一个推理测试“团灭”各大语言模型

□ 张立英

逻辑放大镜

近日，德国某非营利人工智能研究学机构的几位研究者发表了一项研究成果，揭示了当下各大语言模型在推理能力上的短板。他们设计了一系列简单的推理问题，用来测试大语言模型的推理能力，结果GPT-4、Claude、Gemini、Llama、Mistral等模型几乎全线崩塌。但是，这些大语言模型仍然展现出“迷之自信”，宣称自己的“思考过程”非常符合逻辑。

究竟是什么样的简单推理测试，难倒了这些大语言模型？

被研究者们命名为爱丽丝漫游奇境（AIW）的推理测试可概括如下：

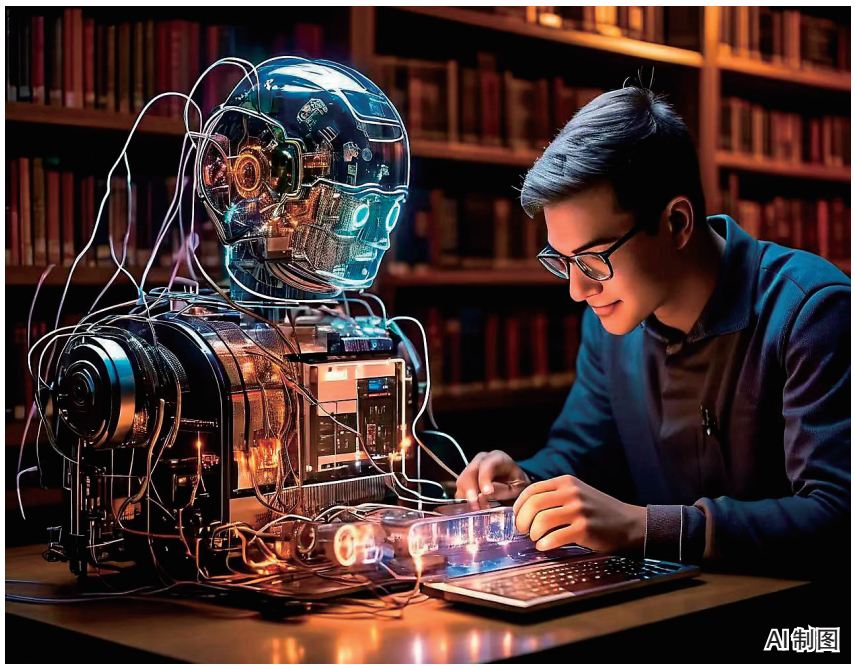
爱丽丝有N个兄弟，她还有M个姐妹。爱丽丝的兄弟有多少个姐妹？

从人类视角来看，这个问题其实并不复杂，也许稍加思考就可以得出结论：答案是M+1，即爱丽丝的姐妹数量再加上爱丽丝自己。

不过，如果拿出逻辑放大镜来仔细观察，就会发现：这个推理暗含了一些人们默认的背景知识，如果没有这些背景知识为前提，推理是无法完成的。

完成这个推理需要的背景知识有：

第一，爱丽丝是个女孩；第二，爱丽丝的姐妹都是女孩；第三，爱丽丝的兄弟都是男孩；第四，男孩不是女孩；第五，就爱丽丝和他的兄弟姐妹这个讨论范围中的每个人而言，每个女孩都是（除这位女孩自己以外的）其他人的姐妹。



有了这些背景知识，就可以比较顺畅地进行推理了。基于三、四、五，运用排除法和整体推出部分原理，可以得出第六点——爱丽丝和她的兄弟姐妹这个讨论范围中，所有女孩都是爱丽丝兄弟的姐妹。再基于一、二、五、六和题干条件，运用简单的加法运算，就可以得出M+1这个结论了。其实，如果真把这个推理过程完整写出来，需要一些步骤，但是就原理上说，这个推理不算复杂，成人和一定年龄以上的中小学生都可以完成。

那么，人们觉得很简单的推理，为什么大语言模型集体“不会”了？

这要从大语言模型的原理说起。当我们向大语言模型提问时，它们所生成的回答其实是一个字符、一个字符“蹦”出来的，每“蹦”出一个字符之前，模型要进行一番概率运算，看看语料库中哪些字符和前面已经生成的字符关联度大，然后从中做出选择。而之所以大语言模型的回答像“人话”，与这些模型做计算依赖的、海量的、由人们的真实会话所构成的语料库息息相关。

要通过爱丽丝漫游奇境推理测试，需要很多对人类而言非常基础的背景知识，以及从测试问题关联到这些知识的能力。可能恰恰由于这些知识和关联过程对人类而言过于“简单自然”，反而很少有人专门去谈论这些话题，所以相关的内容很少出现在语料库中。没了相关语料的喂养，大语言模型自然就“不会”了。

大语言模型备受关注的原因，就在于出色的说“人话”能力。然而，要想与人类真正对话，逻辑推理能力可不能差。除了让科学家们多想想办法之外，也许我们每个人也可以从自己做起，往语料库中多加点逻辑推理的养分，也能让大语言模型变得更“聪明”。

（作者系中国科学院哲学研究所教授）

不必全盘接受批评

□ 曹大刚

大耳叔叔：

您好！

我每天实在是太累了，好像一直活在批评的声音中。早自习一开始，班主任就开始发射“高音炮弹”：“你看看你们东倒西歪的坐姿！上课铃声响了还在聊天，你们这个劲儿怎么不用在学习上呢？你现在不努力，就是对自己未来不负责任，对不起自己父母的期待……”

回到家，爸爸看到我把校服扔在沙发上，又是劈头盖脸一顿数落，说我邋遢、懒，甚至说我是不好好学习浪费钱的“败家子”，说到激动时还要动手打我。姐姐说我头发油乎乎的很恶心，可我是油性皮肤，即使每天都洗头，到了晚上依然会出很多油……唉！我感觉整天都生活在批评中，自己真是个没用的人……

萌芽（化名）

萌芽同学：

你好！感谢你的信任，你的名字“萌芽”很有寓意，很有生机，渴望在孕育中寻找能量。人们回应批评时，要么是默默忍受要么是怼回去，你似乎更贴近前者。

其实，每个人或多或少都会面对别人的批评。客观的批评声音可以帮助我们发现自身存在的问题；但如果批评得过于苛刻与频繁，就可能转化出现病态的自我批评，对自己的心理和情绪产生负面的影响，形成自己对内攻击。

我们应该尝试减少他人的批评、自我批评带给我们的负面影响。面对批评，如果奋力地怼回去会让人越来越偏激，容易伤害自己或他人；毫无反抗的忍受或让人增加自我评判的强度，导致自尊心下降，无助感增强。当批评发生时，我们需要理性对待。

要客观对待批评，对方不一定说得都对。我们首先要分清“事实”和“观点”，比如你把衣服扔在沙发上是事实，说你邋遢和懒就是你爸爸的观点。我们不能否认事实，但也不必完全以别人的观点来评判自己。其次，批评的重点是要落在发生的事情上，所以我们要客观地区分“行为”和“人”，要学会区分这些声音是评价行为，而不是否定我这个人。

叔叔悄悄告诉你，适当的自我批评具有安慰作用。如果过度自我批评、自我否定，就会产生习得无助性行为。当面对自我批评，我们可以尝试将“我”与这个事件的分开，将“个人”与“想法”分离，从而削弱批评的威力。你需要厘清自己只是某个事情没做好而不是你这个人完全不行，避免陷入全盘否定声音中而无法自拔，这也是我们逐步迈入社会化的一项必修课程。

萌芽同学，放平心态，一切都会好起来！

大耳叔叔

案例反思

中学生的自我社会化意识在不断加强，他们很在意别人的评价。无论在家庭还是在学校中，他们会受到外在的各种批评，这些批评会转化为自我负性批评，造成自尊心下降，容易产生焦虑、抑郁等消极情绪。学校和家庭教育中，要教会他们理性应对批评，正确评估自己，这有助于拥有良好的自尊水平。

（作者系中国科普作家协会会员、高级心理学教师、国家卫生健康委心理治疗师）

青春的路上一个人独自行走，是否有很多心思无人倾诉，很多想法无人理解？那就给大耳叔叔写信3548004514@qq.com，我愿意成为你的朋友！

先睹为快

雨林下的“吸血鬼”

东南亚婆罗洲雨林里，阳光穿过枝叶，在大树脚边洒下点点光斑。厚厚的落叶堆中，隐藏着一些神秘角色：暗红硕大、带有斑纹的大花草，好像一朵从高处掉落的花，无害地躺在树根旁。蛇菰（gū）和帽蕊草，则会被认成枯叶中钻出的菌类。事实上，它们都不是“省油的灯”……

2024年第6期《博物》杂志的主角就是这样一群“异类”植物，它们的长相和习性都十分富有侵略性。让我们跟随本期杂志，一起改变植物“人畜无害”的印象吧。