Turning Dust into Gold: Distilling Complex Reasoning Capabilities from LLMs by Leveraging Negative Data

太长不看版

LLMs的推理能力在实践中不够用，所以做了一个model specialization框架。

之前都是做CoT，但是复杂数学问题又不适应小的模型，为了让复杂的数学推理能在小模型中大规模地应用，利用上了负面的数据样本。

先通过测试验证了负面数据有一定的价值，又开发和运用了一系列的方法让这个框架更高效。这些方法是根据老一套的model specialization来开发的，主要有NAT（使从负面数据中吸收知识更加全面），NCE（让自提取的过程更集中于重要知识）和排名模型（可以给答案集合分配合适的权重）。

接着文章花了大量的笔墨介绍了这些方法的原理。包括CoT、SE、SC、NAT、ANT、DIU、NCE、ASC、RMT、WP。

实验部分通过实证数据证明了本文使用的NAT、NCE、ASC都取得了显著的进步。

摘要

LLMs在各种推理任务上表现不错，但是它们的难以理解性（*accessibility*）和众多的参数阻碍了在实践中的广泛应用。

一个可能的解决方案是通过生成的思维链（*CoT，chain-of-thought*）推理路径来从LLMs中提取推理能力给小的模型应用。但在某些情况下（特别是面对复杂的数学问题时）LLMs会生成错误的推理链。

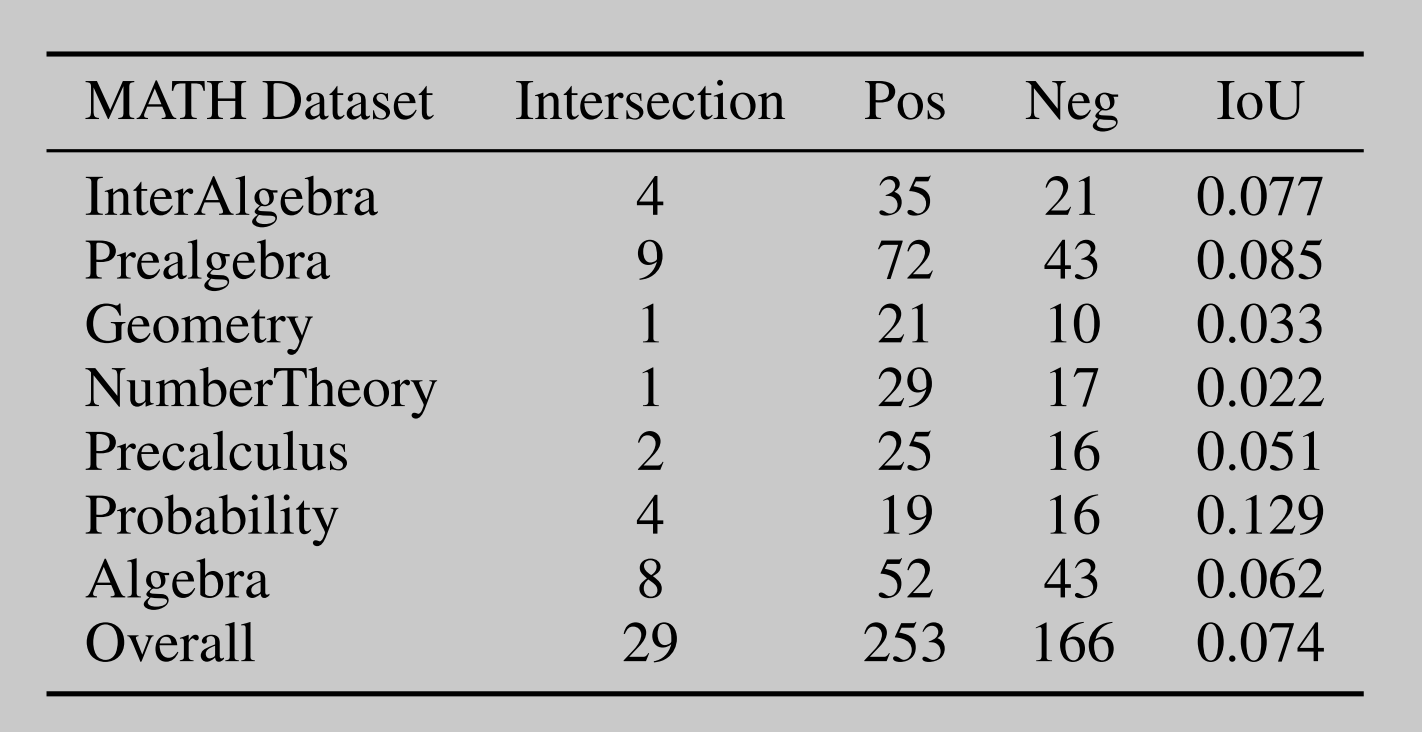
先前的研究只采纳（*transfer*）了正面（*positive*）样本的知识并且删除了有着错误答案的合成数据。在本文研究中证明了负面数据的优势并提出了一个模型专门化框架（*model specialization framework*）来用负面样本提取LLMs（中的能力）。

此框架由三部分组成，包含从训练到推理阶段，再到从负面数据中吸收知识。本文在算术推理任务方面实施了大量实验来演示负面数据在从LLM中提取的角色。

引言

由于CoT，LLMs展现出了出色的推理能力，特别是在复杂数学问题上。但是CoT是有多于100B参数的模型的新兴产物，不适用于更小的模型。在资源有限的情况下，这些模型的繁重的计算需求和大量的推理消耗阻碍了它们在现实应用情况中的发展。因此，本文研究的目标就是让复杂的数学推理能在小模型中大规模地应用。

要把特殊的能力从LLMs转移到更小模型上，知识提取（*Knowledge distillation*）提供了可能的方法。这个过程也被称为模型特殊化（*model specialization*），强制简练的（*compact*）模型专注于特定的技能。先前的研究把上下文学习（*ICL，in-context learning*）应用至LLMs来生成数学问题的推理路径，这对小模型获得复杂推理能力的好处大于对小模型引出推理路径的好处。



表格1：在正负面的数据上分别训练模型，

在他们在数学测试组上的正确答案中有着极小的重叠部分。

尽管负面的模型准确率更低，它也能够解决一些正面模型不能答对的问题，这也证实了负面数据有着有价值的知识。此外，在阻止模型犯相同错误的时候，这些原本不想要的带有负面数据的行为也是有用的。另一个我们应当利用负面数据的缘由是OpenAI基于token（来决定其价值）的定价策略（*token-based pricing strategy of OpenAI*）。即使是GPT-4，在数学数据集上的准确率也小于50%，这也就意味着负面数据的所有token都是免费的（因为依据策略判定其无价值）。因此，与其弃置那些负面的样本，我们不如好好利用负面样本中有价值的知识来增强模型特殊化。

老一套的模型特殊化流程可以分为三步：

1. 提取思维链，用LLMs生成的推理链训练小模型
2. 即self-enhancement（自增强），实施self-distillation（自提取）或者self-augmentation（自增长）来进一步优化模型。
3. 此外，self-consistency（自洽性）在推理任务中被广泛用作增强模型表现的解码战略。

此次工作中，提出了一种虚构的模型特殊化框架（如表一）来利用负面数据提升从LLMs中提取复杂推理能力。特别地，本文开发了负面协助训练（*NAT, the negative assistant training*），使得双LoRA结构能从正面和负面两边获取知识。作为一个辅助模块，负面LoRA的知识可以通过相关的注意力机构（*attention mechanism*）在正面LoRA中被动态调整。

对于自增强，本文设计了负定标增强（*NCE，negative calibrated enhancement*）将负面输出看作基准（*baseline*）来增强临界正理性（*critical positive rationales*）的提取。

除了训练阶段外，本文还在推理中充分利用了负面的信息。传统的自洽性给所有参与的输出分配相同或者基于概率的权重，导致了一些不可靠的答案被投票赞成（*voted up*）。为了缓和这个问题，提出了适应自洽性（*ASC，adaptive self-consistency*）在投票前排名，而且这个排名的模型是被正负面的数据训练的。

本文将LLaMA-7b作为学习模型实施全面的实验和精细的分析。先前的模型特殊化只在原先的数据集上生效，本文是第一次专注于挑战性的数学竞赛问题——MATH dataset。

实验显示：

1. NAT使从负面数据中吸收知识更加全面。
2. NCE让自提取的过程更集中于重要知识。
3. 用正负面原理(*rationales*)训练的排名模型可以给答案集合分配合适的权重。

总之本文贡献如下：

1. 本文证明，在复杂数学推理任务中从LLMs提取知识时，带有不正确答案的负面样本也能独立（*besides positive data*）提供参考价值。
2. 为了充分利用负面数据，本文提出一种由三个步骤组成的模型特殊化框架，包括训练一直到推理阶段。
3. 高难度数学推理数据集上的大量评估证明该框架可以高效利用负面信息并且比基准表现好很多。

方法

**思维链提取**

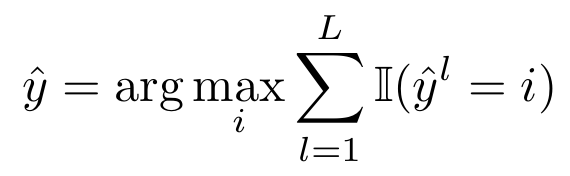
Rajani等人证明在数据集上用清晰的rationales训练语言模型可以提高生成最终答案的能力。因此提出思维链提取来最大化LLMs在更小模型上推理能力的表现形式。

**自增强**

基于人类自我反思以实现进步的思想，人们提出了各种方法来根据自身知识强化语言模型，我们统称为自增强。它包括两种常见的方法：一种是自增长（*self-augmentation*），即模型首先生成具有多样性的数据，然后在这些数据上进行训练，以实现更好的泛化。另一种是自提取（*self-distrillation*），即利用模型本身作为老师，完成迭代提炼，从而利用dark knowledge进一步提高性能。

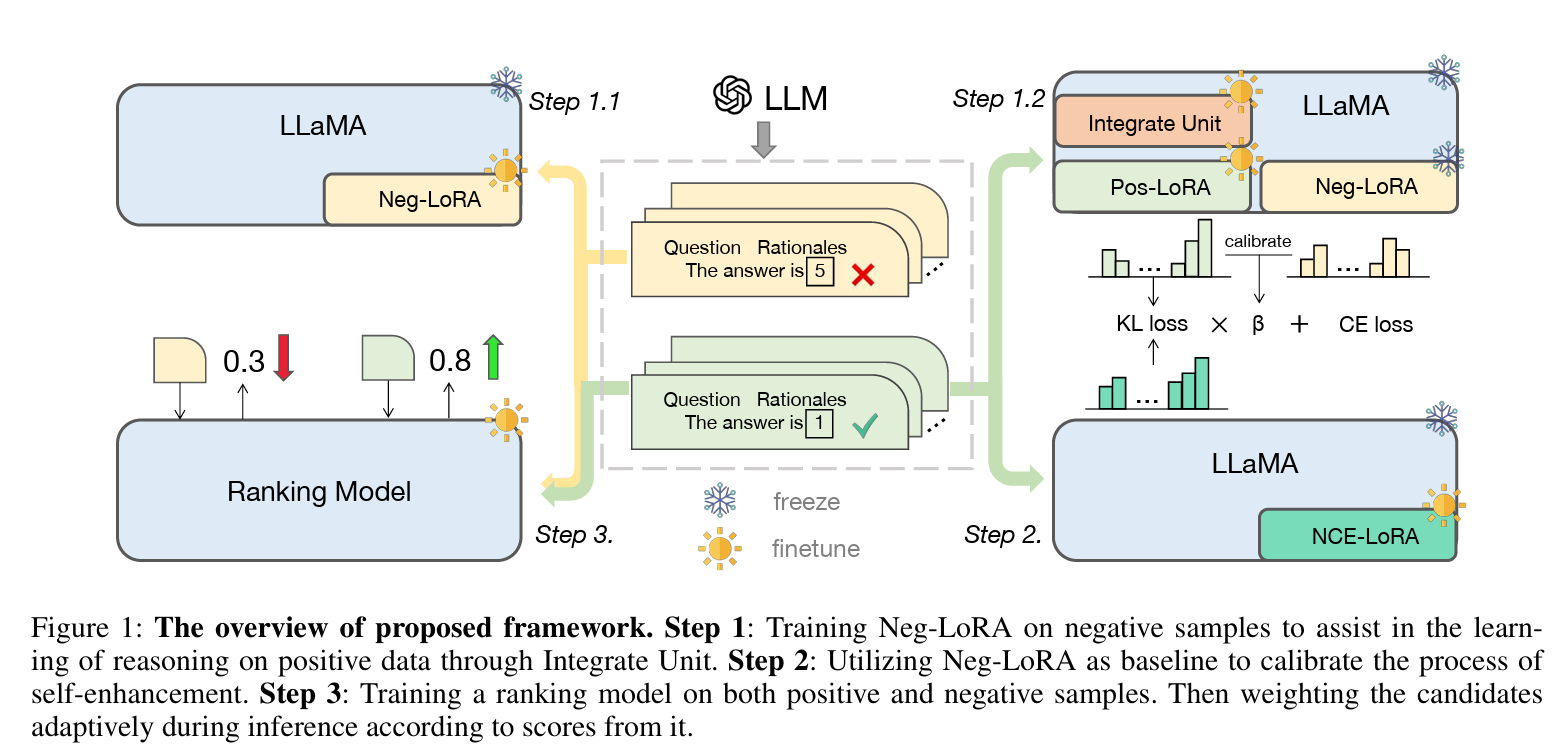
**自洽性**

自洽性充分利用了，一个需要逻辑思考的复杂问题，通常给出不同方法达到相同准确结果，的概念。



**Negative Assistant Training (NAT)**

如表1所示，负面的样本也包含有价值的知识，甚至能作为正面数据的好的补充。然而，≠的情况下，对于与负数据相对应的，推断错误的风险会增加。因此从没有被不想要的行为影响的负面样本中提炼知识是一项有风险的任务。为解决这个问题，本文提出两个阶段的NAT程序。如图一。



第 1 步： 在负面样本上训练 Neg-LoRA，以协助通过集成单元的正面数据的推理的学习。

第 2 步： 利用 Neg-LoRA 作为基线，校准自我增强过程。

第 3 步： 在正面和负面样本上训练排序模型。然后在推理过程中根据其得分对参与者进行自适应加权。

**吸收负面知识（*Absorbing Negative Knowledge*）**

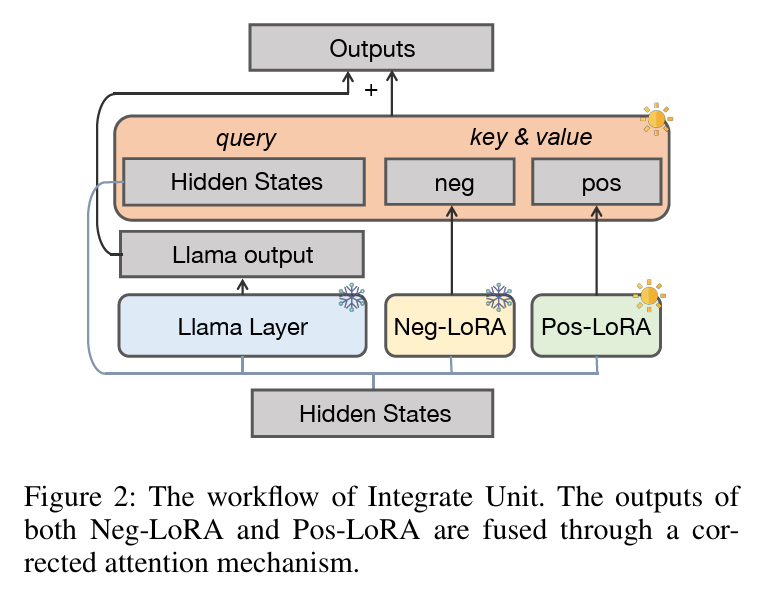
首先，在数学问题上从LLMs中获得（, , ）三个一组。为了推动各种从LLMs中的问题解决方法的全面学习，给每个问题收集八个不同的回答，然后根据是否与相等来把这些样本分为和。如果直接微调在和的集合体上的LLaMA必然引起不想要的行为结果到模型中去，因此考虑先在上训练一个负模型，并随后提取有用的知识。本文通过最大化接下来这个期望值，选择了LoRA模组给它的参数有效特征来微调在上的LLaMA

在此过程中，LLaMA的参数保持冻结，而的知识正在被LoRA吸收。

（Denote LLaMA with as 。）

**动态集成单元**

因为不能预先确定擅长哪个数学问题，本文设计了如图二所示的动态集成单元来在的正面知识的学习过程中，动态地集成从中习得的知识。



冻结以阻止其中的知识被遗忘，此外引入正面LoRA模组。在LLaMA的每一个层级，记从中获得的输出值为给输入的隐状态（*Hidden States*）。理想状态下，如果包含有益的知识，应当正合并和来完善中有关的有益知识。（反之亦然）

）

使用上NAT，能够在多样性（更多样本）和类型（正负面数据）的维度更全面地利用LLMs的知识，来提高复杂推理能力。

**NCE（*Negative Calibrated Enhancement*）**

为了进一步增强模型的推理能力，提出了能加上自增强过程来利用负面知识的NCE。本文首次用来生成对给在中的每个问题作为增强样本并把它们加入。对于self-distillation，本文发现一些样本可能包含更多可以区分带有不同能力的缘由的理性推理步骤。本文初步的目标是识别这些关键的rationales并且在self-distillation过程中增强它们的学习。考虑到已经包含从中汲取的有价值的知识，让其更进一步的关键在它们不一致的rationale之间是肯定的，并在其之间产生分配。因此本文采用KL差异（*KL divergence*）来测量这种不一致并最大化以下的期望。

(记学生模型的LoRA模块)

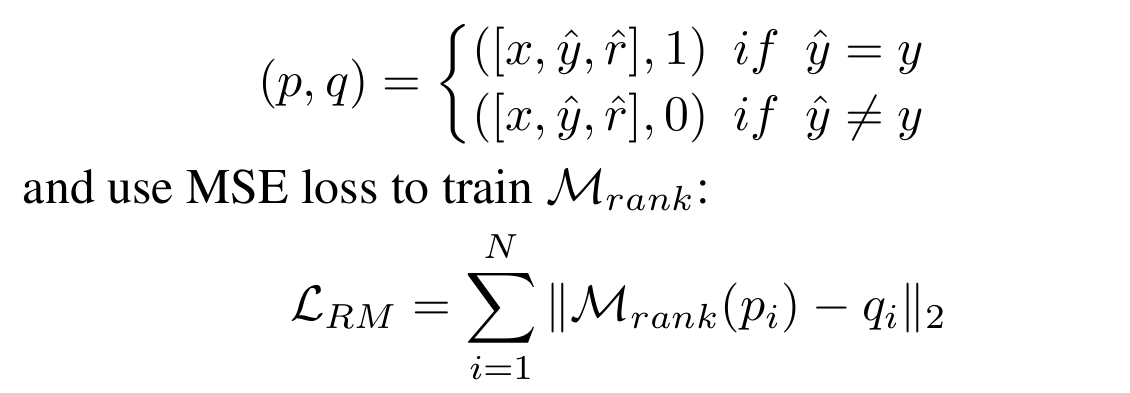
**适应的自洽性（*ASC，Adaptive Self-Consistency*）**

SC技术能进一步提高模型在复杂推理方面的表现。然而现行的方法要么给每个参与量赋予相等的权重，要么基于生成概率赋予权重。这些方法不能在投票阶段根据的数量调整参与量的权重，可能会遮蔽那些正确的参与量。于是提出了ASC利用和训练可以带有判断地，有适应性地再分配权重给参与量的排名模型。

排名模型训练

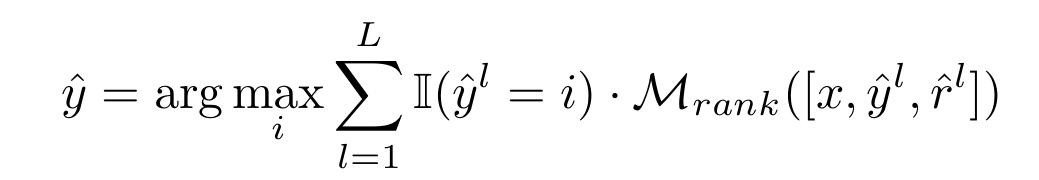
理想状态下，我们希望能给引向正确答案的rationales更高的权重，反之亦然。

于是通过以下方法创建训练样本



**权重政策（*Weighting Policy*）**

建立起的基础之后，本文修订了自洽性中的E。



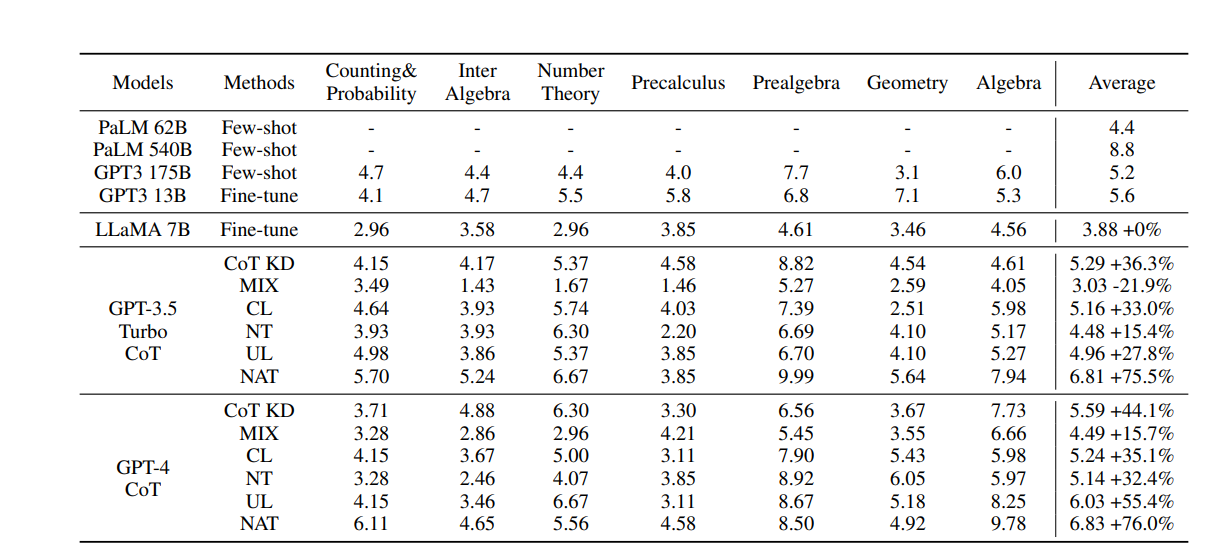
实验部分

对于教师模型，使用 Open AI 的 gpt-3.5-turbo 和 gpt-4 API来生成推理链。对于学生模型，选择 LLaMA-7b。

作者团队研究中有两种基线：一种为大语言模型（*LLMs*），另一种则基于 LLaMA-7b。对于 LLMs，将其与GPT3 和 PaLM进行比较。

对于 LLaMA-7b，将文中方法与三种设置进行比较：Few-shot、Fine-tune（在原始训练样本上）、CoT KD（思维链蒸馏）。在从负向角度学习方面，包括四种基线方法：MIX（直接用正向和负向数据的混合物训练 LLaMA）、CL（对比学习）、NT（负训练）和 UL（非似然损失）

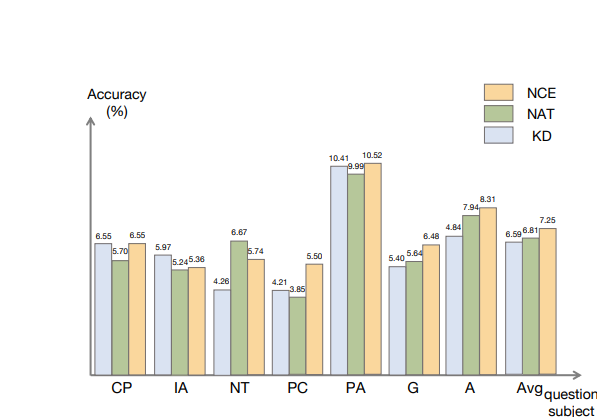
**NAT实验结果**



可以发现，在困难的数学数据集上，NAT表现突出。在原始数据上进行微调相比，NAT 在两种不同的 CoT 来源下实现了约 75.75% 的提升。与 CoT KD 在正样本上的比较，NAT 也显著提高了准确性，体现了负样本的价值。

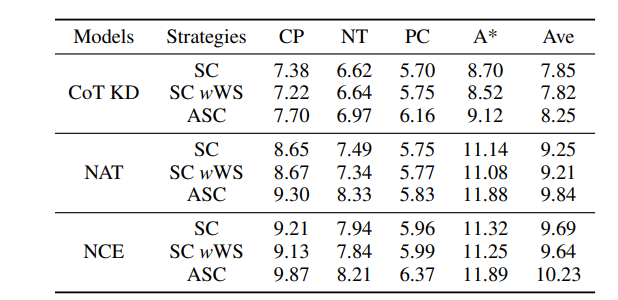
对于利用负向信息基线，MIX 的低性能表明直接训练负样本会使模型效果很差。其他方法也大多不如 NAT，这表明在复杂推理任务中仅在负方向上使用负样本是不够的。

**NCE实验结果**



与知识蒸馏（*KD*）相比，NCE 实现了平均 10%（0.66） 的进步，这证明了利用负样本提供的校准信息进行蒸馏的有效性。与 NAT 相比，尽管 NCE 减少了一些参数，但它依然有 6.5% 的进步，实现压缩模型并提高性能的目的。

**ASC 实验结果**



A\* 这一栏是 InterAlgebra, Prealgebra 和Algebra三个方向的平均值.

为了评估 ASC，将其与基础 SC 和 加权（WS）SC 进行比较，使用采样温度 T = 1 生成了 16 个样本。如图所示，结果表明，ASC 从不同样本聚合答案，是一种更有前景的策略。

总结：

本文主要探讨利用负样本从大语言模型中提炼复杂推理能力，迁移到专业化小模型的有效性 。作者团队提出了一个全新框架，由三个步骤组成，并在模型专业化的整个过程中充分利用负向信息。负向协助训练（NAT）从两个角度提供利用负向信息的方法。负向校准增强（NCE）校准自蒸馏过程，使其针对性地掌握关键知识。而基于两种观点训练的排序模型为答案聚合分配更适当的权重，以实现动态自洽性（ASC） 。