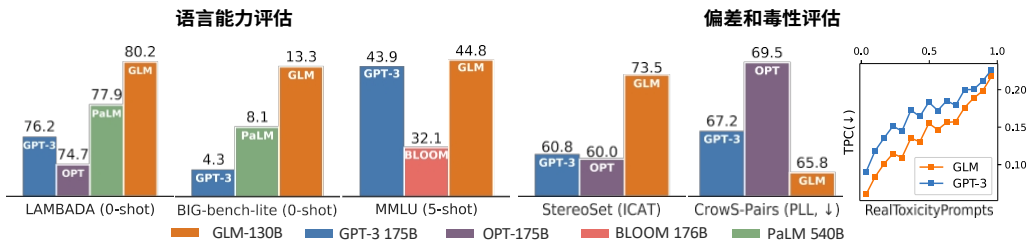


图 1：绩效评估和伦理研究概要。

表 1：GLM-130B 与其他 100B 级 LLM 和 PaLM 540B 的比较。(LN：层规范；FPF：浮点格式；MIP：多任务指令预训练；CN：中文)

模型	开放来源	架构与数据			培训		推理	
		目标	LNMajor		Lang.	FPF	稳定化所需	GPU
GPT-3 175B	×	GPT	LN 预备班	英语	FP16	*未披露*	*未披露*	*未披露*
OPT-175B	✓	GPT	LN 预备班	英语	FP16	手动调节	INT8	8 × 3090
BLOOM-176B	✓	GPT	LN 预备班	多种语言	BF16	嵌入规范	INT8	8 × 3090
PaLM 540B	×	GPT	LN 预备班	英语	BF16	手动调节	*未披露*	*未披露*
GLM-130B	✓	GLM（深度和 MIP）规范		双语（中文和英文）	FP16	嵌入式梯度收缩	INT4	4 × 3090 或 8 × 1080 Ti

在这项工作中，我们从工程设计工作、模型设计选择、提高效率和稳定性的训练策略以及实现可负担推理的量化等方面介绍了 100B 尺度模型--GLM-130B 的预训练。由于人们普遍认识到，通过经验枚举训练 100B 级 LLM 的所有可能设计在计算上是难以承受的，因此我们不仅介绍了训练 GLM-130B 的成功部分，还介绍了许多失败的方案和经验教训。特别是，训练稳定性是成功训练如此大规模模型的决定性因素。与 OPT-175B 中手动调整学习率、BLOOM-176B 中牺牲性能使用嵌入规范等做法不同，我们尝试了多种方案，发现嵌入梯度收缩策略可以显著稳定 GLM-130B 的训练。

具体来说，GLM-130B 是一个具有 130 个双向参数的双语（英语和中文）密集模型，于 2022 年 5 月 6 日至 7 月 3 日期间在由 96 个英伟达 DGX-A100 (8×40G) GPU 节点组成的集群上经过 4000 亿 token 的预训练。我们没有使用 GPT 类型的架构，而是采用了通用语言模型（GLM）算法（Du 等人，2022 年），以充分利用其双向保留优势和自回归空白填充目标。表 1 总结了 GLM-130B、GPT-3 和另外两个开源模型--OPT-175B 和 BLOOM-176B，以及 PaLM 540B（Chowdhery 等，2022 年）--一个更大 4 倍的模型--之间的比较。

总之，GLM-130B 在概念上的独特性和工程设计上的努力使其在大量基准测试（共 112 项任务）中表现出超越 GPT-3 的性能，并在许多情况下优于 PaLM 540B，而在 OPT-175B 和 BLOOM-176B 中未观察到优于 GPT-3 的性能（参见图 1 左侧）。在 LAMBADA（Pa-LM 540B）上，GLM-130B 的零点性能优于 GPT-3 175B（+5.0%）、OPT-175B（+6.5%）和 BLOOM-176B（+13.0%）。

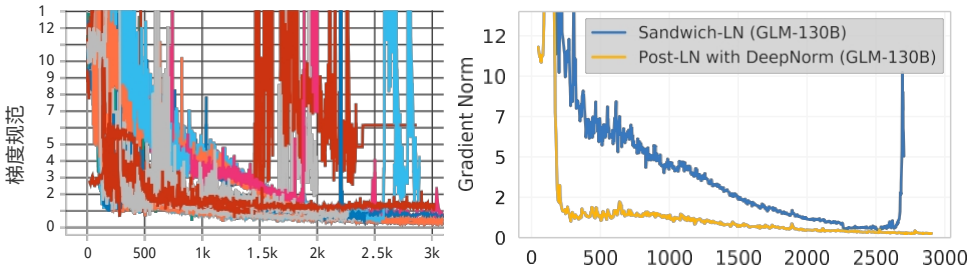
perno等人，2016），在Big-bench-lite（Srivastava等人，2022）上的性能比GPT-3高3倍。在 5 次 MMLU（Hendrycks 等人，2021 年）任务中，它优于 GPT-3 175B（+0.9%）和

和历史研究国际会议（ICLR 2023）。

BLOOM-176B（+12.7%）。作为一种双语 LLM（也是中文），它在 7 个零拍 CLUE（Xu 等人，2020 年）数据集（+24.26%）和 5 个零拍 FewCLUE（Xu 等人，2021 年）数据集（+12.75%）上的结果明显优于 ERNIE TITAN 3.0 260B（Wang 等人，2021 年）--最大的中文 LLM。重要的是，如右图 1 所示，GLM-130B 作为一个开放模型，其*偏差和生成毒性明显低于 100B 尺度的同类*模型。

最后，我们设计 GLM-130B 的目的是让尽可能多的人能够进行 100B 规模的 LLM 研究。首先，我们没有像 OPT 和 BLOOM 那样使用 175B+ 的参数，而是决定使用 130B 的规模，因为这样的规模可以在单台 A100（8×40G）服务器上支持推理。其次，为了进一步降低对 GPU 的要求，我们将 GLM-130B 量化为 INT4 精度，而 OPT 和 BLOOM 只能达到 INT8。由于 GLM 架构的独特属性，GLM-130B 的 INT4 量化带来的性能下降可以忽略不计，例如，在 LAMBADA 上的性能下降为 -0.74%，在 MMLU 上的性能下降甚至为 +0.05%，因此它仍然优于未压缩的 GPT-3。这使得GLM- 130B能够在4×RTX 3090（24G）或8×RTX 2080 Ti（11G）的服务器上实现快速推理，并保证性能，这是*迄今为止使用100B规模LLM所需的最经济实惠的GPU。*

和历史研究国际会议（ICLR 2023）。



(a) 30 多次 100B 级初步试验失败 (b) 最终决定性试验：Sandwich-LN 对 DeepNorm

图 3：GLM-130B 训练中不同 LayerNorms 的试验。结果表明，DeepNorm 是最稳定的一种，因为它的梯度规范较小，在训练初期不会出现尖峰。

我们将模型检查点、代码、培训日志、相关工具包和经验教训开源。

2 GLM-130B 的设计选择

机器学习模型的架构决定了它的归纳偏差。然而，人们已经意识到，探索 LLM 的各种架构设计在计算上是难以承受的。我们介绍并解释了 GLM-130B 的独特设计选择。

2.1 GLM-130B的结构

以 GLM 为骨干。最近大多数百亿规模的 LLM，如 GPT-3、PaLM、OPT 和 BLOOM，都遵循传统的 GPT 式（Radford 等人，2019 年）解码器纯自回归语言建模架构。在 GLM-130B 中，我们尝试探索双整式 GLM--通用语言模型（Du 等人，2022 年）作为其骨干的潜力。

GLM 是一种基于转换器的语言模型，利用自回归空白填充作为训练目标。简言之，对于一个文本序列 $\boldsymbol{x} = [x_1, \ldots, x_n]$，从中采样文本跨度 $\{\boldsymbol{s}_1, \ldots, \boldsymbol{s}_m\}$，其中每个跨度 $\boldsymbol{s}_i$ 表示连续标记符 $[s_{i,1}, \ldots, s_{i,l_i}]$，并用一个掩码标记符替换（即损坏），形成 $\mathrm{xcorrupt}$。要求模型自动恢复它们。为了允许被破坏的跨度之间相互影响，它们之间的可见性是由随机采样的排列顺序决定的。

GLM 对未屏蔽（即未损坏）上下文的双向关注使 GLM-130B 有别于使用单向关注的 GPT 式 LLM。为了支持理解和生成，它混合了两个损坏目标，每个目标都由一个特殊的屏蔽标记来表示：

- **[MASK]**：句子中的短空白，其长度等于输入的某一部分。
- **[gMASK]**：在句末提供随机长度的长空白，并提供前缀上下文。

从概念上讲，空白填充目标与双整时注意力能够比GPT式模型更有效地理解语境：当使用[MASK]时，GLM-130B的表现与BERT（Devlin等人，2019）和T5（Raffel等人，2020）相似；当使用[gMASK]时，GLM-130B的表现与Pre- fixLM相似（Liu等人，2018；Dong等人，2019）。

根据经验，在图 2 中，GLM-130B 超越了 GPT-3 和 PaLM 540B，在零射频 LAMBADA 上实现了 80.2% 的创纪录高准确率。通过设置注意力掩码，GLM-

和历史研究国际会议（ICLR 2023）。

130B 的单向变异性与 GPT-3 和 OPT-175B 相当。我们的研究结果与现有研究结果一致（Liu 等人，2018；Dong 等人，2019）。

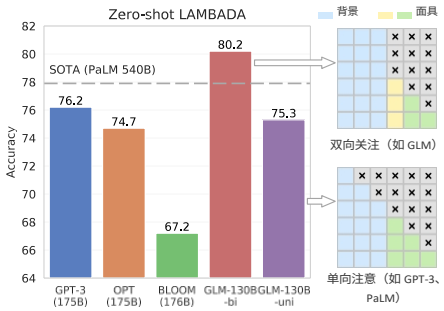


图 2：GLM-130B 和类似规模的 LLM 在零拍 LAMBADA 语言建模上的表现。有关 GLM 双向注意力的详细信息，请参见 Du 等人（2022 年）。

层归一化（LN，Ba 等人（2016 年））。训练的不稳定性是训练 LLMs 的一个主要挑战（Zhang 等人，2022；Scao 等人，2022；Chowdhery 等人，2022）（参见附录图 10 中几个 100B 尺度模型训练时的崩溃情况）。适当选择 LNs 有助于稳定 LLM 的训练。我们尝试了现有的做法，如预 LN（Xiong 等，2020 年）、

不幸的是，Post-LN（Ba 等人，2016 年）和 Sandwich-LN（Ding 等人，2021 年）无法稳定我们的 GLM-130B 测试运行（详见图 3 (a) 和附录 B.2）。

由于 Post-LN 在初步实验中取得了良好的下游结果，尽管它并不能稳定 GLM-130B，但我们后来还是将搜索重点放在了 Post-LN 上。幸运的是，在使用新提出的 DeepNorm（Wang 等，2022b）对 Post-LN 进行初始化的一次尝试中，[1] 产生了很好的训练稳定性。具体来说，考虑到 GLM-130B 的层数 N，我们采用 $\text{DeepNorm}(\boldsymbol{x}) = \text{LayerNorm}(\alpha \cdot \boldsymbol{x} + \text{Network}(\boldsymbol{x}))$，其中 $\alpha = (2N)^{2}$，并对 `ffn`、`v_proj` 和 `out_proj` 应用 Xavier 正则初始化，缩放因子为 $(2N)^{-1}$。此外，所有偏置项都初始化为零。图 3 显示，这大大提高了 GLM-130B 的训练稳定性。

位置编码和 FFN。我们从训练稳定性和下游性能两方面对位置编码（PE）和 FFN 改进的不同方案进行了实证测试（详见附录 B.3）。对于 GLM-130B 中的 PE，我们采用旋转位置编码（RoPE，Su 等人，2021 年），而不是 ALiBi（Press 等人，2021 年）。为了改进 Transformer 中的 FFN，我们选择了带有 GeLU（Hendrycks & Gimpel，2016 年）激活的 GLU 作为替代。

2.2 GLM-130B的训练前设置

受近期工作（Aribandi 等人，2022；Wei 等人，2022a；Sanh 等人，2022）的启发，GLM-130B 的预训练目标不仅包括自监督 GLM 自回归空白填充，还包括一小部分标记的多任务学习。预计这将有助于提高其下游零点性能。

自监督空白填充（95% 标记）。回顾一下，GLM-130B 在这项任务中同时使用了 [MASK] 和 [gMASK]。每个训练序列一次单独使用其中一个。具体来说，[MASK] 用于屏蔽 30% 训练序列中的连续跨度，以进行空白填充。跨度的长度遵循泊松分布（$\lambda = 3$），加起来占输入的 15%。对于其他 70% 的序列，每个序列的前缀被保留为上下文，[gMASK] 用于屏蔽其余部分。屏蔽长度从均匀分布中采样。

预训练数据包括 1.2T Pile（训练拆分）（Gao 等人，2020 年）英语语料库、1.0T 中文五道口语料库（Yuan 等人，2021 年）以及我们从网上抓取的 250G 中文语料库（包括在线论坛、百科全书和问答），这些语料库均衡地包含了英语和中文内容。

多任务指令预训练（MIP，5% 代币）。 T5（Raffel 等人，2020 年）和 ExT5（Aribandi 等人，2022 年）表明，预训练中的多任务学习比微调更有帮助，因此我们建议在 GLM-130B 的预训练中加入语言理解、生成和信息提取等多种教学提示数据集。

最近的研究（Wei 等，2022a；Sanh 等，2022）利用多任务提示微调来改善零点任务转移，与之相比，MIP 只占 5%的标记，并且设置在预训练阶段，以防止破坏 LLMs 的其他一般能力，例如无条件自由生成。具体来说，我们包括了 74 个来自（Sanh 等人，2022；Wang 等人，2022a）的提示数据集，列于附录 C 和表 12。建议 GLM-130B 用户根据第 5 节中说明的标准，避免在这些数据集上评估其零次和少次功能。

和历史研究国际会议（ICLR 2023）。

2.3 平台感知并行策略和模型配置

GLM-130B 在由 96 台 DGX-A100 GPU（8×40G）服务器组成的集群上进行训练，访问时间为 60 天。目标是通过尽可能多的令牌，因为最近的一项研究（Hoffmann 等人，2022 年）表明，大多数现有的 LLM 在很大程度上训练不足。

三维并行策略。数据并行（Valiant，1990）和张量模型并 行（Shoeybi等人，2019）是训练十亿规模模型的实际做法（Wang & Komatsuzaki，2021；Du等人，2022）。为了进一步处理巨大的 GPU 内存需求以及节点间应用张量并行所导致的 GPU 整体利用率的降低（因为训练 GLM-130B 时使用的是 40G 而不是 80G 的 A100），我们将管道模型并行与其他两种策略结合起来，形成了一种三维并行策略。

流水线并行性将模型划分为每个并行组的顺序阶段，为了进一步减少流水线带来的气泡，我们利用 DeepSpeed（Rasley 等人，2020 年）的 PipeDream-Flush （Narayanan 等人，2021 年）实现来训练 GLM-130B，其相对于 GLM-130B 和 GLM-130B 所产生的气泡数量减少了 20%。

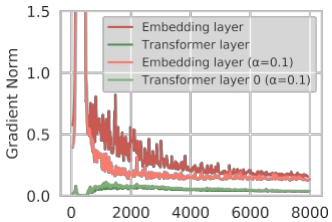
大的全局批量大小（4 224），以减少时间和 GPU 内存的浪费。通过数值和经验检验，我们采用了 4 路张量并行和 8 路流水线并行（详见附录 B.4）。根据（Chowdhery et al., 2022）中的计算，我们报告硬件 FLOPs 利用率（HFU）为 43.3%，模型 FLOPs 利用率（MFU）为 32.5%。

GLM-130B 配置。我们的目标是让我们的 100B 规模 LLM 能够在 FP16 精度下运行单个 DGX-A100 (40G) 节点。根据我们从 GPT-3 中采用的 12,288 个隐藏状态维度，由此产生的模型大小不得超过 130B 个参数，即 GLM-130B。为了最大限度地利用 GPU，我们根据平台及其相应的并行策略来配置模型。为了避免两端额外的字嵌入导致中间阶段内存利用率不足，我们通过去掉一层来平衡流水线分区，使得 GLM-130B 的变压器层数为 9×8-2=70。

在访问集群的 60 天时间里，我们对 GLM-130B 进行了 4000 亿个词块（中文和英文各约 2000 亿个）的训练，每个样本的固定序列长度为 2048 个。对于[gMASK]训练目标，我们使用了一个 2,048 个词组的上下文窗口。对于[MASK]和多任务目标，我们使用 512 的上下文窗口，并将四个样本串联起来，以满足 2,048 个序列长度的要求。在最初的 2.5% 样本中，我们将批量大小从 192 提升到 4224。我们使用 AdamW（Loshchilov & Hutter，2019 年）作为优化器，将 β_1 和 β_2 设置为 0.9
和 0.95，权重衰减值为 0.1。我们将学习率从 10^{-7} 升温到 8×10^{-5}
在前 0.5%的样本上，然后按 10 倍余弦值衰减。我们使用 0.1
并使用 1.0 的剪切值剪切梯度（全部配置见表 11）。

3 GLM-130B 的训练稳定性

训练稳定性是影响 GLM-130B 质量的决定性因素，而训练稳定性在很大程度上也受其通过的标记数量的影响（Hoffmann 等人，2022 年）。因此，考虑到计算量的限制，必须在浮点（FP）格式的效率和稳定性之间做出权衡：低精度 FP 格式（如 16 位精度-FP16）可提高计算效率，但容易出现溢出和下溢错误，导致训练崩溃。



混合精度。我们采用了混合精度（Micikevicius 等人，2018 年）策略（Apex O2）的常见做法，即对前向和后向采用 FP16，对优化器状态和大规模权重采用 FP32，以减少 GPU 内存使用量并提高训练效率。与 OPT-175B 和 BLOOM-176B 类似（参见附录中的图 10），GLM-130B 的训练也面临着因这一选择而导致的频繁损失尖峰，而且随着训练的进行，这种损失尖峰的频率会越来越高。与精度相关的尖峰
这些问题往往没有明确的原因：有些会自行恢复；有些则预示着梯度法线会突然飙升，甚至会出现峰值，甚至损失为零。OPT-175B 试图通过手动跳过数据和调整超参数来修复；BLOOM- 176B 则通过嵌入规范技术来修复（Dettmers 等人，2021 年）。我们花了几个月的时间对尖峰现象进行了实证研究，发现当变压器规模扩大时，会出现一些问题

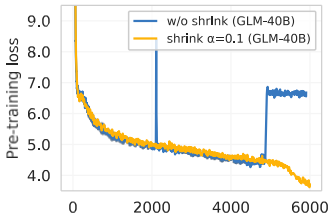
和历史研究国际会议（ICLR 2023）。

：

首先，如果使用前 LN，变压器主分支的数值标度在较深层可能会非常大。在 GLM- 130B 中，通过使用基于 DeepNorm 的 Post-LN（参见第 2.1 节）解决了这一问题，它使值标度始终有界。

其次，关注度分数增长过快，以至于超过了 FP16。

(a) 梯度规范，EGS α = 0.1



(b) 40B 规模测试中的 EGS

图 4：EGS 减少了梯度尺度和方差，以稳定 LLM 的预训练。

随着模型规模的扩大。在 LLM 中，有几种方法可以克服这个问题。在 CogView（Ding 等人，2021 年）中，PB-Relax 被提出来去除偏差项，并在注意力计算中扣除极值以避免该问题，但遗憾的是，这无助于避免 GLM-130B 中的不收敛。在 BLOOM-176B 中，由于 BF16 格式在 NVIDIA Ampere GPU（即 A100）上的取值范围较广，因此采用了 BF16 格式而非 FP16 格式。然而，在我们的实验中，BF16 在运行时比 FP16 多消耗 15% 的 GPU 内存，这是因为它在渐变过程中转换成了 FP32。

更重要的是，其他 GPU 平台（如 NVIDIA Tesla V100）并不支持这种方法，从而限制了所生成 LLM 的可用性。BLOOM-176B 提出的另一个方案是在 BF16 中应用嵌入规范，但会对模型性能造成重大影响，因为他们注意到嵌入规范会损害模型的零点学习（参见 Scao 等人，2022 年）中的第 4.3 节）。

嵌入层梯度收缩（EGS）。我们的实证研究发现，梯度规范可以作为训练崩溃的信息指标。具体来说，我们发现训练崩溃通常会滞后于梯度准则的 "尖峰 "几个训练步骤。这种峰值通常是由嵌入层的异常梯度引起的，因为我们观察到，在 GLM-130B 的早期训练中，它的梯度常态往往比其他层的梯度常态大几个量级（参见图 4 (a)）。此外，在早期训练中，它往往会出现剧烈波动。视觉模型（Chen 等人，2021 年）通过冻结贴片投影层来解决这个问题。遗憾的是，我们无法冻结语言模型中嵌入层的训练。

最后，我们发现嵌入层的梯度收缩可以克服损失尖峰，从而使 GLM-130B 的训练更加稳定。它首次应用于多模态变换器 CogView（Ding 等人，2021 年）。设α为收缩因子，该策略可通过 word_embedding = word_embedding $*$ α +word_embedding.detach() $*$ (1 - α) 轻松实现。图 4 (b) 表明，根据经验，设置 $\alpha = 0.1$ 可以消除我们可能遇到的大多数尖峰，延迟时间可以忽略不计。

事实上，最终的 GLM-130B 训练运行只经历了三次晚期损失发散情况，尽管由于硬件故障而失败了无数次。对于这三个意外的峰值，事实证明进一步缩小嵌入梯度仍然有助于稳定 GLM-130B 的训练。详情请查看我们代码库中的训练笔记和 Tensorboard 日志。

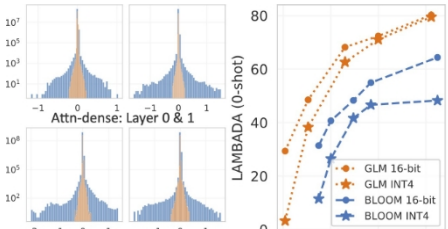
4 RTX 2080 Ti 上的 glm-130b 推断

GLM-130B 的主要目标之一是降低访问 100B 规模 LLM 的硬件要求，而不会降低效率和效果。

如前所述，130B 的模型大小是为在一台普通 A100（40G×8）服务器上运行完整的 GLM-130B 模型而确定的，而不是 OPT-175B 和 BLOOM-176B 所需的高端 A100（80G×8）机器。为了加速 GLM-130B 推断，我们还利用 FasterTransformer（Ti-monin 等人，2022 年）用 C++ 实现了 GLM-130B。与 Huggingface 中 BLOOM-176B 的 PyTorch 实现相比，在同样的单 A100 服务器上，GLM-130B 的解码推断速度快了 7-8.4 倍（参见附录 B）。(详见附录 B.5）。

针对 RTX 3090s/2080s 的 INT4 量化。为了进一步支持流行的 GPU，我们尝试在保持性能优势的同时尽可能压缩 GLM-130B，特别是通过量化（Zafrir 等人，2019；Shen 等人，2020；Tao 等人，2022），这对生成式语言模型的任务区分性能下降很小。

通常的做法是将模型权重和激活度量化为 INT8。然而，我们在附录 B.6 中的分析表明，LLM 的激



活可能包含极端离群值。与此同时，我们还发现了 OPT-175B 和 BLOOM-176B 中出现的异常值（Dettmers 等人，2022 年），这些异常值只影响了约 0.1% 的特征维度，因此可以通过矩阵乘法分解来解决异常维度。不同的是，在 GLM-130B 的激活中存在约 30% 的离群值，这使得上述技术的效率要低得多。因此，我们决定将重点放在模型权重的量化上（即主要是线性层），同时保持 FP16 的精度用于激活。量化模型在运行时动态转换为 FP16 精度，这虽然会产生少量计算开销，但却大大减少了用于存储模型权重的 GPU 内存使用量。

图 5：（左）`attn-dense` 和 `w2` 的权重分布；（右）GLM-130B 的 INT4 权重量化缩放规律。

表 2：左图量化 GLM-130B 在多个基准测试中的性能；右图：INT4 量化 GLM-130B 与 FasterTransformer 的推理速度（编码和解码）。

精密型	GLM-130B			GPT-3
	fp16	int8	int4	fp16
MMLU (acc, 个)	44.75	44.71	44.80	43.9
林巴达（acc，个）	80.21	80.21	79.47	76.2
桩（a 部分，BPB，↓）	0.634	0.638	0.641	0.74

GPU 类型	128 Enc./Dec.	512 Enc./Dec.
8 × A100 (40G)	15.15s	4.29s
8 × V100 (32G)	0.31 秒	6.97 秒
4 × RTX 3090 (24G)	0.37 秒	8.16 秒
8 × RTX 2080 Ti (11G)	0.39 秒	6.77 秒

令人兴奋的是，我们设法实现了 GLM-130B 的 INT4 权重量化，而现有的成功案例迄今为止只达到了 INT8。在内存方面，与 INT8 相比，INT4 版本额外节省了所需 GPU 内存的一半，达到 70GB，从而允许在 4 × RTX 3090 Ti（24G）或 8 × RTX 2080 Ti（11G）上进行 GLM-130B 推断。性能方面，表 2 左侧显示，在完全没有后训练的情况下，INT4 版本的 GLM-130B 几乎没有性能下降，因此在常见基准测试中保持了相对于 GPT-3 的性能优势。

GLM 的 INT4 权重量化缩放规律。我们在图 5 右侧研究了这一独特的 INT4 权重量化缩放规律的内在机制。我们在图 5 左侧绘制了权重值分布图，结果发现它直接影响量化质量。具体来说，分布较广的线性层需要用较大的分区进行量化，从而导致更多的精度损失。因此，分布较广的 `attn-dense` 和 `w2` 矩阵可以解释 GPT 式 BLOOM 的 INT4 量化失败。相反，与类似大小的 GPT 相比，GLM 的分布往往要窄得多，而且随着 GLM 模型规模的扩大，INT4 和 FP16 版本之间的差距会进一步缩小（详情参见附录中的图 15）。

5 结果

我们按照 GPT-3 和 PaLM 等 LLM 的常见设置，对 GLM-130B 的英文[1] 进行了评估。作为一款双语 LLM，GLM-130B 也在中文基准上进行了评估。

关于 GLM-130B 零点学习范围的讨论。由于 GLM-130B 已经使用 MIP 进行了训练，因此我们在此澄清其零点评估的范围。事实上，对于 "零点学习 "的解释似乎存在争议，社会各界尚未达成共识。我们遵从其中一个相关调查（Xian 等人，2018）的观点，即 *"在测试时，在零镜头学习设置中，目的是将测试图像分配给一个未见的类标签"*，其中涉及未见的类标签是一个关键。因此，我们选择 GLM-130B 的零镜头（和少镜头）数据集的标准是：

- **英语**：1) 对于有固定标签的任务（如*自然语言推理*）：不应对此类任务中的数据集进行评估；2) 对于无固定标签的任务（如*（多选）质量保证*、*主题分类*）：只应考虑与 MIP 中的数据集有明显领域转移的数据集。
- **中文**：所有数据集均可进行评估，因为存在零镜头跨语言传输。

过滤测试数据集。按照之前的做法（Brown 等人，2020 年；Rae 等人，2021 年）和上述标准，我们过滤并避免报告可能受到污染的数据集的评估结果。对于 LAMBADA 和 CLUE，我们发现在 13 个词组的设置下重叠现象极少。Pile、MMLU 和 BIG-bench 要么被搁置，要么发布时间晚于语料抓取时间。

5.1 语言建模

LAMBADA.LAMBADA（Paperno 等人，2016 年）是测试最后一个词语言建模能力的数据集。图 2 中显示的结果表明，GLM-130B 的双向关注零点准确率达到了 80.2，创造了 LAMBADA 的新纪录。

PilePile 测试集（Gao 等人，2020 年）包括一系列语言建模基准。平均而言，与 GPT-3 和 Jurassic- 1（Lieber 等人，2021 年）相比，GLM- 130B 在其 18 个共享测试集中的加权 BPB 表现最好，而 GPT-3 和 Jurassic- 1（Lieber 等人，2021 年）的结果被直接采用。

表 3：GLM-130B 在桩基评估中的平均 BPB（18 个子数据集）。

	侏罗纪-1	GPT-3	GLM-130B
平均值BPB	0.650	0.742	**0.634**

后者的语言能力很强（详见附录 C.4）。

[1]报告 OPT-175B 论文的结果，是因为访问该论文的申请数月来一直未获批准。

和历史研究国际会议（ICLR 2023）。

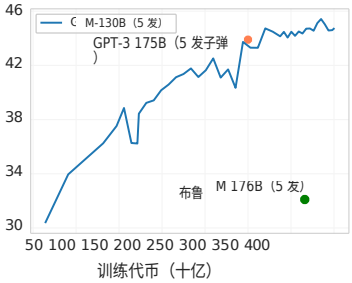


图 6：多功能监测单元上的 GLM-130B

(57 项任务）。

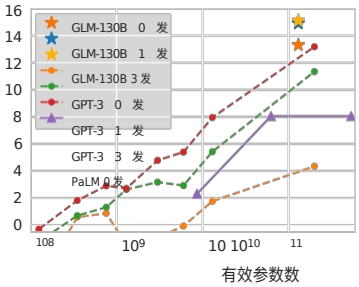


图 7：BIG-bench-lite 评估在不同尺度上的任务（24 项）。

	0 发	1 发	3 发
GPT-3 2.6B	0.60	0.71	1.83
GPT-3 6.7B	-0.06	2.93	5.40
GPT-3 13B	1.77	5.43	7.95
GPT-3 175B	4.35	11.34	13.18
PaLM 540B	8.05	**37.77**	-
GLM-130B	**13.31**	14.91	**15.12**

表 4：BIG-bench-lite 的详细信息轻型工作台（24 项任务）。

5.2 大规模多任务语言理解（MMLU）

MMLU（Hendrycks 等人，2021 年）是一个多样化的基准，包括 57 道多选题，涉及从高中到专家级的人类知识。它是在 Pile 抓取之后发布的，是 LLMs 少量学习的理想测试平台。GPT-3 采用了 MMLU 的结果，BLOOM-176B 则使用与 GLM-130B 相同的提示进行测试（详见附录 C.6 和表 15）。

在图 6 中，GLM-130B 在 MMLU 上的少发（5 发）性能在查看约 300B 标记后接近 GPT-3（43.9）。随着训练的进行，GLM-130B 的准确率继续上升，当训练结束时（即总共查看了 400B 标记），准确率达到 44.8。这与（Hoffmann 等人，2022 年）的观察结果一致，即大多数现有的 LLM 都没有经过充分的训练。

5.3 超越模仿游戏基准（BIG-BENCH）

BIG-bench（Srivastava 等人，2022 年）是关于模型推理、知识和常识能力的挑战性任务基准。鉴于评估其 150 项任务对 LLM 来说非常耗时，我们暂时报告 BIG-bench--一个官方的 24 项任务子集。从图 7 和表 4 中可以看出，GLM-130B 的性能优于 GPT-3 175B，甚至在零点测试环境中也优于 PaLM 540B（大 4 倍）。这可能归功于 GLM-130B 的双向上下文关注和 MIP，这已被证明可以改善未见任务中的零镜头结果（Wei 等人，2022a；Sanh 等人，2022）。随着镜头数量的增加，GLM-130B 的性能也在不断提高，并保持着优于 GPT-3 的性能（参见附录 C.5 和表 14，了解每个模型和任务的详细信息）。

局限与讨论。在上述实验中，我们观察到 GLM-130B 的性能增长（从 13.31 到 15.12）与 GPT-3 的增长（从 4.35 到 13.18）相比并不显著。以下是我们对这一现象的直观理解。

首先，GLM-130B 的双向性可能会导致强大的零次预测性能（正如零次语言建模所显示的），从而比单向 LLM 更接近类似规模（即 100B 规模）模型的少次预测 "上限"。其次，这也可能归因于现有 MIP 范式的不足（Wei 等人，2022a；Sanh 等人，2022），这些范式在训练中只涉及零点预测，很可能会使 GLM-130B 偏向于更强的零点学习，而相对较弱的上下文少点性能。为了纠正这种偏差，我们想出了一个潜在的解决方案，那就是在 MIP 中使用不同的上下文样本镜头，而不仅仅是零镜头样本。

和历史研究国际会议（ICLR 2023）。

最后，尽管 GPT 架构与 GPT-3 几乎相同，但 PaLM 540B 在少镜头上下文学习方面的相对增长要比 GPT-3 显著得多。我们推测，这种性能增长的进一步加速是 PaLM 高质量和多样化的私人收集训练语料的结果。通过将我们的经验与（Hoffmann 等人，2022 年）的见解相结合，我们认识到应该进一步投资更好的架构、更好的数据和更多的训练 FLOPS。

5.4 汉语理解能力评估（CLUE）

我们在已建立的中文 NLP 基准 CLUE（Xu 等人，2020 年）和 FewCLUE（Xu 等人，2021 年）上评估了 GLM-130B 的中文零射性能。到目前为止，我们已经完成了这两个基准的部分测试，包括

和历史研究国际会议（ICLR 2023）。

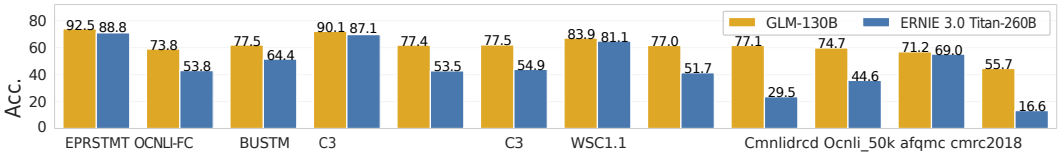


图 8：GLM-130B 和 ERNIE Titan 3.0 260B 在零镜头 CLUE 和 FewCLUE 上的评估结果。

7 个 CLUE 和 5 个 FewCLUE 数据集（详见附录 C.7）。我们将 GLM-130B 与现有最大的中文单语语言模型--260B ERNIE Titan 3.0（Wang 等人，2021 年）进行了比较。我们按照它的设置报告了 dev 数据集上的零次结果。在 12 项任务中，GLM-130B 的表现始终优于 ERNIE Titan 3.0（参见图 8）。有趣的是，在两个抽象 MRC 数据集（DRCD 和 CMRC2018）上，GLM-130B 的表现比 ERNIE 至少好 260%，这可能是由于 GLM-130B 的预训练目标与抽象 MRC 的形式产生了自然共鸣。

6 相关工作

在本节中，我们将回顾与 GLM-130B 有关的预训练、转移和预训练 LLM 的推理等主题的相关工作（Qiu 等人，2020 年；Bommasani 等人，2021 年）。

预训练。香草语言建模指的是纯解码器自回归模型（如 GPT（Radford 等人，2018 年）），但它也可以识别文本上任何形式的自监督目标。最近，基于变换器（Vaswani 等人，2017 年）的语言模型呈现出迷人的缩放规律：随着模型规模的扩大，出现了新的能力（Wei 等人，2022b），从 1.5B （Radford 等人，2019 年）、10B 规模的语言模型（Raffel 等人，2020 年；Shoeybi 等人，2019 年；Black 等人，2022 年）到 100B 规模的 GPT-3 （Brown 等人，2020 年）。后来，尽管出现了许多 100B 尺度的中英文 LLM（Lieber 等人，2021 年；Thop- pilan 等人，2022 年；Rae 等人，2021 年；Smith 等人，2022 年；Chowdhery 等人，2022 年；Wu 等人，2021 年；Zeng 等人，2021 年；Wang 等人，2021 年），但它们并不对公众开放，或只能通过有限的 API 访问。LLMs 的封闭性严重阻碍了其发展。GLM-130B 与最近的 ElutherAI、OPT-175B（Zhang 等人，2022 年）和 BLOOM-176B（Scao 等人，2022 年）一起，旨在为我们的社区提供高质量的开源 LLM。

迁移。尽管微调已成为迁移学习的一种*事实上的*方法，但由于 LLMs 的规模巨大，对它们的评估主要集中在提示和上下文学习上（Brown 等人，2020；刘等人，2021a）。不过，最近也有人尝试对语言模型进行参数高效学习（Houlsby 等人，2019 年）和提示调整（即 P-tuning，李和梁（2021 年）；刘等人（2021b）；Lester 等人（2021 年）；刘等人（2022 年））。目前，我们暂不关注它们，并将在今后的研究中对它们在 GLM-130B 上进行全面测试。

推理。如今，大多数可公开访问的 LLM 都是通过有限的 API 提供服务的。在这项工作中，我们努力的一个重要部分就是 LLM 的高效和快速推断。相关工作可能包括蒸馏（Sanh

和历史研究国际会议（ICLR 2023）。

等人，2019 年；Jiao 等人，2020 年；Wang 等人，2020 年）、量化（Zafrir 等人，2019 年；Shen 等人，2020 年；Tao 等人，2022 年）和剪枝（Michel 等人，2019 年；Fan 等人，2019 年）。最近的研究（Dettmers 等人，2022 年）表明，由于离群维度的特殊分布，OPT-175B 和 BLOOM-176B 等 LLM 可以量化为 8 位。在这项工作中，我们展示了 GLM 对 INT4 权重量化的缩放规律，这使得 GLM-130B 可以在 4×RTX 3090 (24G) GPU 或 8×RTX 2080 Ti (11G) GPU 上进行推理。

7 结论与教训

我们介绍了 GLM-130B，这是一个双语预训练语言模型，旨在促进开放和包容性的 LLM 研究。GLM-130B 的技术和工程工作使我们对 LLM 的架构、预训练目标、训练稳定性和效率以及可负担得起的推断有了深入的了解。总之，GLM-130B 在 112 项任务中的语言表现以及在偏差和毒性基准方面的道德结果，都为 GLM-130B 的高质量做出了贡献。我们的成功和失败经验已浓缩为 100B 级 LLM 培训的教训，见附录 B.10。

和历史研究国际会议（ICLR 2023）。

鸣谢

本研究得到了中国自然科学基金（NSFC）61825602、62276148 和 Zhipu.AI 的支持。我们感谢来自清华大学知识工程组（KEG）、移动、加速和网络系统并行架构与编译器技术组（PACMAN）、自然语言处理组（THUNLP）和知璞人工智能的所有合作者和伙伴。

道德规范声明

我们在此确认，本作品的所有合著者均了解所提供的《ICLR 道德准则》，并遵守该行为准则。本作品介绍了一种开源的大语言模型（LLM），它可用于生成合成文本，用于有害应用，如电话营销诈骗、政治宣传和个人骚扰（Weidinger 等人，2021 年；Sheng 等人，2021 年；Dev 等人，2021 年）。我们预计在使用该模型后，不会产生任何有害输出，尤其是对弱势群体和历史上处于不利地位的群体。

为了更好地与社区合作，从技术上预防并最终消除风险，我们在这项工作中做出了以下重要的公开努力：

用于道德风险研究的开源法律硕士。有些人认为限制法学硕士的获取可以防止此类有害应用，而我们则认为，促进法学硕士的包容性可以更好地防御法学硕士造成的潜在危害。目前，只有政府和大型企业才有能力承担对法律硕士进行预先培训的巨额费用。不能保证拥有雄厚财力的组织不会利用法律硕士造成伤害。如果没有机会接触这些法律硕士，个人甚至无法认识到法律硕士在危害中的作用。

反之，发布开放的 LLM 可以为所有研究人员提供访问权和透明度，并促进减少 LLM 潜在危害的研究，如识别合成文本的算法 Gehrmann 等人（2019）。另外，众所周知，LLMs 可能存在公平性、偏差、隐私和真实性等问题 Zhang 等人（2021）；Lin 等人（2022）；Liang 等人（2021）；Bender 等人（2021）。开放式 LLM 可以揭示与特定输入相对应的模型参数和内部状态，而不是为黑箱模型提供 API。总之，研究人员可以深入分析 LLM 的缺陷，并提出改进算法来解决问题。

伦理评估与改进。我们还在广泛的英语伦理评估基准上评估了我们的模型，包括偏见测量（Nadeem 等人，2021 年；Nangia 等人，2020 年）、仇恨言论检测（Mollas 等人，2020 年）和毒性生成估计（Gehman 等人，2020 年）。尽管存在不足（Blodgett 等人，2021 年；Jacobs & Wallach，2021 年），但这些数据集是迈向开放式定量评估 LLMs 的有意义的第一步。

我们的评估结果表明，我们的算法设计，尤其是对 LLM 的双语预训练，可以显著减轻 LLM 可能出现的偏差和毒性，同时与其他使用单语英语语料库训练的 LLM（Brown 等人，2020 年；Zhang 等人，2022 年）相比，保持其强大的语言性能（详情参见附录 A）。

和历史研究国际会议（ICLR 2023）。

重现性

与 GPT-3 175B（Brown 等人，2020 年）、PaLM 540B（Chowdhery 等人，2022 年）、Gopher（Rae 等人，2021 年）、Chinchilla（Hoffmann 等人，2022 年）、LaMDA（Thoppilan 等人，2022 年）、FLAN（Wei 等人，2022a、GLM-130B采用开放源码，从一开始就致力于促进 LLM 研究的开放性和包容性。

我们为确保评估的可重复性付出了巨大努力。在预训练部分，尽管目前重现所需的成本难以承受，但我们仍然尽最大努力公开 GLM-130B 预训练的代码、细节和整个过程。我们努力让GLM-130B的推理在3090/2080 Ti等少数几种流行的GPU上进行，这也与可重现性承诺相一致，因为它允许大多数学术研究人员在他们的离线机器上重现GLM-130B的结果。我们还提供免费的应用程序接口，供个人用户测试 GLM-130B 的能力。

和历史研究国际会议（ICLR 2023）。

预培训。我们在存储库中提供了完整的训练笔记、Tensorboard 日志和预训练代码（参见摘要）。预训练的超参数和集群配置见第 2.3 节和表 11。多任务指令预训练的训练语料组成和详情见第 2.2 节和附录 C.1 和 C.2。

评估。我们将所有的评估，包括语言基准（LAMBADA、Pile、MMLU、BIG-bench、CLUE 和 FewCLUE）和伦理基准（CrowS-Pairs、StereoSet、ETHOS、RealToxicPrompts），整理成代码库中可单指令运行的 bash 脚本。语言建模基准的数据处理细节见第 5.1 节和附录 C.4，MMLU 的数据处理细节见第 5.2 节和附录 C.6，BIG-bench 的数据处理细节见第 5.3 节和附录 C.5，CLUE 和 FewCLUE 的数据处理细节见第 5 .4 节。所有伦理评估详情请参见附录 A。

参考资料

Oshin Agarwal、Heming Ge、Siamak Shakeri 和 Rami Al-Rfou。基于知识图谱的合成语料库生成，用于知识增强型语言模型预训练。In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguis- tics：人类语言技术*》，第 3554-3565 页，2021 年。

Vamsi Aribandi, Yi Tay, Tal Schuster, Jinfeng Rao, Huaixiu Steven Zheng, Sanket Vaibhav Mehta, Honglei Zhuang, Vinh Q Tran, Dara Bahri, Jianmo Ni, et al. Ext5: Towards extreme multi-task scaling for transfer learning.*国际学习表征大会*，2022 年。

Mikel Artetxe、Shruti Bhosale、Naman Goyal、Todor Mihaylov、Myle Ott、Sam Shleifer、Xi Victo- ria Lin、Jingfei Du、Srinivasan Iyer、Ramakanth Pasunuru 等：《利用专家混合物进行高效大规模语言建模》，*arXiv 预印本 arXiv:2112.10684*, 2021 年。

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. 层归一化》，*arXiv preprint arXiv:1607.06450*, 2016.

Stephen Bach, Victor Sanh, Zheng Xin Yong, Albert Webson, Colin Raffel, Nihal V Nayak, Abheesht Sharma, Taewoon Kim, M Saiful Bari, Thibault Févry, et al. Promptsource：自然语言提示的集成开发环境和资源库。*计算语言学协会第 60 届年会论文集：系统演示*、pp.93-104, 2022.

Emily M. Bender、Timnit Gebru、Angelina McMillan-Major 和 Shmargaret Shmitchell。随机鹦鹉的危险：语言模型会太大吗？In *FAccT '21: 2021 ACM Con- ference on Fairness, Accountability, and Transparency, Virtual Event / Toronto, Canada, March 3-10, 2021*, pp.ACM, 2021.

Jonathan Berant、Andrew Chou、Roy Frostig 和 Percy Liang。从问答对在 freebase 上进行语义解析。*2013 年自然语言处理实证方法会议论文集*》，第 1533-1544 页，2013 年。

Yonatan Bisk, Rowan Zellers, Jianfeng Gao, Yejin Choi, et al. Piqa：用自然语言推理物理常识

和历史研究国际会议（ICLR 2023）。

。In *Proceedings of the AAAI conference on artificial intelligence, volume* 34, pp.

Sidney Black、Stella Biderman、Eric Hallahan、Quentin Anthony、Leo Gao、Laurence Golding、Ho- race He、Connor Leahy、Kyle McDonell、Jason Phang 等。Gpt-neox-20b：开源自回归语言模型。In *Proceedings of BigScience* Episode\# *5-Workshop on Challenges & Perspectives in Creating Large Language Models*, pp.

Su Lin Blodgett、Gilsinia Lopez、Alexandra Olteanu、Robert Sim 和 Hanna Wallach。挪威三文鱼的刻板印象：公平性基准数据集的陷阱盘点。In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp.