2024年11月29日 星期五
编辑/吴 琼 投稿邮箱：kxjy@kepu.gov.cn
科普时报

数学推理，大语言模型真的会了吗

□ 张立英

逻辑放大镜

当下，大语言模型似乎具备了与人交流对话的能力，不仅如此，大语言模型还能飞速完成很多人类需要花更多时间才能完成的文字生成任务，比如，写篇总结文章、作首诗、写段Rap。然而，大语言模型的推理能力却引起了很多争议，近两年来的很多测试显示，大模型在计数、符号推理、算术推理、子集求和、几何推理等方面的表现都不理想。

反复“刷题”，或导致数据污染

为了提高大模型的推理能力，Open AI发布了一个名为GSM8K的数据集，这个由人类写手创造的数据集包含了8000多个小学数学问题和答案，其中有7473个训练问题和1319个测试问题。对于人类而言，这些问题只需用到简单的加、减、乘、除运算，通过2-8个步骤，就可以得出最终答案。

经过不断地训练和调整，大语言模型在面对GSM8K时，性能已经有了显著提高。但这是否真的意味着大模型的数学推理能力变强了？一种质疑是，由于这个数据集的题目固定且被拿来反复使用，很可能出现数据污染——数据完整性的破坏。所以，即使测试结果变得更好了，也不能确认这些大语言模型的数学推理能力真的提高了。

微调题库，测试应变能力

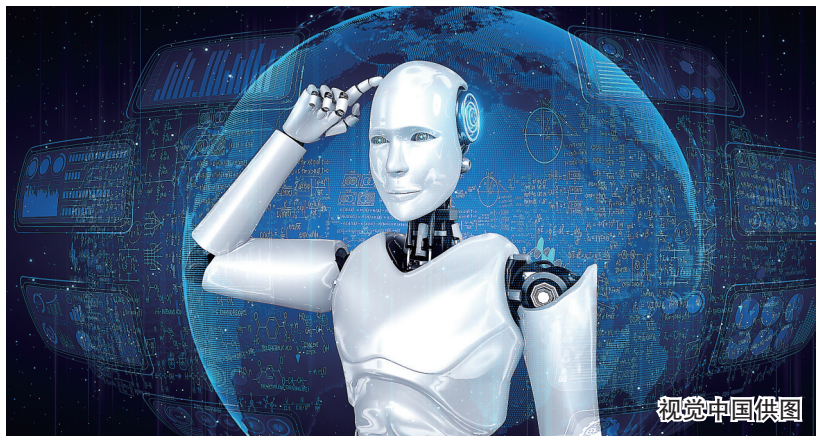
今年10月，苹果公司发布的一项测评证实了这一质疑的合理性。为避开GSM8K可能导致的数据污染，苹果公司的研究人员想出了一个好办法，他们给出了一个在GSM8K基础上进行微调的测试系统GSM-Symbolic。微调方式主要有3种：替换题目中的专有名词；改变其中的数字；添加无关信息。

举个例子，假设原题库中的题目是这样的：

小明周五钓了6条鱼，周六钓了15条鱼，周日钓到的鱼是周五的2倍，问小明总共收获了多少条鱼？

GSM-Symbolic对这道题采取以下3种方式进行微调：或是把原题中的小明换成小丽；或是把原题中6换成9，15换成23；或是增加一些无关信息，比如增加条件“周日钓到的鱼中，有5条鱼的重量低于平均值”。当然，还可能把这几种微调综合在一起。基于这些微调，从GSM8K数据集中的题目出发，GSM-Symbolic可以千变万化出更多题目来对大语言模型进行评估。

从人类的视角来看，这些微调策略就是我们常说的“换汤不换药”，做过小学数学题的读者们再熟悉不过了。所谓“不换药”，是说微调完全没有涉及这些数学问题的逻辑结构，只是调整了一些无关参数。



正确率大幅下滑

但正是这样的微调，却造成了大语言模型输出答案正确率的大幅下滑。其中，无关信息的添加会导致所有最先进的大语言模型的性能大幅下降，降幅高达65%。

苹果公司的研究人员基于这些测评得出结论：大语言模型既不理解这些问题中的数学概念，也不能进行逻辑推理，而仅仅是将面对的问题和训练数据中的问题进行比较而已。

目前来看，大语言模型所得出的正确答案，主要体现了系统的记忆和匹配能力，这种应答机制更像一种模式匹配，这与人类推理的机制完全不同，也没有遵循逻辑。

人类才懂“万变不离其宗”

那么，人类在做小学数学推理题时，究竟启用了哪些隐藏技能？

笔者理解至少有两条，一是透过现象看本质的能力：人类能够抓取或识别表层语言背后的一般性运算和推理的规律；二是由内及外、活学活用的能力：人类能够通过非关键因素（比如前面提到的3种微调因素）的替换和变化展开千变万化的实际应用。

这两条综合起来，就是我们常说的“万变不离其宗”。

（作者系中国科学院哲学研究所教授）

三个臭皮匠，可以赛过诸葛亮吗

□ 池红梅

数学文化

“三个臭皮匠，赛过诸葛亮”，这是我国的一句俗语，典故出自《三国演义》。

诸葛亮为了实施“草船借箭”之计，命令三名部下在草船上插草靶子，然后用布掩盖伪装成士兵。但部下觉得如此布置不够妥善，会露出破绽，他们将每个船头立上稻草人，并套上皮衣皮帽。后来，曹军果然中计。

这句话的意思是指三个普通人的智慧联合起来，就有可能超过一个非常聪明的人。它强调了团队合作的重要性，即使个体的能力有限，但通过合作，可以发挥出更大的力量。诸葛亮是我国历史上著名的智者和战略家，在这里作为杰出的智慧化身。

那么问题来了，三个能力一般的人合作，真的可以超越一个能力出众的人吗？如果你就是那三个“臭皮匠”之一，怎样想办法超越“诸葛亮”呢？

让我们来看一个最简单的数学模型：假设有三个能力一般的人，记为甲、乙、丙，他们各自独立解决某一难题的概率分别是0.3、0.4、0.5；另有一位能力出众的人，记为丁，他解决这一难题的概率是0.8。现在甲乙丙三人组队，为简化模型，他们还是各自独立解决问题，三人都不能将问题解决的概率是$(1-0.3)\times(1-0.4)\times(1-0.5)=0.21$。因此，这个队伍中至少有一人将难题解决的概率是$1-0.21=0.79$。

视觉中国供图

神奇的结果出现了，三人组队后，至少有一人将难题解决的概率已经很接近0.8了。

为了能够超越0.8，我们可以尝试如下两种办法：第一种，尝试换一个能力稍强一些的队友，比如将甲换成独自解决这一难题概率是0.4的人，那么最后的结果会变为$1-(1-0.4)\times(1-0.4)\times(1-0.5)=0.82$；第二种办法，让更多的队友加入，比如再加入一个解决问题的能力为0.3的队友，那团队解决问题的概率将变为$1-(1-0.3)\times(1-0.4)\times(1-0.5)\times(1-0.3)=0.853$。大家可以尝试一下，如果队伍壮大成10人甚至100人，解决问题的概率远远高于0.8，甚至接近1。

以上是简化的数学模型，现实中的“臭皮匠”各自拥有不同的技能和知识，合作往往能够激发新的创意，也会增加解决问题的途径和方法，从而提高解决问题的概率。

“三个臭皮匠，赛过诸葛亮”强调了合作、多样性和累积效应在解决问题中的重要性，正所谓众人拾柴火焰高，在通往成功的道路上，大家可以充分相信团队的力量。

（作者系华中农业大学信息学院大学数学教学学术团队骨干教师）

先睹为快

领略世界“第八大陆”风采

提起“马达加斯加”，或许你立刻会想到可爱的狐猴，自然爱好者还会想到各种独特的两栖动物和爬行动物、昆虫、奇花异木……

作为世界第四大岛，马达加斯加不但有多种复杂生态，动植物种类也是独一无二，故而号称“世界第八大陆”。可惜由于全球气候变化等原因，马达加斯加岛如今正深陷生态危机，岛上的众多特有物种也濒临灭绝。

让我们跟着2024年第11期《博物》杂志一起，领略这座神奇的马达加斯加岛的风采。