

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 ChatGPT-3.5。结果显示，国产大模型中6家答对3家答错，就连 ChatGPT-3.5 也差点“翻车”。



大模型“翻车”小学题？

1 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。

2 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。

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。

3 算法工程师：目前生成式语言模型更像文科生

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。

(据《南方都市报》)