13.11 > 13.8

大模型“翻车”小学题？

专家：复杂推理的关键是要构造过程性内容

13.11和13.8谁大？——这样一道简单的“小升初”数学题难倒一众网友和大模型。

上周五（7月12日）综艺《歌手2024》投票率出炉，孙楠以13.8%的投票率位居第三，Chanté Moore以13.11%的投票率排在第四。一位网友留言质疑“13.8%比13.11%高？”迅速引起热议，随后“13.8 13.11”话题登上微博热搜。

围绕这样一道简单数学题，记者提问了9家国产大模型以及全球大模型标杆ChatGPT3.5。结果显示，国产大模型中6家答对3家答错，就连ChatGPT3.5也差点“翻车”。

13.11和13.8谁大？9家大模型，3家已翻车！

7月16日下午，记者实测阿里通义千问、百度文心一言、智谱清言、腾讯元宝、科大讯飞星火、字节跳动豆包、学而思九章大模型、面壁智能Luca、月之暗面Kimi等9家国产大模型。

当记者输入“13.11和13.8，哪个数值更大？”时，Luca、Kimi、智谱清言均错误地表示13.11更大。其余6家大模型给出了正确答案，认为13.8更大。

当记者进一步追问时：“你确定13.11大于13.8吗？”Kimi回复：“是的，我确定。”然后一本正经地解释：“13.11与13.8的整数部分相同，由于小数部分0.11是两位小数，而0.8只有一位小数，13.11比13.8更接近14，因此13.11是更大的数。”

智谱清言是通过比较小数部分得出错误结论，称“13.11的百分位是1，而13.80（即13.8）的百分位是0。在这种情况下，1大于0。”“简而言之，13.11实际上等于13.80＋0.01＋0.01，这使得它比13.80（即13.8）大”。

Luca则错误地回复称：“因为13.11比13.8多了0.31，所以13.11的数值更大。”

回答正确的6家国产大模型，都在小数的比较上做出正确判断。例如，通义千问表示：“对于13.11而言，十分位是1；而对于13.8而言，十分位是8。由于8大于1，因此可以立即得出结论，13.8比13.11大。所以，13.8是这两个数值中更大的一个。”文心一言还进行了一番简单运算：“0.8是8个0.1，而0.11只是1个0.1加上1个0.01。显然，8个0.1比1个0.1加上1个0.01要大。”

当记者用新的账号重启对话，并尝试变换语序问答：“13.8和13.11，哪个数值更大？”Luca、Kimi、智谱清言等三家大模型也未反应过来，坚持认定13.11更大。

9.11和9.9谁大？12家大模型，8家都答错！

此前，艾伦研究机构（Allen Institute）成员林禹臣在社交平台发帖，显示ChatGPT－4o在回答中认为13.11比13.8更大。

随后Scale AI的提示工程师莱利·古德赛德（Riley Goodside）基于此灵感变换了问法，拷问了大模型ChatGPT－4o、谷歌Gemini Advanced以及Claude 3.5 Sonnet——9.11和9.9哪个更大？几家主流大模型通通答错。

随后，有记者拿“9.11和9.9哪个大？”的问题一一测试了ChatGPT以及目前国内的主流大模型，包括阿里、百度等5家大厂模型，月之暗面等6家AI独角兽的模型。阿里通义千问、百度文心一言、Minimax和腾讯元宝4家大模型答对，其他8家则答错。

大模型ChatGPT在被问到“9.11和9.9哪个大？”时回复称，小数点后面的数字“11大于9”，因此9.11大。记者追问ChatGPT有没有其他比较方法，它将小数转化成分数比较，得出“11／100比90／100小”，这一步是对的，但它接着下结论称“因此9.11比9.9大”。有人提出，大模型回答错误可能是语境问题，比如从软件版本迭代的语境来说，9.11可能就会比9.9版本更大。因此记者加上限定词“从数学上”比较，ChatGPT仍然回答错误。

再看国内的大模型，询问kimi，它在比较小数部分时认为，9.11的第一位小数是1，而9.9的第一位小数是0，错误地给出了小数，得出结论9.11更大。当记者质疑并提出常识后，kimi转而开始表示自己回答有误，并给出了正确的比较方法。

询问字节豆包，它不仅给出了答案，还举了生活中的例子方便理解，单看似有理有据实则胡说八道。豆包举例认为，假设有两笔钱，“9.11元比9.9元多0.21元”，并且测量长度时“9.11米要比9.9米长”。

智谱清言在答题中，成功提到了9.11的十分位是1，而9.9的十分位是9，但仍然得出结论“9.11整体大于9.9”。并且还特意强调，“这个结果可能让人感到意外，因为直觉上可能会认为9.9更大，但根据数学规则，9.11确实是更大的数字”。在记者质疑答案后，智谱清言首先表示“您的理解是常见的误解”，随后自己推演了一遍后，得出了正确的答案，并承认自己之前的回答错误。

商汤商量大模型首先给出了错误答案，记者追问具体是如何比较的，它在推演过程中成功得出小数0.11小于0.9，但话锋一转称“所以9.11大于9.9”。记者指出了这个前后逻辑问题，商量随后承认“解释有误”。

阶跃星辰跃问同样给出了错误答案“9.11比9.9大”，错误地比较了小数点大小，记者进一步质疑，有趣的是，在解释中，跃问前后语言表达逻辑开始混乱，似乎没有意识到自己答案发生了变化。跃问在解释中首先称“理解你的困惑”，并表示日常生活中9.9确实比9.11大，但是在数学中“需要更精确地比较两个数的大小”，结果跃问随后推演得出结论称根据数学规则“9.11小于9.9”，丝毫没有提及自己之前回答错误。

还有两家大模型百川智能和零一万物，首先给出了错误答案，但在记者追问“为什么”的时候，就在推演后默默改变了答案。

算法工程师：

目前生成式语言模型更像一个文科生

为什么号称智能的大模型答不好小学生数学题？

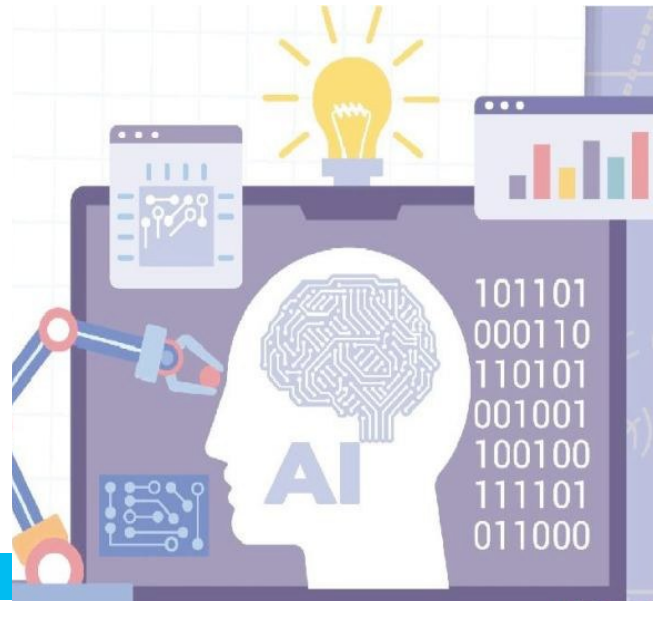
一些行业人士将数学不好的原因归结于LLM（大语言模型）的架构问题，大语言模型往往是通过预测下一个词的监督学习方式进行训练。简单来说，向大模型输入大规模的文本数据集，模型在训练学习后会根据当前输入的文本来预测下一个词的概率分布。通过不断比较模型预测和实际的下一个词，语言模型逐步掌握了语言规律，学会了预测并生成了下一个词。

一位算法工程师认为，生成式的语言模型更像文科生而不是理科生。实际上语言模型在这样的数据训练过程中学到的是相关性，使得AI在文字创作上达到人类平均水平，而数学推理更需要的是因果性，数学是高度抽象和逻辑驱动的，与语言模型处理的语言数据在本质上有所不同。这意味着大模型要学好数学，除了学习世界知识外，还应该有思维的训练，从而具备推理演绎能力。

此外，针对简单数学题出现的大模型集体错误，大部分行业人士都会第一时间想到Tokenizer（分词器）的数字切分问题。在大语言模型中，Tokenizer会将输入文本拆分转换成更小的部分（词元tokens）供模型处理。而Tokenizer并没有专门为数学设计，这导致数字在分割时可能被拆成不合理的部分，破坏了数字的整体性，使得模型难以理解和计算这些数字。

新浪微博新技术研发负责人张俊林对此解释道，早期LLM的Tokenizer一般不会对数字进行特殊处理，经常把连续的若干数字切在一起形成一个Token，比如“13579”，可能被切成3个Token，“13”是一个，“57”是一个，“9”是一个，哪些数字被切在一起组成Token，这取决于数据集合里的统计情况，在这种不确定哪些数字片段组成一个Token的情况下，LLM要想做多位数字数值计算，是非常困难的。

针对大模型复杂推理能力的短板，上海人工智能实验室领军科学家林达华此前在采访中对媒体记者表示，复杂推理的关键是要构造很多过程性的内容。例如，构造上亿条解几何题具体过程的数据，拿去给大模型训练后，模型就能逐渐学会解题过程。而从互联网上很难去大量获取这些数据，“未来在模型的训练数据上面，尤其是突破更高层次的智能的过程中，会越来越依赖构造型的数据，不是直接爬取下来的数据。”林达华认为。



据南方都市报、第一财经