

Learning Distribution-Wise Control in Representation Space for Language Models

Chunyuan Deng¹ Ruidi Chang¹ Hanjie Chen¹

英寸

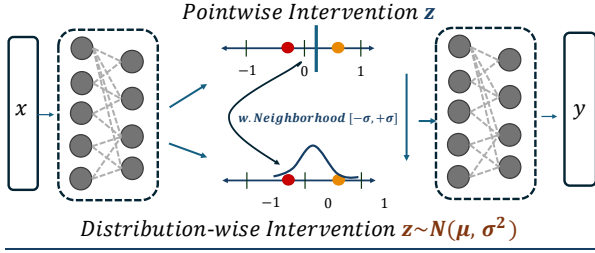


Figure 1. 整体分布干预与点对点干预。这是一种直观而有效的适应，因为先前的研究表明概念空间是连续的 (Gandikota et al., 2023)。

随着语言模型 (LMs) 的复杂性和能力不断增长，理解和控制它们的行为变得愈发重要。最近的可解释性研究进展突出展示了模型干预的潜力——即在前向传递过程中对模型行为进行有针对性的修改——作为实现这种控制的一种有前途的方法。这些干预让研究人员能够引导模型行为，通常是通过操纵模型潜在空间中的表示。

基于干预的方法中的一个核心挑战在于从低级控制推进到高级控制。低级控制包括引导模型输出反义词/同义词或二元标签（例如，积极/消极的情感）。虽然这些任务作为早期干预研究的基础性基准，但该领域需要解决更复杂的高级行为 (Zhang & Nanda, 2024)。这些任务需要在更深层次、更抽象的水平上进行干预，以捕捉模型隐藏表示中的复杂关系和依赖性。

可学习的干预或表示微调 (Wu et al., 2024b; Yin et al., 2024) 已成为一种很有前途的解决方案。它可以实现更强大的高级控制，通常在常识问答 (QA)、数学推理和对齐任务等任务上表现优于参数高效微调 (PEFT) 方法，同时使用的参数更少，约为 10 倍至 100 倍 (Houlsby et al., 2019; Hu et al., 2022; 2023; Liu et al., 2024)。这突显了以更细粒度方式修改模型行为的潜力。

这些方法的一个潜在改进是探索一个理想的“概念空间”。这一空间应是连续的，因为之前的研究表明，

¹ 莱斯大学计算机科学系。Correspondence to: Chunyuan Deng <chunyuan.deng@rice.edu>, Hanjie Chen <hanjie@rice.edu>.

一旦确定了干预向量，其大小可以调整以控制其效果 (Gandikota et al., 2023; Zou et al., 2023; Turner et al., 2023)。这意味着即使学习到了一种干预方法，其邻近区域也应产生相关效果。那么关键的研究问题就出现了——如何有效地探索这一区域。

如图 1 所示，一种直观的方法是将确定性节点替换为随机节点，以直接学习潜在分布。一种实现这目标的常见方法是通过重参数化 (Kingma & Welling, 2014)，这通过将随机性与模型参数解耦，实现了高效的基于梯度的优化。具体来说，随机节点被表示为基础分布和模型参数的可微分函数。这允许标准的反向传播，同时保留从学习到的分布中采样的能力。

基于这些见解，我们提出了一个简单而有效的改进方法，以增强干预方法中概念空间的探索。具体来说，我们用两个独立的网络替换了一个确定性节点（神经网络）。这些网络通过梯度下降独立学习潜在分布的均值 (μ) 和对数方差 ($\log \sigma^2$)。这种方法可以作为现有方法的有效替代。

我们遵循先前工作的实验设置 (Hu et al., 2023; Wu et al., 2024b)，并进行了综合实验，涵盖了八个常识推理基准和七个数学推理基准。我们在 Llama 系列模型 (Touvron et al., 2023b; Dubey et al., 2024) 上测试其分层配置和全层配置下的性能。

在我们的逐层实验中，我们观察到一个有趣的性能提升：在早期层中用随机节点替换确定性节点显著提高了模型性能，获得了从 +4% 到 +6% 的增益。此外，我们发现这些增益与随机节点的学习方差密切相关，这表明训练期间更广泛的邻域探索会导致性能增强。

然后，我们将干预层分析中的见解应用于跨所有层的实验。通过改变根据分布进行干预的层数比例，我们识别了一种有效策略：用随机节点替换前几层，同时在后续层中保持确定性。这种方法在所有 15 个基准测试中实现了一致的性能提升，并且与逐点干预相比，在鲁棒性方面有显著改善。这些发现突出了学习的分布式干预优于其逐点对应方法的优势。

我们的主要贡献是：

- 我们提出了一种简单而有效的干预方法，通过用随机节点替换确定性节点，从而能够更好地探索概念空间。
- 我们证明了这种方法通过干预早期层显著提高了

模型性能。

- 我们发现，一种混合策略——仅替换前几层为随机节点，同时保持后面的层为确定性节点——在性能和鲁棒性方面均产生了最佳结果。

重参数化技巧是一种被广泛使用的技术，它使神经网络能够通过梯度下降从采样中学习。通常应用于变分自编码器（VAE）和变分信息瓶颈（VIB），以帮助模型学习潜在分布。VIB 已被证明能够有效控制语言模型的词嵌入层或下游任务中的表现。VAE/VIB 中的术语“变分”是指变分推断，其中使用 KL 散度来测量近似与真实分布的接近程度。然而，在我们的设定中，我们去掉了 KL 损失项，使其不再是变分推断。相反，它变成了一种放松的方式，允许语言模型在没有约束的情况下学习分布。

最近，表示微调作为一种方法出现，用于提供对语言模型（LMs）行为的高级控制。两项并行的研究工作，ReFT (Wu et al., 2024b) 和 LoFiT (Yin et al., 2024)，从不同的角度解决了这个问题。ReFT 建立在分布式对齐搜索（DAS）(Geiger et al., 2021; 2023) 理论之上，证明了在潜在子空间中正交属性对于在 transformer 块之间实现任务无关的控制至关重要。而 LoFiT 则从传统干预研究中汲取灵感，专注于本地化和编辑的范式 (Meng et al., 2022; Stolfo et al., 2023)。这两种方法都有一个共同的目标：将可解释性研究引入到高级行为控制领域，与参数高效微调（PEFT）方法如适配器 (Houlsby et al., 2019; Pfeiffer et al., 2020; Fu et al., 2021; Hu et al., 2023; Zhang et al., 2023) 和 LoRA (Hu et al., 2022; Liu et al., 2024; Zhang et al., 2024) 相媲美。他们的结果表明，可学习的干预可以匹配甚至超过 PEFT 方法，通常只需显著更少的资源。我们的方法更接近于 ReFT，因为它涉及到学习编码器和解码器之间（即 transformer 块之间）的潜在分布。因此，我们将主要在 ReFT 框架下评估我们随机与确定性节点方法。

基于干预的可解释性。 基于干预的可解释性侧重于操纵模型的内部状态，以理解语言模型如何表现出各种行为 (Subramani et al., 2022; Zou et al., 2023; Turner et al., 2023; Li et al., 2023a)。通过干预潜在表示的特定线性子空间，研究人员揭示了人类可解释的概念 (Rumelhart et al., 1986)，如语言特征（例如，性别、数量）(Hewitt & Manning, 2019; Lasri et al., 2022; Wang et al., 2023; Hanna et al., 2023; Arora et al., 2024; Huang et al., 2024) 和逻辑推理，通常在线性编码在这些模型中 (Wu et al., 2023; Deng et al., 2024; Gur-Arieh et al., 2025)。概念擦除和子空间干预等技术在解开这些属性上发挥了重要作用 (Belrose et al., 2023; Ravfogel et al., 2022)，使得目标修改能够改善模型的公平性、可解释性和任务性能 (Nanda et al., 2023; Park et al., 2024)。这些研究展示了语言模型的表示空间编码了与任务高度相关的丰富、结构化信息，支持更有效和针对性的干预。

1. 预备知识

我们现在首先介绍我们干预方法的背景。¹ 首先，我们概述基于变压器的解码语言模型及其分层隐藏表示的公式。然后，我们将从信息论的角度提供一个统一的干预视角。

基于 Transformer 的自回归语言模型 (Vaswani et al., 2017) 旨在预测一个符号序列的概率。设 $X = \{x_1, x_2, \dots, x_n\}$ 表示输入序列，其中每个 x_i 表示序列中的一个符号。设 $Y = \{y_1, y_2, \dots, y_m\}$ 表示输出序列。总体而言，语言建模中下一个符号的预测目标可以形式化表示为估计 $P(Y|X)$ 。

模型每一层的隐藏表示充当潜在变量 Z ，编码中间抽象以桥接 X 和 Y ：

$$Z^{(l)} = \{z_1^{(l)}, z_2^{(l)}, \dots, z_n^{(l)}\}, \quad l = 1, 2, \dots, L,$$

，其中 l 索引语言模型的层， $z_i^{(l)}$ 表示层 l 的令牌 x_i 的隐藏状态。

1.1. 逐层表示变换

语言模型中的每一层旨在使用来自邻近标记的上下文信息来转换潜在表示 $Z^{(l)}$ 。这些转换可以表示为：

$$Z^{(l+1)} = \text{Attn}(Z^{(l)}) + \text{FFN}(\text{Attn}(Z^{(l)})),$$

其中 FFN 表示一个从 $Z^{(l)}$ 输入的前馈网络 (FFN)，而 $\text{Attn}(\cdot)$ 表示 transformer 块内的自注意力模块。

在这项工作中，我们提供了关于模型干预的信息理论视角。在成熟的信息理论研究中，估计互信息的常用方法是在特定层插入辅助网络（例如，变分自编码器）。有趣的是，在干预研究中，语言模型中的干预被形式化为一个函数 f_ϕ ，由 ϕ 参数化，它将特定层 l 的隐藏表示 $Z^{(l)}$ 转换为修改后的表示 $\hat{Z}^{(l)}$ ：

$$\hat{Z}^{(l)} = f_\phi(Z^{(l)}). \quad (1)$$

我们在此观察到一个联系：信息论中的互信息估计所使用的辅助网络可以被视为一种特定形式的干预。这代表了一种特殊情况，其中变换函数 f_ϕ 是一个可学习的变分自编码器。

在实际中，介入的目标可以视为提高下游任务的表现。对于可学习的介入，这转化为最小化预测输出和真实输出之间的交叉熵（CE）损失：

$$\mathcal{L}_{CE} = -\mathbb{E}_{(X,Y)} \left[\log P(Y|f_\phi(Z^{(l)})) \right]. \quad (2)$$

交叉熵损失可以直接被识别为条件熵：

$$\mathcal{L}_{CE} = -\mathbb{E}_{(X,Y)} \left[\log P(Y|f_\phi(Z^{(l)})) \right] = H(Y|f_\phi(Z^{(l)})). \quad (3)$$

¹在这项工作中，我们主要关注层（即 transformer 块）之间的可学习干预。

Y 和变换后的表示之间的互信息定义为：

$$I(Y; f_\phi(Z^{(l)})) = H(Y) - H(Y|f_\phi(Z^{(l)})). \quad (4)$$

重新排列这个表达式：

$$H(Y|f_\phi(Z^{(l)})) = H(Y) - I(Y; f_\phi(Z^{(l)})). \quad (5)$$

将其代入交叉熵损失：

$$\mathcal{L}_{CE} = H(Y|f_\phi(Z^{(l)})) = H(Y) - I(Y; f_\phi(Z^{(l)})). \quad (6)$$

鉴于 $H(Y)$ 对于介入参数 ϕ 来说是常数，最小化交叉熵损失等价于最大化互信息：

$$\arg \min_{\phi} \mathcal{L}_{CE} \equiv \arg \max_{\phi} I(Y; f_\phi(Z^{(l)})). \quad (7)$$

该公式抓住了介入的根本目标：转换内部表示，使它们对目标输出具有最大的信息性。概念上，这种优化寻求能够最好地保持和放大与任务相关信号同时可能过滤掉无关信息的介入。

在这一部分中，我们首先介绍分布式控制的动机，并详细描述改进之处，特别是将确定性节点替换为能够通过采样学习的随机节点。

1.2. 动机

许多先前的研究发现，通过调整干预的幅度可以控制其效果 (Gandikota et al., 2023; Turner et al., 2023; Han et al., 2023)。干预效果不应限于单一点；相反，其周围邻域也必须展示相关效果。这表明干预的影响会传播到相关区域。一个有用的类比是从自动编码器 (AE) 到变分自动编码器 (VAE) 的过渡。VAE 用随机采样替换确定性节点，使模型可以直接学习潜在分布 (Kingma & Welling, 2014)。我们探索将该技术应用于干预研究，调查其是否可以帮助学习更好的干预措施。

1.3. 随机干预重参数化

为了通过随机节点有效地学习分布，我们采用重新参数化技巧 (Kingma & Welling, 2014)。该技术通过将随机采样过程重新表述为分布参数和辅助噪声变量的确定性函数，使基于梯度的优化能够通过采样实现。

考虑一个简单的确定性 MLP 层，它通过以下方式转换输入表示 Z ：

$$\hat{Z} = \text{MLP}(Z) = W^T Z + b. \quad (8)$$

我们将此替换为学习分布 $\mathcal{N}(\mu, \sigma^2)$ 的随机层。我们不直接从该分布中采样，因为那样会破坏梯度流，我们重新参数化采样过程：

$$\mu = \text{MLP}_\mu(Z), \quad (9)$$

$$\log \sigma^2 = \text{MLP}_{\log \sigma^2}(Z), \quad (10)$$

$$\sigma = \exp\left(\frac{1}{2} \log \sigma^2\right), \quad (11)$$

$$\epsilon \sim \mathcal{N}(0, I), \quad (12)$$

$$\hat{Z} = \mu + \sigma \odot \epsilon. \quad (13)$$

在这里， MLP_μ 和 $\text{MLP}_{\log \sigma^2}$ 学习分布参数，而 $\odot$ 表示逐元素相乘。随机性来自 ϵ （随机噪声），它允许梯度在反向传播过程中通过 μ 和 σ ，同时保持通过 ϵ 转换的随机性质。

1.4. 训练目标

给定一个被冻结的基础语言模型 $\mathcal{M}$ ，并在变压器块之间插入了可训练的随机干预层，我们将下一个令牌预测任务的交叉熵损失最小化：

$$\mathcal{L} = -\mathbb{E}_{(X,Y)} \left[\log P_{\mathcal{M} \circ \mathcal{I}}(Y|f_\phi(Z^{(l)})) \right],$$

其中 $\mathcal{M} \circ \mathcal{I}$ 表示由冻结语言模型和我们的随机干预层组成的系统。在训练过程中，梯度通过重新参数化的随机网络流回到 MLP_μ 和 MLP_σ 网络的可学习参数 $\{\phi_\mu^{(l)}, \phi_\sigma^{(l)}\}_{l=1}^L$ 。

为了处理因大抽样方差而产生的数值不稳定问题，我们引入基于目标语言模型权重分布的模型特定夹紧。给定具有第 l 层干预的语言模型 $\mathcal{M}$ ，我们使用邻近层权重的统计数据定义夹紧边界。令 $W^{(l)}$ 和 $W^{(l+1)}$ 分别表示干预层之前和之后的权重矩阵。我们定义夹紧边界为：

$$v_{\min} = \min(\min(W^{(l)}), \min(W^{(l+1)})), \quad (14)$$

$$v_{\max} = \max(\max(W^{(l)}), \max(W^{(l+1)})). \quad (15)$$

。这个模型特定夹紧确保了干预保持在模型权重分布的自然范围内，有助于在维护模型学习表示的同时保持稳定。边界在训练之前计算一次，并在整个干预过程中保持固定。

为了评估我们的分布级干预方法与逐点干预相比的效果，我们在十多个数据集上评估了我们的方法，并结合了不同超参数调优的全部组合。通常情况下，我们遵循之前 SOTA 方法的标准设置，如 ReFT (Wu et al., 2024b)，我们的代码库是基于 pyenve (Wu et al., 2024c) 构建的。ReFT 的评估框架也源自于之前的工作，如 (Hu et al., 2023; Liu et al., 2024; Wu et al., 2024a)。与这些之前的工作类似，我们评估了 Llama 系列模型 (Touvron et al., 2023a;b; Dubey et al., 2024)，范围从 Llama-7B/13B 到 Llama-3-8B。我们所有的实验都是在启用混合精度 (bfloat16) 的单个 NVIDIA RTX A6000 GPU 上进行的。

我们的评估分为两个部分：(i) 逐层设置和 (ii) 全局设置。首先，我们通过实验不同类型的干预来分析逐层控制。然后，我们探讨将这些干预替换为分布级控制如何影响性能。最后，我们在全局设置中评估这些干预，并与以前文献中的结果进行比较。

1.5. 基线

对于逐层设置，除了 RED (Wu et al., 2024a) 和 ReFT，我们还包括简单的 MLP 和 SwiGLU (Shazeer, 2020)

作为基线，以评估分布级干预的影响。² 具体格式如下所示。

对于所有层设置，我们对 ReFT 和以前的参数高效微调 (PEFT) 方法进行了比较分析。这些方法包括：Prefix-tuning (Li & Liang, 2021)，RED (Wu et al., 2024a)，LoRA (Hu et al., 2022)，DoRA (Liu et al., 2024) 和 ReFT (Wu et al., 2024b)。

Intervention Functions

Pointwise intervention function f_ϕ :

- MLP: $\hat{Z} = W^T Z + b$
- 红色: $\hat{Z} = W \odot Z + b$
- SwiGLU: $\hat{Z} = (W \odot Z + b) \odot GELU(Z)$
- ReFT: $\hat{Z} = Z + R(W^T Z + b - R^T Z)$

Distribution-wise intervention function f'_ϕ :

- D-MLP: $\hat{Z} = \mu + \sigma \odot \epsilon$
- D-RED: $\hat{Z} = \mu + \sigma \odot \epsilon$
- D-SwiGLU: $\hat{Z} = \mu \odot GELU(Z) + \sigma \odot \epsilon$
- D-ReFT: $\hat{Z} = Z + R(\mu + \sigma \odot \epsilon - R^T Z)$

1.6. 基准测试

我们在七个常识推理基准和七个算术推理基准上评估我们的方法。

对于常识推理，我们有 BoolQ (Clark et al., 2019)、PIQA (Bisk et al., 2020)、SIQA (Sap et al., 2019)、HellaSwag (Zellers et al., 2019)、WinoGrande (Sakaguchi et al., 2020)、ARC-e、ARC-c (Clark et al., 2018) 和 OBQA (Mihaylov et al., 2018)。输入格式是多项选择问答，给定一个背景或多个答案选择的问题。输出则是选择的答案，没有 CoT 推理。

对于算术推理，我们有 AddSub (Hosseini et al., 2014)、SingleEQ (Koncel-Kedziorski et al., 2015)、MultiArith (Roy & Roth, 2015)、AQuA (Ling et al., 2017)、GSM8K (Cobbe et al., 2021)、MAWPS (Koncel-Kedziorski et al., 2016) 和 SVAMP (Patel et al., 2021)。在算术推理基准测试中，CoT 推理会在最终答案之前给出。

对于所有基准测试，我们使用相同的提示模板作为 Hu et al. (2023); Wu et al. (2024b) 中的模板。我们还去除了数据集中的首尾空白。

对于常识推理基准测试，我们使用 Commonsense170K 数据集训练模型。对于算术推理基准测试，我们使用 Math10K 数据集。这些数据集是从原始基准测试中组合而成的训练集。我们使用 GSM8K 训练集的一部分作为开发集，用于调整最佳超参数，并应用这组超参

²我们将 ReFT 的分布级变体表示为 D-ReFT，类似的符号也适用于其他方法

数来报告测试分数。我们不在测试集上直接进行优化。此设置与 Wu et al. (2024b) 所使用的设置相同。

关键参数包括干预层 (l)、噪声尺度 (ϵ)、子空间秩 (r)、干预位置 (p)、批大小 (bs)、训练轮数 (e) 和学习率 (lr)。这些参数在开发集上进行了调整，但正文中未包含消融研究。详细数值在附录 7 中提供。

2. 逐层干预

以往的研究经常使用全层设置 (即，干预应用于所有层) 报告结果。在这项研究中，我们首先进行逐层消融实验，以确定性能提升来自哪里。

2.1. 消融研究：D-干预层 l

我们在七个基准上评估算术推理中的分布级可控性：AddSub、SingleEQ、MultiArith、AQuA、GSM8K、MAWPS 和 SVAMP。所有方法均在开发集上进行调整，并报告在使用不同种子运行三次后的平均结果。

如图 2 所示，在所有四种方法中，ReFT 相较于 RED、SwiGLU 和 MLP 保持了卓越的性能。无论是标准干预还是其分布级变体，都揭示了一个一致的模式：更深的层会导致显著的性能下降。当干预移动到后期层时，准确性从 0.7 下降到约 0.4，这表明后期层的干预比早期层的干预更具挑战性。

这来源于数据处理不等式 (见附录 ??)，该不等式指出深层无法恢复在较早层中丢失的信息。因此，应尽早进行干预以在变换降低信息之前保留有用信息。在 MLP、SwiGLU、RED 和 ReFT 中，他们的 D-变体在早期层中始终将准确率提高约 +4%，这比 LoRA $\rightarrow$ DoRA $\rightarrow$ ReFT (Liu et al., 2024; Wu et al., 2024b) 的改进更高。这突显了在网络早期应用的分布级干预是一个有效的改进。

我们随后对随机节点 ϵ 进行了一项消融研究。在我们的方法中， $\epsilon \odot \sigma$ 在将确定性节点转变为随机节点以学习分布方面起到了关键作用。这里， σ 代表潜在变量 Z 的标准差矩阵，而 ϵ 控制这种学习效果的强度。调整 ϵ 使我们能够探索真实分布或理想概念空间。

我们在 D-ReFT 设置中，使用步长为 0.2，将应用于 ϵ 的缩放因子 λ 从 0 变化到 3.0。如图 3 所示，性能从 ReFT 到 D-ReFT 随着 (缩放因子为 $\epsilon = 1$) 取得最佳结果而改善。这表明默认设置带有 $\epsilon \sim \mathcal{N}(0, I)$ 仍然具有稳定的性能。然而，进一步增加 ϵ 引入了更高的方差，使训练更加困难并导致数值不稳定。

我们研究了准确度与学习分布的标准差之间的相关性。使用所有测试集样本，我们计算出不同标准差下的平均准确度。如图 ?? 所示，准确度和标准差表现出正相关关系：标准差较高的分布往往表现更好，而标准差较低则与较差的表现相关。这表明，分布级别的干预效果显著，因为它们探索了逐点干预的邻域区域。

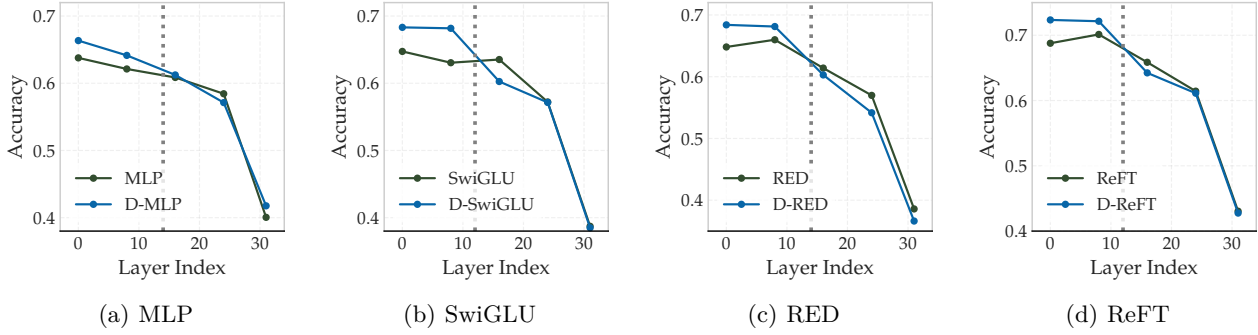


Figure 2. 不同层次的 D 干预在干预层方面的表现。我们报告了 Llama-3-8B 在七个算术基准上的平均得分：AddSub、SingleEQ、MultiArith、AQuA、GSM8K、MAWPS 和 SVAMP。值得注意的是，早期层的 D 干预表现最佳，突显了各层之间显著的差异。

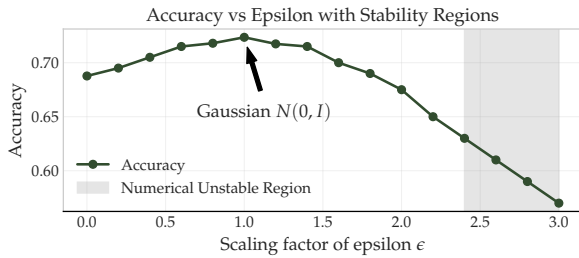


Figure 3. 对于 D-ReFT 中 ϵ 使用不同缩放因子的准确性。当 $\epsilon = 0$ 的缩放因子时，该方法简化为原始的 ReFT。而对于 $\epsilon > 2.4$ 的缩放因子，D-ReFT 进入一个由于方差过大而容易出现数值不稳定的区域。

3. 全局干预

我们评估了 D-ReFT 在所有网络层次上的表现，重点关注层分配如何影响结果。由于在早期层次上改进最为显著，我们进行了一个消融研究，以测试逐点干预和分布级干预的混合策略。具体而言，我们改变了用 D-ReFT 替换多少个早期层次。我们比较了四种配置：D-ReFT 25%，D-ReFT 50%，D-ReFT 75%，和 D-ReFT 100%，其中下标表示用 D-ReFT 替换的早期层的百分比。表 1 显示了在八个不同常识推理基准上的性能比较。结果表明，D-ReFT 25% 一贯优于其原始逐点版本 ReFT 和现有的 PEFT 方法，适用于所有 LLaMA 模型。这些结果强调了 D-ReFT 在提高推理性能同时保持效率方面的有效性。表 ?? 显示了 ReFT 和 D-ReFT 混合策略的结果。我们观察到一个一致的趋势：将 D-ReFT 应用于前 25% 的层，并将 ReFT 应用于剩余层会得到最佳性能，而在训练期间将 D-ReFT 应用于某一阈值以上（即引入过多随机性）会导致模型难以收敛。这表明一种混合策略——在早期层中使用随机节点，在后期层中使用确定性节点——对于语言模型是最优的。

这个现象可以理解成语言模型的早期层具有输入 X 的丰富相关信息。通过对这些层施加分布级干预，模型

保留了对不确定性的更丰富、更灵活的表示，避免过早地过于专注于具体特征。这种随机性允许模型向下传播多样的假设，后来的层——专注于高级推理和特定任务逻辑——可以使用确定性的、逐点的干预来加以优化。

3.1. 算术推理

我们基于 ReFT 对现有 PEFT 和介入方法（LoRA、RED 和 ReFT）在算术推理数据集上的表现进行评价，使用 LLaMA-1 7B、LLaMA-2 7B 和 LLaMA-3 8B（参见表 ??）。我们的方法在保持参数效率的同时，始终优于基线。值得注意的是，D-ReFT 25% 在所有模型尺寸中取得了最高的平均准确率，并在 GSM8K 和 Singleequation 等关键任务中表现出色。这些结果证明了其在以最少参数开销提高算术推理方面的有效性，使其成为现有 ReFT 方法的有前途的轻量替代方案。

3.2. 参数 (%) 在干预中的影响

一个潜在的问题是，D-干预会引入额外的参数用于方差计算，可能解释了性能的提升。我们在子空间秩 (r) 上进行了一项消融研究，以控制实验中使用的参数数量。我们将秩设置为 8、16、32、64 和 128，每个设置中参数数量增加 $2 \times$ 。

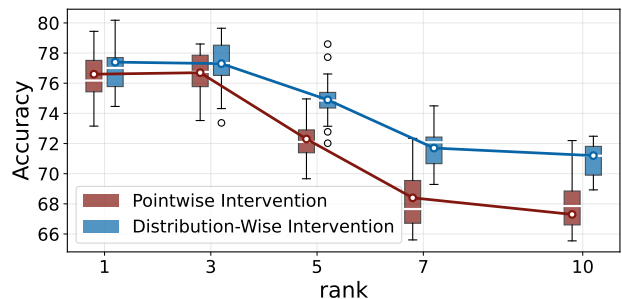


Figure 4. 在 Llama-3-8B 中对算术推理任务选择不同秩的准确性。

Table 1. LLaMA-1 7B/13B, Llama-2 7B 和 Llama-3 8B 在八个常识推理数据集上的性能与现有的 PEFT 方法进行比较。* 中所有基线方法的性能结果均取自 Wu et al. (2024b); Liu et al. (2024)。例如，在 Llama-3-8B 中，D-ReFT_{25%} 表示将前 8 层替换为 D-ReFT，并保持剩下的 24 层为 ReFT 干预。

Model	PEFT	Params (%)	Accuracy (↑)								
			BoolQ	PIQA	SIQA	HellaS.	WinoG.	ARC-e	ARC-c	OBQA	Avg.
ChatGPT *	—	—	73.1	85.4	68.5	78.5	66.1	89.8	79.9	74.8	77.0
LLaMA-7B	PrefT *	0.039 %	64.3	76.8	73.9	42.1	72.1	72.9	54.0	60.6	64.6
	Adapter ^S *	1.953 %	63.0	79.2	76.3	67.9	75.7	74.5	57.1	72.4	70.8
	Adapter ^P *	3.542 %	67.9	76.4	78.8	69.8	78.9	73.7	57.3	75.2	72.3
	LoRA *	0.826 %	68.9	80.7	77.4	78.1	78.8	77.8	61.3	74.8	74.7
	DoRA *	0.838 %	68.5	82.9	79.6	84.8	80.8	81.4	65.8	81.0	78.1
	ReFT *	0.031 %	69.3	84.4	80.3	93.1	84.2	83.2	68.2	78.9	80.2
	D-ReFT _{25 %} (Ours)	0.046 %	72.1	87.4	81.1	93.7	85.4	84.7	71.7	80.4	82.2
LLaMA-13B	PrefT *	0.031 %	65.3	75.4	72.1	55.2	68.6	79.5	62.9	68.0	68.4
	Adapter ^S *	1.586 %	71.8	83.0	79.2	88.1	82.4	82.5	67.3	81.8	79.5
	Adapter ^P *	2.894 %	72.5	84.9	79.8	92.1	84.7	84.2	71.2	82.4	81.5
	LoRA *	0.670 %	72.1	83.5	80.5	90.5	83.7	82.8	68.3	82.4	80.5
	DoRA *	0.681 %	72.4	84.9	81.5	92.4	84.2	84.2	69.6	82.8	81.5
	ReFT *	0.025 %	72.1	86.3	81.8	95.1	87.2	86.2	73.7	84.2	83.3
	D-ReFT _{25 %} (Ours)	0.037 %	74.3	87.1	83.3	95.2	89.3	87.1	73.6	85.9	85.1
Llama-2 7B	LoRA *	0.826 %	69.8	79.9	79.5	83.6	82.6	79.8	64.7	81.0	77.6
	DoRA *	0.838 %	71.8	83.7	76.0	89.1	82.6	83.7	68.2	82.4	79.7
	ReFT *	0.031 %	71.1	83.8	80.8	94.3	84.5	85.6	72.2	82.3	81.8
	D-ReFT _{25 %} (Ours)	0.046 %	71.3	86.7	81.8	94.1	87.3	86.1	73.0	84.2	83.6
Llama-3 8B	LoRA *	0.700 %	70.8	85.2	79.9	91.7	84.3	84.2	71.2	79.0	80.8
	DoRA *	0.710 %	74.6	89.3	79.9	95.5	85.6	90.5	80.4	85.8	85.2
	ReFT *	0.026 %	75.1	90.2	82.0	96.3	87.4	92.4	81.6	87.5	86.6
	D-ReFT _{25 %} (Ours)	0.039 %	78.3	93.4	83.7	96.1	89.7	94.9	83.1	89.4	89.1

分析低秩设置的结果 (图 4)，我们首先发现 ReFT 的性能并没有随着秩的增加 (更多的参数) 而提高; 相反，所有方法的性能都有所下降。这表明，仅靠参数数量无法推动性能的提升。ReFT 和 D-ReFT 在秩为 8 和 16 时达到峰值，这强调了在低维子空间中进行有针对性的干预比单纯增加参数更有效。

在我们的初步研究中，我们尝试了同义词替换 (使用 WordNet (Miller, 1994)) 和释义生成 (使用回译)。然而，我们的实证分析表明，这些语义保持的转换产生的扰动幅度不足以有效区分干预方法。因此，我们实现了一种更具挑战性的设置，通过从基准中随机删除 N 个非算术词并观察对准确性的影响，来评估 ReFT 和 D-ReFT 变体在算术任务上的鲁棒性。

当 D-ReFT 学习到 ReFT 干预的分布时，我们发现其表现出更大的鲁棒性以应对这些扰动 (参见图 5)。当删除少于 8 个单词时，D-ReFT 的准确率保持稳定，而其点对点变体下降了大约 30%。虽然删除超过 10 个单词会导致两种方法的准确率显著下降，但分布级变体在对抗攻击中仍表现出显著的鲁棒性。这表明，分布级干预在鲁棒性方面提供了额外的优势。

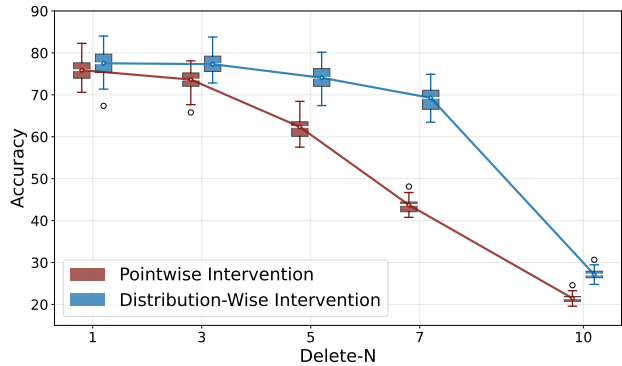


Figure 5. Llama-3-8B 在算术推理任务测试集随机删除 N 个单词时的准确率。

4. 测试时的随机性：受控温度缩放

虽然我们的分布性干预在训练过程中显示出显著的改进，但一个关键问题仍然存在：在推理时应该如何利用学到的随机分布？随后，我们通过温度缩放研究了可控的随机性，这使我们能够在测试时精细控制随机程度。

我们引入了一个温度参数 τ ，在推理过程中对学习到的

的方差进行缩放:

$$\hat{Z} = \mu + (\tau \cdot \sigma) \odot \epsilon, \quad \epsilon \sim \mathcal{N}(0, I) \quad (16)$$

, 其中 $\tau \geq 0$ 控制随机性水平:

$$\tau = 0 : \text{Deterministic inference} \quad (\hat{Z} = \mu) \quad (17)$$

$$\tau = 1 : \text{Training-time stochasticity} \quad (18)$$

$$\tau > 1 : \text{Increased exploration} \quad (19)$$

$$0 < \tau < 1 : \text{Reduced stochasticity} \quad (20)$$

。在这一节的消融研究中, 我们将 τ 设置为 0, 0.5, 1, 2。具体来说, 我们引入了指令微调的实验, 以观察在不同设置下的任务差异。我们使用 Alpaca-Eval v1.0 (Li et al., 2023b) 进行指令微调。默认情况下, 版本 1.0 计算相对于 text-davinci-003 的胜率, 同时使用 GPT-4 作为评判。提示模板由 Alpaca-Eval 提供, Alpaca-Eval 基准中的所有模型都使用该模板进行评估。在训练中, 我们使用 UltraFeedback (Cui et al., 2024), 这是一个高质量的指令微调数据集, 涵盖了各个方面, 如一般 IT 知识、真实性、诚实性和有用性, 以评估模型表现。这个设置与之前关于 RED 和 ReFT 的工作一致。

我们采用论文推荐的超参数设置, 用于基线方法如 ReFT。对于 D-ReFT, 我们直接应用了在教学算术学习数据集中使用的参数。所有结果均报告为三次运行的平均值。

表格

结果。 ?? 揭示了温度缩放如何影响不同任务类别的不同模式。较低的温度值 ($\tau < 1$) 在常识和算术推理任务上稳定提高性能, 通过确定性推理 ($\tau = 0$) 分别实现了 +0.7 和 +0.6 点的最大增益。

相反, 这些相同的低温设置会降低指令遵循性能, 表明确定性处理对于指令型任务所需的多样化响应模式可能过于僵化。较高的温度值 ($\tau \geq 1$) 则扭转了这一趋势, $\tau = 2$ 在显著改善指令遵循的同时也降低了在需要更集中、一致处理的推理任务上的性能。

在这项工作中, 我们介绍了一种分布干预框架, 其拓展了传统的点干预方法, 用于修改语言模型的表示。通过用随机节点代替确定性节点, 我们的方法能够在潜在空间中实现更稳健和细粒度的控制。通过在多个常识和算术推理基准测试中的全面评估, 我们展示了分布级干预显著提高了可控性和稳健性, 特别是在模型的早期层。我们的结果表明, 将分布感知的修改纳入模型训练可能是提高可解释性和更精确引导模型行为的一个有希望的方向, 能够实现更细致的控制。

5.

致谢 我们感谢匿名审稿人的宝贵意见。感谢吴正文对 ReFT 代码库的热心协助, 该代码库便于干预研究。我们还感谢 Chili Lab 成员对这项工作和写作提出的宝贵建议。

6.

影响声明 本文呈现的工作旨在推进机器学习领域的发展。我们的工作有许多潜在的社会影响, 我们认为这里无需特别强调。

References

- Alemi, A. A., Fischer, I., Dillon, J. V., and Murphy, K. Deep variational information bottleneck. In 5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings . OpenReview.net, 2017. URL <https://openreview.net/forum?id=HyxQzBceg>.
- Arora, A., Jurafsky, D., and Potts, C. Causalgym: Benchmarking causal interpretability methods on linguistic tasks, 2024. URL <https://arxiv.org/abs/2402.12560>.
- Behjati, M., Fehr, F., and Henderson, J. Learning to abstract with nonparametric variational information bottleneck. In Bouamor, H., Pino, J., and Bali, K. (eds.), Findings of the Association for Computational Linguistics: EMNLP 2023, pp. 1576-1586, Singapore, 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.106. URL <https://aclanthology.org/2023.findings-emnlp.106>.
- Belrose, N., Schneider-Joseph, D., Ravfogel, S., Cotterell, R., Raff, E., and Biderman, S. LEACE: perfect linear concept erasure in closed form. In Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., and Levine, S. (eds.), Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023, 2023.
- Bisk, Y., Zellers, R., LeBras, R., Gao, J., and Choi, Y. 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, pp. 7432-7439. AAAI Press, 2020. URL <https://aaai.org/ojs/index.php/AAAI/article/view/6239>.
- Chen, H. and Ji, Y. Learning variational word masks to improve the interpretability of neural text classifiers. In Webber, B., Cohn, T., He, Y., and

- Liu, Y. (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 4236–4251, Online, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.347. URL <https://aclanthology.org/2020.emnlp-main.347>.
- Clark, C., Lee, K., Chang, M.-W., Kwiatkowski, T., Collins, M., and Toutanova, K. BoolQ: Exploring the surprising difficulty of natural yes/no questions. In Burstein, J., Doran, C., and Solorio, T. (eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 2924–2936, Minneapolis, Minnesota, 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1300. URL <https://aclanthology.org/N19-1300>.
- Clark, P., Cowhey, I., Etzioni, O., Khot, T., Sabharwal, A., Schoenick, C., and Tafford, O. Think you have solved question answering? try arc, the ai2 reasoning challenge, 2018. URL <https://arxiv.org/abs/1803.05457>.
- Cobbe, K., Kosaraju, V., Bavarian, M., Chen, M., Jun, H., Kaiser, L., Plappert, M., Tworek, J., Hilton, J., Nakano, R., Hesse, C., and Schulman, J. Training verifiers to solve math word problems, 2021. URL <https://arxiv.org/abs/2110.14168>.
- Conmy, A., Mavor-Parker, A. N., Lynch, A., Heimersheim, S., and Garriga-Alonso, A. Towards automated circuit discovery for mechanistic interpretability. In Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., and Levine, S. (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023.
- Cui, G., Yuan, L., Ding, N., Yao, G., He, B., Zhu, W., Ni, Y., Xie, G., Xie, R., Lin, Y., Liu, Z., and Sun, M. ULTRA FEEDBACK: boosting language models with scaled AI feedback. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net, 2024. URL <https://openreview.net/forum?id=BOorDpKHjJ>.
- Deng, C., Li, Z., Xie, R., Chang, R., and Chen, H. Language models are symbolic learners in arithmetic, 2024. URL <https://arxiv.org/abs/2410.15580>.
- Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Yang, A., Fan, A., Goyal, A., Hartshorn, A., Yang, A., Mitra, A., Sravankumar, A., Korenev, A., Hinsvark, A., Rao, A., Zhang, A., Rodriguez, A., Gregerson, A., Spataru, A., Roziere, B., Biron, B., Tang, B., Chern, B., Caucheteux, C., Nayak, C., Bi, C., and et al. The llama 3 herd of models, 2024. URL <https://arxiv.org/abs/2407.21783>.
- Fu, C., Huang, H., Chen, X., Tian, Y., and Zhao, J. Learn-to-share: A hardware-friendly transfer learning framework exploiting computation and parameter sharing. In Meila, M. and Zhang, T. (eds.), *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pp. 3469–3479. PMLR, 2021. URL <http://proceedings.mlr.press/v139/fu21a.html>.
- Gandikota, R., Materzynska, J., Zhou, T., Torralba, A., and Bau, D. Concept sliders: Lora adaptors for precise control in diffusion models, 2023. URL <https://arxiv.org/abs/2311.12092>.
- Geiger, A., Lu, H., Icard, T., and Potts, C. Causal abstractions of neural networks. In Ranzato, M., Beygelzimer, A., Dauphin, Y. N., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pp. 9574–9586, 2021.
- Geiger, A., Wu, Z., Potts, C., Icard, T., and Goodman, N. D. Finding alignments between interpretable causal variables and distributed neural representations, 2023. URL <https://arxiv.org/abs/2303.02536>.
- Geva, M., Schuster, R., Berant, J., and Levy, O. Transformer feed-forward layers are key-value memories. In Moens, M.-F., Huang, X., Specia, L., and Yih, S. W.-t. (eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 5484–5495, Online and Punta Cana, Dominican Republic, 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.446. URL <https://aclanthology.org/2021.emnlp-main.446>.
- Ghandeharioun, A., Caciularu, A., Pearce, A., Dixon, L., and Geva, M. Patchscopes: A unifying framework for inspecting hidden representations of language models. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net, 2024. URL <https://openreview.net/forum?id=5uwBzcn885>.

- Gur-Arieh, Y., Mayan, R., Agassy, C., Geiger, A., and Geva, M. Enhancing automated interpretability with output-centric feature descriptions, 2025. URL <https://arxiv.org/abs/2501.08319>.
- Han, C., Xu, J., Li, M., Fung, Y., Sun, C., Jiang, N., Abdelzaher, T., and Ji, H. Word embeddings are steers for language models, 2023. URL <https://arxiv.org/abs/2305.12798>.
- Hanna, M., Belinkov, Y., and Pezzelle, S. When language models fall in love: Animacy processing in transformer language models. In Bouamor, H., Pino, J., and Bali, K. (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 12120–12135, Singapore, 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.744. URL <https://aclanthology.org/2023.emnlp-main.744>.
- Hewitt, J. and Manning, C. D. A structural probe for finding syntax in word representations. In Burstein, J., Doran, C., and Solorio, T. (eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4129–4138, Minneapolis, Minnesota, 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1419. URL <https://aclanthology.org/N19-1419>.
- Hosseini, M. J., Hajishirzi, H., Etzioni, O., and Kushman, N. Learning to solve arithmetic word problems with verb categorization. In Moschitti, A., Pang, B., and Daelemans, W. (eds.), *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 523–533, Doha, Qatar, 2014. Association for Computational Linguistics. doi: 10.3115/v1/D14-1058. URL <https://aclanthology.org/D14-1058>.
- Houlsby, N., Giurugu, A., Jastrzebski, S., Morrone, B., de Laroussilhe, Q., Gesmundo, A., Attariyan, M., and Gelly, S. Parameter-efficient transfer learning for NLP. In Chaudhuri, K. and Salakhutdinov, R. (eds.), *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2790–2799. PMLR, 2019. URL <http://proceedings.mlr.press/v97/houlsby19a.html>.
- Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. Lora: Low-rank adaptation of large language models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022. URL <https://openreview.net/forum?id=nZeVKeeFYf9>.
- Hu, Z., Wang, L., Lan, Y., Xu, W., Lim, E.-P., Bing, L., Xu, X., Poria, S., and Lee, R. LLM-adapters: An adapter family for parameter-efficient fine-tuning of large language models. In Bouamor, H., Pino, J., and Bali, K. (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 5254–5276, Singapore, 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.319. URL <https://aclanthology.org/2023.emnlp-main.319>.
- Huang, J., Wu, Z., Potts, C., Geva, M., and Geiger, A. Ravel: Evaluating interpretability methods on disentangling language model representations, 2024. URL <https://arxiv.org/abs/2402.17700>.
- Kingma, D. P. and Welling, M. Auto-encoding variational bayes. In Bengio, Y. and LeCun, Y. (eds.), *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014. URL <http://arxiv.org/abs/1312.6114>.
- Koncel-Kedziorski, R., Hajishirzi, H., Sabharwal, A., Etzioni, O., and Ang, S. D. Parsing algebraic word problems into equations. *Transactions of the Association for Computational Linguistics*, 3:585–597, 2015. doi: 10.1162/tacl_a_00160. URL <https://aclanthology.org/Q15-1042>.
- Koncel-Kedziorski, R., Roy, S., Amini, A., Kushman, N., and Hajishirzi, H. MAWPS: A math word problem repository. In Knight, K., Nenkova, A., and Rambow, O. (eds.), *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 1152–1157, San Diego, California, 2016. Association for Computational Linguistics. doi: 10.18653/v1/N16-1136. URL <https://aclanthology.org/N16-1136>.
- Lasri, K., Pimentel, T., Lenci, A., Poibeau, T., and Cotterell, R. Probing for the usage of grammatical number. In Muresan, S., Nakov, P., and Villavicencio, A. (eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8818–8831, Dublin, Ireland, 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.603. URL <https://aclanthology.org/2022.acl-long.603>.
- Li, K., Patel, O., Viégas, F. B., Pfister, H., and Wattenberg, M. Inference-time intervention: Eliciting truthful answers from a language model. In

- Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., and Levine, S. (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023a.
- Li, X., Zhang, T., Dubois, Y., Taori, R., Gulrajani, I., Guestrin, C., Liang, P., and Hashimoto, T. B. AlpacaEval: An automatic evaluator of instruction-following models. https://github.com/tatsu-lab/alpaca_eval, 2023b.
- Li, X. L. and Eisner, J. Specializing word embeddings (for parsing) by information bottleneck. In Inui, K., Jiang, J., Ng, V., and Wan, X. (eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 2744–2754, Hong Kong, China, 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1276. URL <https://aclanthology.org/D19-1276>.
- Li, X. L. and Liang, P. Prefix-tuning: Optimizing continuous prompts for generation. In Zong, C., Xia, F., Li, W., and Navigli, R. (eds.), *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. 4582–4597, Online, 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.353. URL <https://aclanthology.org/2021.acl-long.353>.
- Ling, W., Yogatama, D., Dyer, C., and Blunsom, P. Program induction by rationale generation: Learning to solve and explain algebraic word problems. In Barzilay, R. and Kan, M.-Y. (eds.), *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 158–167, Vancouver, Canada, 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-1015. URL <https://aclanthology.org/P17-1015>.
- Liu, S., Wang, C., Yin, H., Molchanov, P., Wang, Y. F., Cheng, K., and Chen, M. Dora: Weight-decomposed low-rank adaptation. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net, 2024. URL <https://openreview.net/forum?id=3d5CIRG1n2>.
- Mahabadi, R. K., Belinkov, Y., and Henderson, J. Variational information bottleneck for effective low-resource fine-tuning. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL <https://openreview.net/forum?id=kvhzKz-DMF>.
- Meng, K., Bau, D., Andonian, A., and Belinkov, Y. Locating and editing factual associations in GPT. In Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., and Oh, A. (eds.), *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022.
- Mihaylov, T., Clark, P., Khot, T., and Sabharwal, A. Can a suit of armor conduct electricity? a new dataset for open book question answering. In Riloff, E., Chiang, D., Hockenmaier, J., and Tsujii, J. (eds.), *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2381–2391, Brussels, Belgium, 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1260. URL <https://aclanthology.org/D18-1260>.
- Miller, G. A. WordNet: A lexical database for English. In *Human Language Technology: Proceedings of a Workshop held at Plainsboro, New Jersey, March 8-11, 1994*, 1994. URL <https://aclanthology.org/H94-1111/>.
- Nanda, N., Lee, A., and Wattenberg, M. Emergent linear representations in world models of self-supervised sequence models. In Belinkov, Y., Hao, S., Jumelet, J., Kim, N., McCarthy, A., and Mohebbi, H. (eds.), *Proceedings of the 6th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pp. 16–30, Singapore, 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.blackboxnlp-1.2. URL <https://aclanthology.org/2023.blackboxnlp-1.2>.
- Olah, C. et al. Understanding lstm networks. 2015.
- Park, K., Choe, Y. J., and Veitch, V. The linear representation hypothesis and the geometry of large language models. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net, 2024. URL <https://openreview.net/forum?id=UGpGkLzwpP>.
- Patel, A., Bhattamishra, S., and Goyal, N. Are NLP models really able to solve simple math word problems? In Toutanova, K., Rumshisky, A., Zettlemoyer, L., Hakkani-Tur, D., Beltagy, I., Bethard, S., Cotterell, R., Chakraborty, T., and Zhou, Y.

- (eds.), *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 2080–2094, Online, 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.168. URL <https://aclanthology.org/2021.naacl-main.168>.
- Pfeiffer, J., Vulić, I., Gurevych, I., and Ruder, S. MAD-X: An Adapter-Based Framework for Multi-Task Cross-Lingual Transfer. In Webber, B., Cohn, T., He, Y., and Liu, Y. (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 7654–7673, Online, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.617. URL <https://aclanthology.org/2020.emnlp-main.617>.
- Ravfogel, S., Twiton, M., Goldberg, Y., and Cotterell, R. Linear adversarial concept erasure. In Chaudhuri, K., Jegelka, S., Song, L., Szepesvári, C., Niu, G., and Sabato, S. (eds.), *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pp. 18400–18421. PMLR, 2022. URL <https://proceedings.mlr.press/v162/ravfogel22a.html>.
- Roy, S. and Roth, D. Solving general arithmetic word problems. In Márquez, L., Callison-Burch, C., and Su, J. (eds.), *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 1743–1752, Lisbon, Portugal, 2015. Association for Computational Linguistics. doi: 10.18653/v1/D15-1202. URL <https://aclanthology.org/D15-1202>.
- Rumelhart, D. E., McClelland, J. L., Group, P. R., et al. *Parallel distributed processing, volume 1: Explorations in the microstructure of cognition: Foundations*. The MIT press, 1986.
- Sakaguchi, K., Bras, R. L., Bhagavatula, C., and Choi, Y. Winogrande: An adversarial winograd schema challenge at scale. 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*, pp. 8732–8740. AAAI Press, 2020. URL <https://aaai.org/ojs/index.php/AAAI/article/view/6399>.
- Sap, M., Rashkin, H., Chen, D., Le Bras, R., and Choi, Y. Social IQa: Commonsense reasoning about social interactions. In Inui, K., Jiang, J., Ng, V., and Wan, X. (eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 4463–4473, Hong Kong, China, 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1454. URL <https://aclanthology.org/D19-1454>.
- Shazeer, N. Glu variants improve transformer, 2020. URL <https://arxiv.org/abs/2002.05202>.
- Stolfo, A., Belinkov, Y., and Sachan, M. A mechanistic interpretation of arithmetic reasoning in language models using causal mediation analysis. In Bouamor, H., Pino, J., and Bali, K. (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 7035–7052, Singapore, 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.435. URL <https://aclanthology.org/2023.emnlp-main.435>.
- Subramani, N., Suresh, N., and Peters, M. Extracting latent steering vectors from pretrained language models. In Muresan, S., Nakov, P., and Villavicencio, A. (eds.), *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 566–581, Dublin, Ireland, 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-acl.48. URL <https://aclanthology.org/2022.findings-acl.48>.
- Tian, R., Mao, Y., and Zhang, R. Learning VAE-LDA models with rounded reparameterization trick. In Webber, B., Cohn, T., He, Y., and Liu, Y. (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1315–1325, Online, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.101. URL <https://aclanthology.org/2020.emnlp-main.101>.
- Tishby, N. and Zaslavsky, N. Deep learning and the information bottleneck principle, 2015. URL <https://arxiv.org/abs/1503.02406>.
- Tishby, N., Pereira, F. C., and Bialek, W. The information bottleneck method, 2000. URL <https://arxiv.org/abs/physics/0004057>.
- Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., and Lample, G. Llama: Open and efficient foundation language models, 2023a. URL <https://arxiv.org/abs/2302.13971>.

- Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., Bikel, D., Blecher, L., Ferrer, C. C., Chen, M., Cucurull, G., and et al. Llama 2: Open foundation and fine-tuned chat models, 2023b. URL <https://arxiv.org/abs/2307.09288>.
- Turner, A. M., Thiergart, L., Leech, G., Udell, D., Vazquez, J. J., Mini, U., and MacDiarmid, M. Steering language models with activation engineering, 2023. URL <https://arxiv.org/abs/2308.10248>.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. Attention is all you need. In Guyon, I., von Luxburg, U., Bengio, S., Wallach, H. M., Fergus, R., Vishwanathan, S. V. N., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pp. 5998–6008, 2017.
- Wang, K. R., Variengien, A., Conmy, A., Shlegeris, B., and Steinhardt, J. Interpretability in the wild: a circuit for indirect object identification in GPT-2 small. 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=NpsVSN6o4ul>.
- White, A. S., Rastogi, P., Duh, K., Van, B., and Inference, D. . Improving fine-tuning on low-resource corpora with information bottleneck. 2020. URL <https://api.semanticscholar.org/CorpusID:219978318>.
- Wu, M., Liu, W., Wang, X., Li, T., Lv, C., Ling, Z., Zhu, J., Zhang, C., Zheng, X., and Huang, X. Advancing parameter efficiency in fine-tuning via representation editing, 2024a. URL <https://arxiv.org/abs/2402.15179>.
- Wu, Z., Geiger, A., Icard, T., Potts, C., and Goodman, N. D. Interpretability at scale: Identifying causal mechanisms in alpaca. In Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., and Levine, S. (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023.
- Wu, Z., Arora, A., Wang, Z., Geiger, A., Jurafsky, D., Manning, C. D., and Potts, C. Reft: Representation finetuning for language models. In Globerson, A., Mackey, L., Belgrave, D., Fan, A., Paquet, U., Tomczak, J. M., and Zhang, C. (eds.), *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024b.
- Wu, Z., Geiger, A., Arora, A., Huang, J., Wang, Z., Goodman, N., Manning, C., and Potts, C. pyvene: A library for understanding and improving PyTorch models via interventions. In Chang, K.-W., Lee, A., and Rajani, N. (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 3: System Demonstrations)*, pp. 158–165, Mexico City, Mexico, 2024c. Association for Computational Linguistics. URL <https://aclanthology.org/2024.naacl-demo.16>.
- Yin, F., Ye, X., and Durrett, G. Lofit: Localized fine-tuning on LLM representations. In Globerson, A., Mackey, L., Belgrave, D., Fan, A., Paquet, U., Tomczak, J. M., and Zhang, C. (eds.), *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024.
- Yu, Y., Buchanan, S., Pai, D., Chu, T., Wu, Z., Tong, S., Haeffele, B. D., and Ma, Y. White-box transformers via sparse rate reduction. In Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., and Levine, S. (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023.
- Zellers, R., Holtzman, A., Bisk, Y., Farhadi, A., and Choi, Y. HellaSwag: Can a machine really finish your sentence? In Korhonen, A., Traum, D., and Màrquez, L. (eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4791–4800, Florence, Italy, 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1472. URL <https://aclanthology.org/P19-1472>.
- Zhang, F. and Nanda, N. Towards best practices of activation patching in language models: Metrics and methods. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL <https://openreview.net/forum?id=Hf17y6u9BC>.
- Zhang, R., Han, J., Liu, C., Gao, P., Zhou, A., Hu, X., Yan, S., Lu, P., Li, H., and Qiao, Y. Llama-adaptor: Efficient fine-tuning of language models

with zero-init attention, 2023. URL <https://arxiv.org/abs/2303.16199>.

Zhang, R., Qiang, R., Somayajula, S. A., and Xie, P. AutoLoRA: Automatically tuning matrix ranks in low-rank adaptation based on meta learning. In Duh, K., Gomez, H., and Bethard, S. (eds.), Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pp. 5048–5060, Mexico City, Mexico, 2024. Association for Computational Linguistics. URL <https://aclanthology.org/2024.naacl-long.282>.

Zou, A., Phan, L., Chen, S., Campbell, J., Guo, P., Ren, R., Pan, A., Yin, X., Mazeika, M., Dombrowski, A.-K., Goel, S., Li, N., Byun, M. J., Wang, Z., Mallen, A., Basart, S., Koyejo, S., Song, D., Fredrikson, M., Kolter, J. Z., and Hendrycks, D. Representation engineering: A top-down approach to ai transparency, 2023. URL <https://arxiv.org/abs/2310.01405>.

数据处理不等式 (DPI) 是信息论中的一个基本概念, 用于描述信息在处理流水线中的丢失。在深度神经网络中, 互信息满足: 其中, $I(\cdot; \cdot)$ 表示两个随机变量之间的互信息。

直观上, 这个不等式暗示了数据处理 (即从 Z 到 Y 的转换) 不能增加关于原始变量 X 的信息。换句话说, 对 Z 的任何转换或操作都不能恢复已经丢失的关于 X 的信息。在语言模型的背景下, 这一概念尤为重要。第 l 层的隐藏表示 $Z^{(l)}$ 编码了关于输入序列 X 的信息。随着这些表示通过注意力机制和前馈网络逐层转化, 数据处理不等式 (DPI) 表明输入 X 和最终输出 Y 之间的互信息在网络深度增加时是不增加的:

$$I(Y; X) \geq I(X; Z^{(L)}) \geq \dots \geq I(X; Z^{(1)}) \geq I(Y; \hat{Y}) \quad (21)$$

这突显了深层架构中的一种权衡: 尽管更深的层可能会为特定任务优化表示, 但它们无法恢复在较早层中丢失的信息。

根据互信息的链式法则, 最小化 $\mathcal{L}_{CE}$ 的同时最大化了 $I(Y; \hat{Y})$, 并隐式地鼓励学习到的表示 $Z^{(L)}$ 保留关于 Y 的足够任务相关信息。

7. 超参数配置

在这一部分中, 我们将讨论以前方法的所有超参数设置。

Hyperparameter Configuration

We investigate the following hyperparameters for our experiments:

- 干预层 (l): 在模型中应用干预的特定层。这是根据架构和对模型行为的期望影响来选择的。
- 噪声尺度 (ϵ): 干预过程中添加的噪声大小。这控制了对模型激活引入的扰动水平。
- 子空间秩 (r): 用于干预的子空间的秩。这决定了干预操作所在子空间的维数。
- 干预位置 (p): 干预应用于层中的位置 (例如, 在特定操作如激活或归一化之前或之后)。
- 批量大小 (bs): 训练过程中每批处理的样本数量。这会影响训练过程的稳定性和速度。
- 训练轮数 (e): 模型在整个数据集上训练的总次数。这会影响模型的收敛性和泛化能力。
- 学习率 (lr): 训练过程中模型参数更新的步长。它控制学习过程的速度和稳定性。

7.1. ReFT

ReFT Hyperparameter Configuration

We investigate the following hyperparameters for our experiments:

- 学习率 (lr): $9e-4$ 。
- 子空间秩 (r): 8 或 16 效果最佳。更高的秩如 256 并不会带来提升。
- 干预位置 (p): $f7+l7$: 前七个词和后七个词。
- 批大小 (bs): 8。我们还启用了梯度检查点, 累计步骤 = 4。由于内存限制, 我们无法消融这些参数。
- 训练轮次 (e): 12 效果最佳。在我们的实验中, 将训练轮次减少到 9 会导致性能下降。
- 干预层 (l): 取决于实验设置, 但我们发现较早的层效果最好。

7.2. D-ReFT

ReFT Hyperparameter Configuration

We investigate the following hyperparameters for our experiments:

- 学习率 (lr): $1e-3$ 或 $3e-3$ 。
- 子空间秩 (r): 8 或 16 效果最佳。较高的秩如 256 并没有带来提升。
- 干预位置 (p): $f7+l7$: 前七个词元和后七个词元。
- 批量大小 (bs): 8。我们还启用了梯度检查点, 累计步骤为 4。由于内存限制, 我们无法对这些参数进行消融实验。
- 训练周期数 (e): 9 效果最佳, 这表明 D-ReFT 也具有更好的收敛特性。
- 干预层 (l): 取决于实验设置, 但我们发现早期层效果最佳。

7.4. 红色。

RED Hyperparameter Configuration

We investigate the following hyperparameters for our experiments:

- 学习率 (lr): $7e-4$ 。
- 干预位置 (p): all : 所有符号位置都被干预。
- 批量大小 (bs): 8。我们还启用了累计步骤为 4 的梯度检查点。由于内存限制, 我们无法去除这些参数。
- 训练轮数 (e): 9 效果最好。
- 干预层 (l): 取决于实验设置, 但我们发现早期层效果最好。

7.3. LoRA

ReFT Hyperparameter Configuration

We investigate the following hyperparameters for our experiments:

- 学习率 (lr): $4e-4$ 。
- 阿尔法 (α): 16。
- 子空间秩 (r): 16 效果最好。
- 干预位置 (p): all : 所有标记位置都进行了干预。
- 批量大小 (bs): 8。我们还启用了梯度检查点, 累积步数为 4。由于内存限制, 我们无法对这些参数进行消融。
- 训练周期 (e): 6 效果最好。
- 干预层 (l): 取决于实验设置, 但我们发现早期的层效果最好。

7.5. SwiGLU

SwiGLU Hyperparameter Configuration

We investigate the following hyperparameters for our experiments:

- 学习率 (lr): $6e-4$ 。
- 干预位置 (p): all : 所有的符号位置都被干预。
- 批量大小 (bs): 8。我们还通过使用累积步数 = 4 启用梯度检查点。由于内存限制, 我们无法对这些参数进行消融。
- 训练周期 (e): 9 效果最佳。
- 干预层 (l): 取决于实验设置, 但我们发现早期层效果最好。

7.6. 多层感知器 (MLP)。

MLP Hyperparameter Configuration

We investigate the following hyperparameters for our experiments:

- 学习率 (lr): $5e-4$ 。
- 干预位置 (p): *all*: 所有的标记位置都被干预。
- 批大小 (bs): 8。我们还启用了具有累积步骤 = 4 的梯度检查点。由于内存限制, 我们无法消融此参数。
- 训练轮数 (e): 9 效果最佳。
- 干预层 (l): 取决于实验设置, 但我们发现早期层效果最佳。

8. 数据集。

在本节中, 我们将介绍本文中使用的基准测试。通常, 我们遵循 [Hu et al. \(2023\)](#); [Wu et al. \(2024b\)](#) 的设置, 使用八个常识基准测试和七个算术基准测试进行评估。

8.1. 常识推理

BoolQ。BoolQ ([Clark et al., 2019](#)) 数据集是一个由自然语言问题和相应段落组成的集合, 旨在进行二元问题回答任务。该数据集包含超过 15,000 个示例, 其中每个问题都可以根据所提供段落中的信息用简单的“是”或“否”来回答。数据集来源于真实用户查询和网页, 使其成为训练和评估模型理解上下文和进行文本推理的宝贵资源。BoolQ 在自然语言处理研究中被广泛用于对需要理解、推理和二元分类任务的模型性能进行基准测试。其挑战性在于需要模型掌握问题和段落之间的细微关系, 使其成为推动问答系统进步的关键数据集。

An example of data from the BoolQ dataset

Instructions: Please answer the following question with true or false, question: does ethanol take more energy make that produces?
Answer format: true/false.

PIQA ([Bisk et al., 2020](#)) 数据集是一个基准, 旨在评估模型在日常场景中对物理常识推理的理解。它由需要推理物体在物理世界中如何互动、使用或操控的问题组成。每个问题都针对一个实际问题提供两个可能的解决方案, 任务是基于现实世界的物理和直觉选择最合适的一个。PIQA 挑战模型超越文本知识, 并结合对物理属性、因果关系和可供性理解, 这使其成为在需要基于现实世界推理的任务中推进 AI 系统的宝

贵资源。

An example of data from the PIQA dataset

Instructions: Please choose the correct solution to the question: How do I ready a guinea pig cage for it's new occupants?

Solution1 : Provide the guinea pig with a cage full of a few inches of bedding made of ripped paper strips, you will also need to supply it with a water bottle and a food dish.

Solution2: Provide the guinea pig with a cage full of a few inches of bedding made of ripped jeans material, you will also need to supply it with a water bottle and a food dish.

Answer format: solution1 or solution2

HellaSwag。The HellaSwag ([Zellers et al., 2019](#)) 数据集是一个基准测试, 旨在评估自然语言理解模型的常识推理能力。它于 2019 年推出, 由多个选择题组成, 需要模型预测给定情境的最合理继续, 依赖于日常知识和情境理解。与许多其他数据集不同, HellaSwag 强调真实世界的情况和细致的推理, 使得最先进的模型特别具有挑战性。创建该数据集是为了应对之前基准测试的局限性, 这些测试往往依赖于数据中的肤浅模式或偏见。通过专注于需要更深入理解和推理的情境, HellaSwag 已成为推进人工智能研究和提高语言模型鲁棒性的重要工具。

An example of data from the HellaSwag dataset

Instructions: Please choose the correct ending to complete the given sentence: Roof shingle removal: A man is sitting on a roof. he

Ending1: is using wrap to wrap a pair of skis.

Ending2: is ripping level tiles off.

Ending3: is holding a rubik's cube.

Ending4 : starts pulling up roofing on a roof.

Answer format : ending1/ending2/ending3/ending4.

WinoGrande。WinoGrande ([Sakaguchi et al., 2020](#)) 数据集是一个大规模的自然语言推理问题集合, 旨在评估人工智能系统的推理能力, 特别是在常识推理的背景下。作为 Winograd Schema Challenge 的更具挑战性的后继者, WinoGrande 包含超过 44,000 个精心设计的代词解析问题, 这些问题需要理解上下文、世界知识和微妙的语言线索。每个问题提供一段简短的文字, 其中包含一个模糊的代词, 任务是从两个可能的选项中确定正确的指代对象。为了解决偏见并确保稳健性, 数据集是通过众包方法创建的, 随后进行系统的对抗性过滤过程。WinoGrande 已成为测试机器学习模型在处理复杂推理任务方面极限的基准, 推动了人工智能系统朝更类似人类的理解和决策方向发展。

An example of data from the WinoGrande dataset

Instructions: Please choose the correct answer to fill in the blank to complete the given sentence: Sarah was a much better surgeon than Maria so always got the easier cases.

Option1 : Sarah

Option2 : Maria

Answer format : option1/option2

ARC-e. ARC-e (AI2 推理挑战-简单) (Clark et al., 2018) 数据集是一个包含小学层次科学问题的集合, 旨在评估人工智能系统的推理和理解能力。ARC-e 由艾伦人工智能研究所开发, 专注于要求对科学概念有基本理解的多项选择题, 使其成为人工智能模型可用但具有挑战性的基准。与更高级的对标 ARC (AI2 推理挑战) 不同, ARC-e 旨在评估基本知识和简单推理, 通常从早期教育中的主题中获取信息。通过提供简化但多样化的问题集, ARC-e 作为评测 AI 系统处理和回答科学相关问题基础能力的有价值工具, 为日后更复杂的推理任务铺平了道路。

An example of data from the ARC-e dataset

Instructions: Please choose the correct answer to the question: Which statement best explains why photosynthesis is the foundation of most food webs?

Answer1 : Sunlight is the source of energy for nearly all ecosystems.

Answer2 : Most ecosystems are found on land instead of in water.

Answer3 : Carbon dioxide is more available than other gases.

Answer4 : The producers in all ecosystems are plants.

Answer format : answer1/answer2/answer3/answer4

ARC-c ARC-c (Clark et al., 2018) 数据集是 AI2 推理挑战 (ARC) 的一部分, 是一个全面的科学问题集合, 旨在评估人工智能系统的推理和理解能力。该数据集包含来自不同年级水平的多项选择题, 强调复杂的推理, 要求模型超越简单的检索, 参与更深层次的理解和推论。这些问题来自多种科学领域, 包括生物学、化学、物理学和地球科学, 使之成为评估 AI 的泛化能力和解决问题技巧的坚实基准。通过专注于具有挑战性和课程对齐的内容, ARC-c 数据集作为一种关键工具推动能够进行细致和具备上下文意识推理的 AI 系统的发展。

An example of data from the ARC-c dataset

Instructions: Please choose the correct answer to the question: A group of engineers wanted to know how different building designs would respond during an earthquake. They made several models of buildings and tested each for its ability to withstand earthquake conditions. Which will most likely result from testing different building designs?

Answer1 : buildings will be built faster

Answer2 : buildings will be made safer

Answer3 : building designs will look nicer

Answer4: building materials will be cheaper

Answer format : answer1/answer2/answer3/answer4

开放书本问答 开卷问答 (OBQA) (Mihaylov et al., 2018) 数据集是一个基准, 用于评估机器学习模型通过结合开卷事实检索和推理技能来回答基于科学的问题的能力。与传统的问答数据集不同, OBQA 要求系统不仅从提供的知识来源中检索相关信息, 还需应用逻辑推理来推断正确答案。该数据集由多个选择题组成, 涵盖广泛的科学主题, 挑战模型展示其理解和分析能力。通过强调外部知识和推理的整合, OBQA 成为推进问答系统、知识表示和人工智能研究的有价值工具。

An example of data from the OBQA dataset

Instructions: Please choose the correct answer to the question: A person wants to start saving money so that they can afford a nice vacation at the end of the year. After looking over their budget and expenses, they decide the best way to save money is to

Answer1 : make more phone calls

Answer2 : quit eating lunch out

Answer3 : buy less with monopoly money

Answer4 : have lunch with friends

Answer format : answer1/answer2/answer3/answer4

8.2. 算术推理

AQuA. AQuA (Algebraic Question Answering) (Ling et al., 2017) 数据集是一个专门设计的资源, 旨在推动数学推理和自然语言处理领域的研究。它由一组代数问题组成, 每个题目都有多个选择答案和详细的逐步解决方案。该数据集对于训练和评估旨在理解和解决用自然语言表述的复杂数学问题的模型尤其有价值。通过提供问题陈述及其相应的解决方案, AQuA 使研究人员能够开发和测试算法, 这些算法不仅能生成正确答案, 还能解释背后的推理过程。

这使得 AQuA 成为探索数学与人工智能交叉领域的重要工具，促进自动化问题解决和教育技术的进步。

An example of data from the AQuA dataset

说明： A rectangular piece of 150 sq m has a length which is 1m more than the 4 times the breadth. What is the perimeter of the piece?
Answer Choices: (A) 60 m (B) 61 m (C) 62 m (D) 63 m (E) 64 m

Let's assume the breadth of the rectangle to be x . According to the problem, the length of the rectangle is 1 m more than 4 times the breadth. Therefore, the length of the rectangle is $4x + 1$.

We know that the area of the rectangle is 150 sq m. Therefore, we can write:

$$\text{Area} = \text{Length} \times \text{Breadth}$$

$$150 = (4x + 1) \times x$$

Simplifying this equation, we get:

$$150 = 4x^2 + x$$

$$150 - x = 4x^2$$

$$x^2 - \frac{1}{4}x - \frac{15}{4} = 0$$

Solving for x , we get:

$$x = 1.732 \quad \text{or} \quad x = -0.732$$

Since the breadth cannot be negative, we can ignore the negative value of x . Therefore, the breadth of the rectangle is 1.732 m.

Now, we can find the length of the rectangle:

$$\text{Length} = 4x + 1 = 4(1.732) + 1 = 7.148 \text{ m}$$

The perimeter of the rectangle is the sum of the length of all four sides. Therefore, the perimeter is:

$$\text{Perimeter} = 2(\text{Length} + \text{Breadth}) = 2(7.148 + 1.732) = 17.76$$

Rounding off to the nearest integer, we get the answer as 17. Therefore, the answer is

(E) 64 m.

理解问题的语言细微差别，同时也要求模型准确执行必要的计算。通过关注基本的算术运算，AddSub 作为评估 NLP 系统推理和理解能力的重要测试平台，使其成为机器学习和人工智能研究中的宝贵资源。

An example of data from the AddSub dataset

说明： There are 7 crayons in the drawer. Mary took 3 crayons out of the drawer. How many crayons are there now?

Step 1: Start with the total number of crayons in the drawer: 7

Step 2: Subtract the number of crayons Mary took out: 3

Step 3: Perform the subtraction: $7 - 3 = 4$

Answer : There are now 4 crayons in the drawer.

GSM8K. GSM8K (Cobbe et al., 2021) 数据集是一个由 8.5K 高质量、语言多样的小学数学文字题组成的集合，旨在基准测试 AI 系统的问题解决能力。数据集 中的每个问题都需要多个推理步骤才能得出正确解决方案，这使其成为评估语言模型数学和逻辑推理能力的宝贵资源。这些问题的设计反映了现实世界的场景，确保它们既具有挑战性又对小学生可接受。通过提供贯穿各种数学概念的多样化问题集，GSM8K 成为推动 AI 系统开发进步的强大测试平台，使其能够理解和解决复杂的多步骤问题。

An example of data from the GSM8K dataset

指示： Mr Boarden is remodeling his bathroom. For every square foot, he needs 24 mosaic tiles. How many mosaic tiles would Mr Boarden need to cover two thirds of his 36 sq ft bathroom?

Step 1: Find the total area of the bathroom: 36 sq ft

Step 2: Calculate the area to be covered by the mosaic tiles

$$36 \text{ sq ft} \times \frac{2}{3} = 24 \text{ sq ft}$$

Step 3: Calculate the number of mosaic tiles needed

$$24 \text{ sq ft} / 1 \text{ sq ft per 24 tiles} = 1 \text{ tile}$$

Therefore, Mr Boarden would need 1 mosaic tile to cover two thirds of his 36 sq ft bathroom.

The answer in Arabic numerals is 1.

加减法。 The AddSub (Hosseini et al., 2014) 数据集是一个广泛使用的 NLP 基准，旨在评估模型解决算术文字题的能力。它由一对对的问题和答案组成，其中每个问题是一个涉及加法或减法的数学问题的文字描述，相应的答案是数值结果。该数据集不仅挑战模型

MAWPS。MAWPS (数学应用问题求解) (Koncel-Kedziorski et al., 2016)，是一个全面的数学应用问题集合，旨在促进自然语言处理领域的研究。该数据集涵盖了包括算术、代数和几何在内的多种类型问题，以自然语言形式呈现，使其成为开发和评估旨在通过

自动推理和语言理解解决数学问题的算法的宝贵资源。MAWPS 特别适用于训练和测试需要兼具语言理解和数学问题解决能力的模型，弥合了文本数据和定量推理之间的差距。通过提供具有不同复杂程度的问题集，MAWPS 作为一个基准，推进 AI 系统在教育应用及其他领域的能力发展。

An example of data from the MAWPS dataset

说明： Nicole found an orange caterpillar and a green caterpillar in her backyard . The green caterpillar was 3 inches long and the orange caterpillar was 1.1666666666666667 inches long . How much longer was the green caterpillar than the orange caterpillar ?

To find the difference in length between the green and orange caterpillars, we need to subtract the length of the orange caterpillar from the length of the green caterpillar.

Length of green caterpillar - Length of orange caterpillar = Difference in length

3 inches - 1.1666666666666667 inches = 1.8333333333333333 inches

Therefore, the green caterpillar was 1.8333333333333333 inches longer than the orange caterpillar.

The answer in Arabic numerals is 1.8333333333333333

MultiArith. MultiArith (Roy & Roth, 2015) 数据集是一个用于评估机器学习模型性能的基准集合，尤其是解决需要多步推理的算术文字问题。该数据集包含以自然语言呈现的各种数学问题，挑战模型去解释文本，提取相关的数据信息，并执行连续的算术运算以得出正确的解决方案。通过包含涉及加法、减法、乘法和除法的问题，通常是组合在一起的，MultiArith 测试 AI 系统的语言理解能力和数学推理能力。它是推动 NLP 研究前进的宝贵资源，特别是在弥合语言理解与定量问题解决之间的差距方面。

An example of data from the MultiArith dataset

说明： At the schools book fair Sam bought 13 adventure books and 17 mystery books. If 15 of the books were used, how many new books did he buy?

To solve the problem, we need to find the total number of books Sam bought and then subtract the number of used books from that total.

1. Find the total number of books Sam bought:
13 adventure books + 17 mystery books = 30 books

2. Subtract the number of used books from the total:

30 books - 15 used books = 15 new books

Therefore, Sam bought 15 new books.

The answer in Arabic numerals is 15.

单方程 The SingleEq (Koncel-Kedziorski et al., 2015) 数据集是一种专门的资源，旨在支持数学问题求解领域的研究和开发，特别关注单变量线性方程。该数据集包含了一些问题集合，每个问题都涉及一个未知量的线性方程的构建和求解，使其成为培训和评估机器学习模型、教育软件和自动辅导系统的宝贵工具。通过提供结构化和多样化的方程集，SingleEq 数据集使研究人员能够探索从符号推理到基于神经网络的方法等各种计算方法，以提高自动方程求解技术的准确性和效率。其综合性确保了在此数据集上训练的模型能够处理广泛的方程类型，促进数学问题求解应用的稳健性和通用性。

An example of data from the SingleEq dataset

说明： Oceanside Bike Rental Shop charges 17 dollars plus 7 dollars an hour for renting a bike. Tom paid 80 dollars to rent a bike. How many hours did he pay to have the bike checked out ?

Let's assume that Tom had the bike checked out for x hours.

According to the problem, the cost of renting a bike is 17 dollars plus 7 dollars per hour. So, the total cost for x hours can be represented as:

$$17 + 7x$$

We know that Tom paid 80 dollars to rent the bike, so we can set up an equation:

$$17 + 7x = 80$$

Subtracting 17 from both sides, we get:

$$7x = 63$$

Dividing both sides by 7, we get:

$$x = 9$$

Therefore, Tom paid to have the bike checked out for 9 hours.

SVAMP。SVAMP (Patel et al., 2021) (简单算术单词问题变化) 数据集是一个精心策划的算术单词问题集合, 旨在评估和增强机器学习模型的问题解决能力。与传统数据集不同, SVAMP 通过在问题结构、措辞和复杂性上引入变化来测试模型在不同场景中的鲁棒性和泛化能力。每个问题的设计都需要结合数学推理和自然语言理解, 使其成为评估 AI 系统在实际应用中表现的有价值的基准。通过融入多种类型的问题, SVAMP 旨在填补简单算术任务与更复杂、充满上下文挑战之间的空白, 提供一个全面的工具来推动数学推理和自然语言处理研究的发展。

An example of data from the SVAMP dataset

说明： Matthew gave equal numbers of crackers and cakes to his 4 friends. If he had 32 crackers and 98 cakes initially. How many crackers did each person eat?

To solve the problem, we need to divide the total number of crackers by the number of friends.

Step 1: Find the total number of crackers by adding them up.

32 crackers

Step 2: Divide the total number of crackers by the number of friends.

$$32 \text{ crackers} \div 4 \text{ friends} = 8 \text{ crackers per person}$$

Therefore, each person ate 8 crackers.

Answer: 8