AI高考成绩出炉 ChatGPT分数最高

文科成绩超过97.55%考生，理科成绩不太理想

大模型参加高考，能考多少分，上什么大学？

近日，在极客公园最新发布的高考新课标Ⅰ卷大模型评测报告中，GPT-4o以562分排名文科总分第一。本次大模型高考评测与河南省考卷完全相同，河南高考录取分数线显示，文科本科一批录取分数线为521分，有3款国产AI成功冲上一本线。

与文科相比，大模型的理科成绩要差很多，最高分还不到480分，多数大模型的理科总分在400分以下。相比河南理科511分的一本线，大模型尚有较大差距。

另一场让AI进行高考的测试也引发关注。由上海人工智能实验室推出的司南评测体系OpenCompass，选取了零一万物、智谱AI等6个开源模型及GPT-4o进行高考“语数外”全卷能力测试。据悉，参与评测的所有开源模型开源时间均早于高考，以确保“闭卷”性，评测采用全国新课标Ⅰ卷，由具有高考评卷经验的教师人工评判，更加接近真实阅卷标准。

一个明显的趋势是，大部分模型“考生”出现了偏科现象，其中语文、英语科目表现良好，但在数学方面全军覆没，连及格分都拿不到。

2024年高考新课标Ⅰ卷全科目测试

大模型产品	数学	语文	英语	历史	地理	政治	物理	化学	生物	文科总分	理科总分
GPT-4o	66	120	139	81	68	88	51	42	51.5	562	469.5
字节豆包	61.5	125.5	131	82.5	62	80	42.5	49.5	56.5	542.5	466.5
文心4.0	62.5	119	137.5	78	61.5	79	54.5	40	65	537.5	478.5
百小应	44	128	139	72	55	83	24.5	47.5	56	521	439
通义千问	35	111	131.5	82	44	75	18	37	62	478.5	394.5
Kimi	39	100	127	72.5	58.5	65	32	34	41	462	373
腾讯元宝	39	120.5	118	73	39	70	27.5	36	47	459.5	388
MiniMax	38.5	104.5	127	67.5	53.5	63	39	36.5	46.5	454	392
智谱清言	37	102.5	134.5	60.5	39	64	13.5	24	50.5	437.5	362

注：默认所有大模型产品均能得到英语听力满分（30分）。
注2：根据教育考试院官网，2024年河南省高校招生文科和理科一本录取分数线分别为521分、511分

语文：无法理解潜台词 不会引用名人名言

语文、英语的语言类考试，是大模型有能力和人类考生较量的赛场，多家产品能拿到客观题目的满分或接近满分。

除了少数开放性的阅读理解和语言文字运用问题，各家大模型主要丢分在语文写作上。

作为评测的语文作文阅卷人，北京市级骨干教师、怀柔区语文学科带头人夏老师曾多次参加全国高考语文阅卷。

夏老师认为：“AI写出的文章大多有清晰完整的结构，有逻辑性，语言通顺流畅。但其理性有余，感性不足，缺乏感情色彩，自然就缺乏感染力。”

另一场OpenCompass评测的阅卷老师也提出，在语文这一科目上，大模型与人类考生相比，在答题时仍有差距。一是大模型的现代文阅读理解能力普遍较强，但是不同模型的文言文阅读理解能力差距较大；二是大模型作文更像问答题，虽然有针对性但缺乏修饰，几乎不存在人类考生都会使用举例论证、引用论证、名人名言和人物素材等手法。三是多数模型无法理解“本体”“喻体”“暗喻”等语文概念。语言中的一些“潜台词”，大模型尚无法完全理解。

英语：大模型英语作文 普遍超出字数被扣分

英语写作同样是大模型的一大难题。在阅读和语言运用两大项客观问题的考试上，GPT-4o、百小应、通义千问获得80分满分，豆包和文心一言4.0也接近满分。但是在40分的写作考试中，最高分只有29分，分别由GPT-4o和百小应获得，各家模型的英语写作主要丢分在表达空泛、缺少细节上。如果大模型在未来能够提升写作能力，获得高考满分并非难事。

而在OpenCompass评测中，英语科目上，各个大模型整体表现良好，但部分模型由于不适应题型，在七选五、完形填空等题型得分率较低，同时大模型英语作文普遍存在因超出字数限制而扣分的情况，而人类考生多因为字数不够扣分。

文综：得分较为出色

在由历史、地理、政治组成的新课标文综考卷评测中，GPT-4o获得237分的成绩，平均分达到79分，优于多数人类考生。

政治考试中，GPT-4o出人意料地获得了88分的最高分，百小应和豆包得分超过80。地理考卷则有大量图片问题，对一众大模型是不小的挑战，图像理解能力较强的GPT-4o得到最高分，但仅有68分。

河南高考分数段统计数据显示，GPT-4o的562分在文科考生中排名8811名，相当于人类考生的前2.45%。

数理：全线不及格

与人类顶尖考生相比，大模型在数学、物理、化学等数理学科上差距极大，包括GPT-4o在内的所有大模型都无法达到及格水平。尽管在语文、英语两科上能获得高分，大模型的理科最好成绩还无法进入人类考生的前30%。

以数学试卷为例，9款大模型产品中，仅GPT-4o、文心一言4.0和豆包获得60分以上成绩（满分150分），目前的大模型只能正确推理步骤相对简单的问题。据测试机构透露，比如豆包等大模型能准确运用求导公式和三角函数定理，但是面对较为复杂的推导和证明问题就很难继续得分。

在OpenCompass评测中，大模型“考生”数学同样不太行，全部不及格，数学科目各大“考生”平均得分率仅为36%（150分满分）。

阅卷老师分析称，此次参与大考的大模型在数学主观题回答上相对凌乱，且过程具有迷惑性，甚至出现过程错误但得到正确答案的情况。虽然大模型的公式记忆能力较强，但无法在解题过程中灵活运用。

针对大模型答数学题普遍“吃瘪”的问题，国内某头部大模型负责人就曾表示，大模型的指令遵循或者说推理能力通常是把一个指令背后的意思拆解出来，但数学题既包含规则性，又包含对各种思维的考察，解题逻辑和正常用大模型时的推理逻辑不一定完全一样。同时该负责人还提到，从更广泛的大模型应用角度来看，AI能不能精准遵循指令是近一段时间内比较重要的事情，真正的商业价值也比较大可能来自于此，而解数学题对目前的AI来说还是一件比较“炫技”的事情。

大模型在理综考试中的表现同样糟糕。在极客公园测评中，重点考查实验探究能力的化学和物理试卷，各模型平均分更是只有34分和39分（满分为100和110）。

大模型在应对考试的灵活性上也不如人类。例如物理有一道送分题，人类考生根据“时间不会倒流”可以排除错误选项，轻易选对正确答案“C”，大模型则几乎全军覆没。要学会像人类一样思考和解决问题，大模型还有很长的路要走。

另有业内人士向记者表示，目前来看大模型的数理能力相对较差的情况在中外都是一样的，“打个比方可以这样讲，大模型就是偏科，文科强理科弱，这个情况在一段时间内也不会得到明显的改善”。

该人士进一步提出，这种情况与文理科的语料数据情况、推理逻辑情况相关。“第一，文科的语料数据丰富多样，有利于训练大模型，而理科的语料主要是数字和符号，形式单一，数据资源少，不利于训练大模型。第二，文科与理科逻辑不同。文科推理预测，有一两处错误，不会影响长文本理解，但是理科一旦某个数字或符号推理错误，结果就是南辕北辙。”

据《南方都市报》