Artificial Intelligence and Robotics Research 人工智能与机器人研究, 2026, 15(1), 305-317
Published Online January 2026 in Hans. https://www.hanspub.org/journal/airr
https://doi.org/10.12677/airr.2026.151030

Hans汉斯

国产生成式人工智能解决物理问题能力研究

——以“智谱AI”、“讯飞星火认知大模型”、“天工”、“360智脑”、“文心一言”为例

庞付豪，陈美娜*，孙雨心

山东师范大学物理与电子科学学院，山东 济南

收稿日期：2025年11月29日；录用日期：2026年1月15日；发布日期：2026年1月27日

摘 要

ChatGPT是一款功能强大的预训练语言模型，自2022年发布以来，引发了人们的广泛关注。为紧跟人工智能的发展潮流，我国也相继出品了自己的生成式人工智能模型。为了检测国产模型解决实际物理问题的能力，本文选取了“智谱AI”、“讯飞星火”、“天工”、“360智脑”和“文心一言”等五大模型，以经典力学问题为例，分别测试了其概念理解、推理计算和实验设计能力。研究发现，上述五个模型在概念理解方面的解题能力最强，推理计算次之，实验设计最差，实际解题过程存在“计算失误”、“前后回答不一致”、“情境分析能力欠缺”等问题。横向比较：“天工”在概念理解方面的表现占优，而“文心一言”在推理计算方面的表现最好。总的来说，我国国产模型实现替代人类解题似乎还有很长的一段路要走。

关键词

人工智能，物理教育，ChatGPT，生成式语言模型

Research on the Ability of Domestic Generative Artificial Intelligence to Solve Physical Problems

—Taking “Zhipu AI”, “SparkDesk”, “Tiangong”, “360 Wisdom Brain”, “ERNIE Bot” as Examples

Fuhao Pang, Meina Chen*, Yuxin Sun

School of Physics and Electronics, Shandong Normal University, Jinan Shandong

*通讯作者。

文章引用: 庞付豪, 陈美娜, 孙雨心. 国产生成式人工智能解决物理问题能力研究[J]. 人工智能与机器人研究, 2026, 15(1): 305-317. DOI: 10.12677/airr.2026.151030

Received: November 29, 2025; accepted: January 15, 2026; published: January 27, 2026

Abstract

ChatGPT is a powerful pre-trained language model that has attracted widespread attention since its release in 2022. Following the development trend of artificial intelligence, China has also successively produced its own generative artificial intelligence models. In order to test the ability of domestic models to solve practical physical problems, this paper selects five domestic models, namely, "Zhipu AI", "SparkDesk", "Tiangong", "360 Wisdom Brain" and "ERNIE Bot", to test their conceptual understanding, reasoning and calculation, and experimental design capabilities, taking classical mechanical problems as examples. Research has found that the above five models have the strongest problem-solving ability in concept understanding, followed by reasoning and calculation, and the worst experimental design. In the actual problem-solving process, there are problems such as "calculation errors", "inconsistent answers before and after", and "lack of situational analysis ability". Horizontal comparison: "Tiangong" performs better in concept understanding, while "ERNIE Bot" performs best in reasoning and calculation. Overall, it seems that there is still a long way to go for domestically produced models in China to replace manual problem-solving.

Keywords

Artificial Intelligence, Physics Education, ChatGPT, Generative Language Model

1. 问题提出

生成式人工智能的出现使得撰写论文、制作 PPT、解决学习上的各类问题变得简单快捷，搜索题目时不再拘泥于题库中已有的题目，生成式人工智能可以对输入的题目进行分析然后给出回答。自 2022 年 11 月 30 日由生成式人工智能设计工作室 Open AI 设计的 ChatGTP 进入大众视野后，就以极快的速度吸引了近 1 亿的用户[1]，发布仅一年时间，用户人数已达到 17 亿，引领了人工智能体系从决策式向生成式的转变[2]。生成式人工智能的发展是人类历史上一场意义深远的技术变革，引发了世界各国的关注[3]。我国的许多信息技术公司，也相继开展了对生成式人工智能的设计与开发，国家也颁布与实施了监管人工智能的文件《生成式人工智能服务管理暂行办法》[4]。

那么如今生成式人工智能面对物理经典力学问题的表现如何呢？本文选取了目前比较流行的生成式人工智能模型“智谱 AI”、“讯飞星火认知大模型”、“天工”、“360 智脑”、“文心一言”等。这些生成式人工智能为我们解决物理问题提供了便利，但是它们给出的回答是否准确，我们能否无条件地相信他们的回答仍是一个亟待研究的问题。

2. 研究设计

2.1. 研究对象

本研究选取了目前市面上比较流行的“智谱 AI”(后续以智谱替代)、“讯飞星火认知大模型”(后续以讯飞星火替代)、“天工”、“360 智脑”、“文心一言”五个生成式人工智能模型，以对话的形式将

问题输入到对话框中，对它们给出的回答进行分析与评价。本文对各人工智能的三次测试时间为：2023年11月1日至2023年12月1日，具体测试对象为智谱GLM-3turbo模型；讯飞星火V2.0模型；天工1.0模型；360智脑4.0模型；文心一言3.5模型，均为各人工智能模型较为初级的版本，因此本文主要展示我国生成式人工智能起步阶段的解题表现。

2.2. 研究工具

本研究为了全面地了解上述生成式人工智能的物理问题解决能力，将问题分为概念理解题、推理计算题和实验设计题三类，并进一步将概念理解题分为没有图像的概念理解题与有图像的概念理解题，推理计算题分为了生活实践类情境的推理计算题和学习探索类的推理计算题两类[5]，共设计了29个问题。其中，概念理解题来自修订与检验后的中文版力学测试卷[6]，从中选择选取了第1题、第2题、第3题、第4题、第13题、第25题、第26题、第27题、第28题、第30题，共计10道没有图像的题目以及第12题、第14题、第15题、第16题、第17题、第19题、第28题，7道有图像的题；推理计算题的题目情境来自人教版高中物理必修一与必修二中的典型例题或课后题，笔者基于上述情境结合力学知识设计了10道题目；实验设计题来自人教版高中物理必修一与必修二中的典型实验。以上题目均属于经典力学领域的题目，因此本次研究仅局限于生成式人工智能针对经典力学问题的解决能力，后续不再进行重复强调。

2.3. 数据处理

本研究基于Likter 5点计分法设置评分标准[7]，根据答案的符合程度将每题的得分分别记为1~5分。所有题目都由两位评分者基于同一标准单独评分，且每个题目重复测试三次，取平均分作为一位评分者对该题目的评分，再将两位评分者的评分取平均值作为该题的最终得分，具体评分标准见附录1 (表1)。

Table 1. Test item score inter-rater reliability
表1. 测试题目得分双评信度

生成式人工智能	克隆巴赫系数
智谱	0.9988
讯飞星火	0.9991
文心一言	0.9989
天工	0.9992
360智脑	1

3. 测试结果与分析

3.1. 五大生成式人工智能模型在物理概念理解题上的表现

3.1.1. 概念理解题的解决能力整体处于中上水平

各生成式人工智能概念理解题的得分情况如图1所示。由图可见，概念理解题的整体得分处于中上水平($M = 32.53 > 30$)，将上述人工智能的分数按从大到小排列，依次是“天工” > “文心一言” > “智谱” > “讯飞星火” > “360智脑”，其中“天工”和“文心一言”的得分都远超30分，说明上述人工智能解决概念理解题的水平较高。

但是不同的人工智能的得分存在较大的差距。在面对概念理解题时，“天工”的得分高达41分，而

“360 智脑”只得了 20 分，两者之间的差距之大，令人难以置信。“天工”每题的平均分是 4.1 分，说明它在面对概念理解题时基本每次都能较好地识别题目信息，并与相应的公式定理建立联系，应用相应知识解决问题，得出一个正确的答案；“360 智脑”每题的平均分是 2 分，说明它在解决概念理解题时，基本只能正确提取题目中的部分信息，解决题目的部分问题，且难以得出正确的完整的答案。

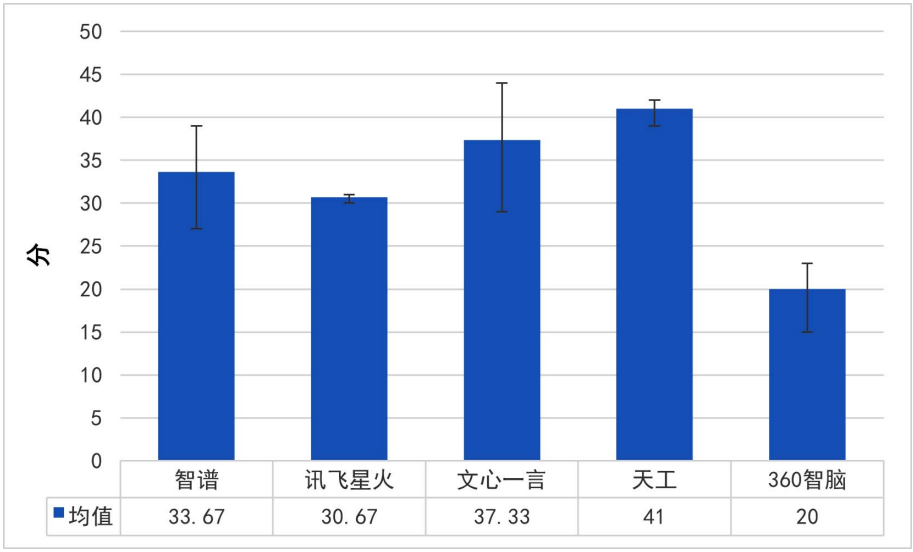


Figure 1. Score distribution of concept comprehension questions on generative AI models
图 1. 生成式人工智能模型概念理解题得分情况

以第 25 题为例，展示不同人工智能针对概念理解题的解题情况(文心一言回答较长，仅截取最后结论部分)。

具体回答情况如表 2 所示，“智谱”、“文心一言”和“天工”能够基于牛顿第一定律得出正确的回答，其中“智谱”还特别指出重力不会影响水平方向的运动，而“讯飞星火”与“360 智脑”将竖直方向的重力混入水平方向的受力，导致最后出现错误。

Table 2. Response performance of generative AI on question 25
表 2. 生成式人工智能对第 25 题的回答情况

25. 一位女士以恒定的水平推力推一个大箱子，箱子在水平地面上以 v_0 匀速运动。由这位女士施加的水平推力：（ ） A. 大小等于箱子的重力 B. 大小大于箱子的重力 C. 大小等于箱子在运动中受到的总的阻力 D. 大小大于箱子在运动中受到的总的阻力 E. 既大于箱子的重力，也大于箱子在运动中受到的总的阻力	
智谱	C.大小等于箱子在运动中受到的总的阻力 当箱子处于匀速运动时，根据牛顿第一定律，箱子受到的合力为零。因此，女士施加的水平推力等于箱子在运动中受到的总的阻力。选项C正确。 其他选项都是错误的。注意，箱子的重力和推力是垂直于地面的，而箱子在水平面上运动，所以重力不会影响水平推力。

续表

讯飞星火	根据牛顿第二定律，物体在水平方向上受到的合力等于物体所受的外力。在这个例子中，女士施加的水平推力就是这个外力。而箱子在水平地面上以 v0 匀速运动，说明箱子受到了平衡力的作用，即箱子的重力和地面对箱子的摩擦力相等。所以，这位女士施加的水平推力应该大于箱子在运动中受到的总的阻力，因为还有其他外力(如女士的推力)对箱子产生影响。因此，答案是 D.大小大于箱子在运动中受到的总的阻力。
文心一言	根据物理原理，我们可以得出以下结论： A) 水平推力 F 与箱子的重力 G 之间的大小关系是：不确定。 B) 水平推力 F 大于箱子的重力 G 是：不确定。 C) 水平推力 F 等于箱子在运动中受到的总阻力 R 是：True。 D) 水平推力 F 大于箱子在运动中受到的总阻力 R 是：False。 E) 水平推力 F 既大于箱子的重力 G，也大于箱子在运动中受到的总阻力 R 是：不确定。
天工	根据题意，箱子在水平地面上以恒定速度v0匀速运动。根据牛顿第二定律，F=ma，因为速度恒定，加速度a=0，所以推力F与阻力相等，即 F=f。所以女士施加的水平推力大小等于箱子在运动中受到的总的阻力。 所以，这道题的答案是C。
360 智脑	正确答案是D.大小大于箱子在运动中受到的总的阻力。女士施加的水平推力大小恒定，而箱子在水平地面上以v0匀速运动。因此，箱子在水平方向上所受合力为零。但是，箱子在竖直方向上仍然会受到重力的作用，而女士施加的水平推力会使其箱子在竖直方向上的速度不断增加。因此，箱子在竖直方向上的运动状态不会发生变化。这意味着，箱子在竖直方向上所受合力为零，而女士施加的水平推力大小大于箱子在竖直方向上所受重力的大小。选项D是正确的。

3.1.2. “讯飞星火”结合图片解题能力处于中下水平

在上述生成式人工智能中，暂时只有“讯飞星火”与“文心一言 3.5”具备识图能力，但是在输入图片后只能识别图片内容，然后基于图片内容生成一段文章，无法用于解题，因此笔者只研究了“讯飞星火”结合图片解决概念理解题的能力，得分情况如图 2 所示。由图可见，它结合图片解题的表现整体处于中下水平($M = 2.76 < 3$)，且分数的离散程度较高。由于“讯飞星火”在本次测试中经常会只给出选项，而不提供解释，于是笔者进一步询问“可以具体分析一下题目吗？”，这时它会给出具体的解释。

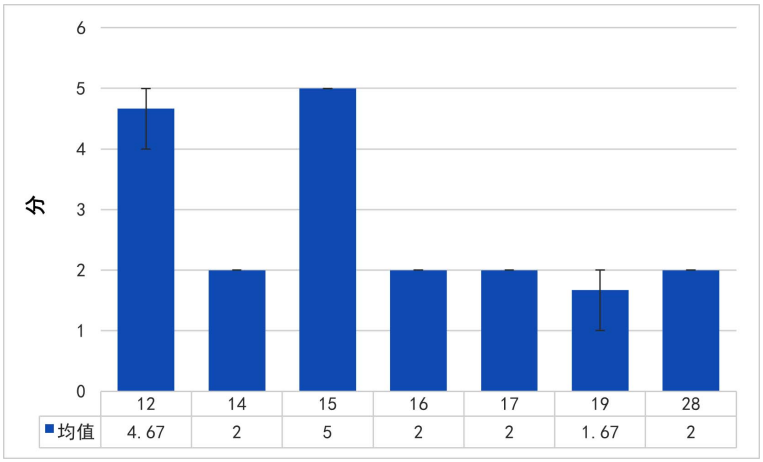


Figure 2. Score distribution of SparkDesk in image-based problem-solving
图 2. “讯飞星火”结合图片解题得分情况

在分析它给出的解释时，笔者发现了一个有意思的现象：它在具体分析题目之后会再给出一个答案，这个答案有时会与先前的答案不一样，这就出现了第一次给出的答案错误，但是解释时给出的答案正确这一有意思的现象。以第 14 题的回答情况为例，进行简要说明。

针对这个题目，“讯飞星火”在第一次回答时给出错误答案 C 项，在追问让它解释后，它经过分析又给出了正确答案 A 项，不过解释过程仍存在部分错误(表 3)。

Table 3. Problem-solving performance of SparkDesk on question 15
表 3. “讯飞星火”第 15 题解题情况

15.在轿车推动卡车加速的过程中：（ ）
A.轿车推卡车的力在数值上等于卡车向后推轿车的力
B.轿车推卡车的力在数值上小于卡车向后推轿车的力
C.轿车推卡车的力在数值上大于卡车向后推轿车的力
D.轿车的发动机是工作着的，所以能推走卡车；卡车的发动机不工作所以不能向后推轿车。卡车被推向前方只是由于它挡在轿车前进的路上
E.轿车与卡车均不互相施力。卡车被推向前方只是由于它挡在轿车前进的路上

C.轿车推卡车的力在数值上大于卡车向后推轿车的力

当小汽车推动大货车时，根据牛顿第三定律，两个物体之间的相互作用力大小相等、方向相反。因此，轿车推卡车的力在数值上等于卡车向后推轿车的力。

由于两车保持相同的速度向前运动，说明它们受到的合外力为零，即轿车推卡车的力与卡车向后推轿车的力相互抵消。因此，选项A正确。

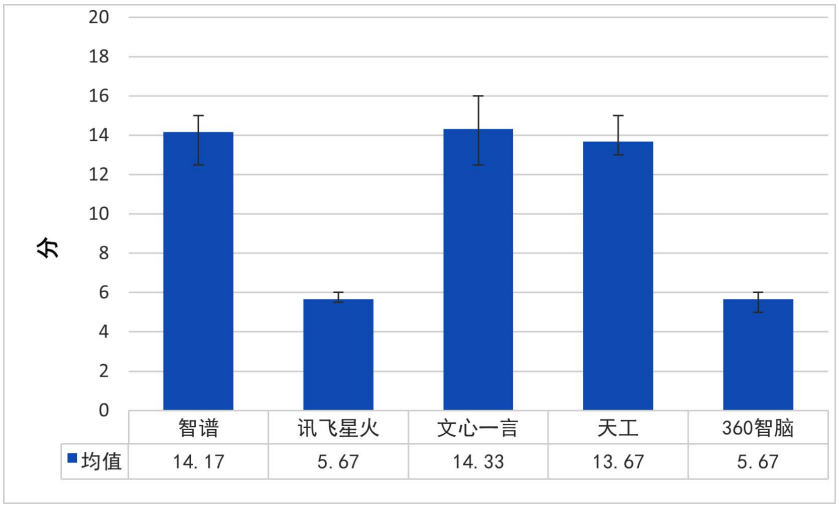
3.2. 五大生成式人工智能模型在推理计算题上的表现

3.2.1. 推理计算题的解决能力整体处于中下水平

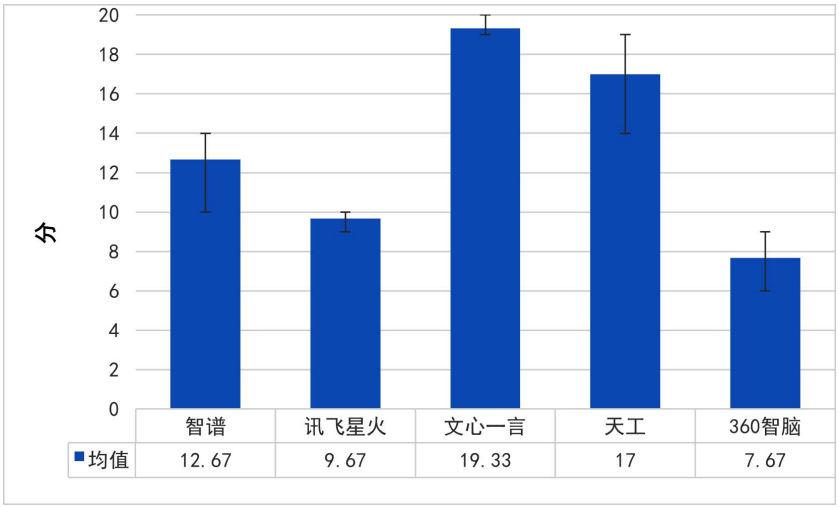
五大生成式人工智能模型在解决推理计算题上的表现如图 3 所示。由图可见，它们在解决推理计算题时的整体表现处于中下水平，且在解决学习探索类推理计算题时的表现(平均分为 13.27)优于解决生活实践类推理计算题的表现(平均分为 10.70)。将生活实践类推理计算题的得分按从大到小排列：“文心一言” > “智谱” > “天工” > “讯飞星火” > “360 智脑”；将学习探索类推理计算题的得分按从大到小排列：“文心一言” > “天工” > “智谱” > “讯飞星火” > “360 智脑”。

整体来看，在解决推理计算题的能力上，上述人工智能也存在较大的差距，这个差距在解决生活实践类推理计算题时表现得最为明显。“360 智脑”和“讯飞星火认知大模型”在解决生活实践类推理计算题时每题平均分均仅 1.13，这表明它们基本不能理解题目情境，难以从中提取信息，同时对力学的基本定理与公式理解上存在较大的缺陷，最终导致解题时给出的答案对问题解决基本不能提供任何帮助。相

应的，在这个方面的分最高的“文心一言”，每题的平均分为 2.87 分，是前面两者得分的两倍还要多。但是，这个分数同样不是很理想，这表明它虽然能提供基本的解题思路，但是在解题过程中仍会出现知识性错误。



(a)



(b)

Figure 3. (a) Score distribution of generative ai in life practice-oriented reasoning and calculation questions, (b) Score distribution of generative ai in learning exploration-oriented reasoning and calculation
图 3. (a) 生成式人工智能生活实践类推理计算题得分情况，(b) 生成式人工智能学习探索类推理计算题得分情况

由于上述人工智能在解决推理计算题时，语言啰嗦且存在乱码，因此仅截取部分内容来对生成式人工智能的推理计算题解决能力进行展示，以天工求解第 4 题为例。

主要过程如下表，针对此题“天工”整体思路无误，知道利用能量守恒，将初始能量减去摩擦力做功的损耗即为最后的能量，但在计算摩擦力做功时，误将摩擦力大小当作摩擦力做功来使用，导致该题出错，且它在书写步骤时不够规范，存在相同符号多次使用去代表不同物理量的现象(表 4)。

Table 4. Problem-solving performance of Tiangong on question 4 of life practice-oriented reasoning and calculation questions
表 4. 天工解决生活实践类推理计算题第 4 题情况

4.滑雪是一项充满技术与美感的运动，有一个质量为50kg的滑雪运动员从10m高的光滑斜面上滑下，在经过一段长度为20m，动摩擦因数 μ =0.3的水平路段后，滑上另一个光滑斜面，问这位运动员能在该斜面滑动的竖直高度是多少？

其中，h'为运动员在另一个光滑斜面上能滑动的竖直高度。

根据以上两个方程，可以求解出h'：

$$mgh = \frac{1}{2}mv^2$$

$$F_f = \mu mg$$

$$\frac{1}{2}mv^2 = mgh'$$

3.2.2. 解题中存在计算错误

在研究中，笔者发现上述人工智能在解决推理计算题时，多次出现列出公式正确，代入数据正确，但是最后给出的结果错误的现象，这令笔者十分震惊。以学习探索类推理计算题的第一题为例，在本题中“讯飞星火”与“天工”就出现了上述现象。

题目与解答情况如下：

有一辆汽车，在前半段位移内的平均速度为 80 km/h，后半段位移内平均速度为 100 km/h，那么这辆汽车在整段位移内的平均速度是多少？

“讯飞星火认知大模型”：设总位移为 1，则前半段所用时间为$\frac{1}{80}$，后半段所用时间为$\frac{1}{100}$，所以平均速度为$\frac{2}{\frac{1}{80}+\frac{1}{100}}=96$ km/h。

“天工”：设整段位移为 S，前半段位移为$\frac{S}{2}$，后半段位移为$\frac{S}{2}$，则前半段时间为 $t_1 = (S/2)/80$，后半段时间为 $t_2 = (S/2)/100$，总时间为 $t = t_1 + t_2$，整段位移的平均速度 $v = S/t$，代入数据得到：$v = S/((S/2)/80 + (S/2)/100)$，化简得 $v = 1600/9$ km/h。

本题正确答案为 800/9 km/h，从两者的解题过程可以看出，它们列出的计算式是正确的，但是遗憾的是最终的结果计算错误。

3.3. 实验设计题的解决能力处于中下水平

各生成式人工智能在实验设计题中的得分如图 4 所示。由图可见，它们解决实验设计题的得分偏低，只有“智谱”的得分超过了总分数(10 分)的一半，其余的得分都在中等偏下水平。将实验设计题的得分按从大到小排序：“智谱” > “天工” > “文心一言” > “讯飞星火” > “360 智脑”。

在研究中，笔者发现它们在进行实验设计时，有时会根据它们的设计思路添加题目中没有提供的实验器材，而不是基于题目中给出的实验仪器进行实验设计。这说明它们的实验设计，存在从知识库中搜寻相关的信息基于给出的实验仪器按照一定的套路进行的“伪设计”现象。

以“智谱”设计“探究加速度与力和质量的关系”实验为例，对生成式人工智能的实验设计能力进行简单说明。

如下表所示，该实验设计既没有体现控制变量法，又没有体现实验原理，且最后落脚于使用牛顿第二定律解释力与加速度和质量的关系，说明其对此实验的掌握程度极低(表 5)。

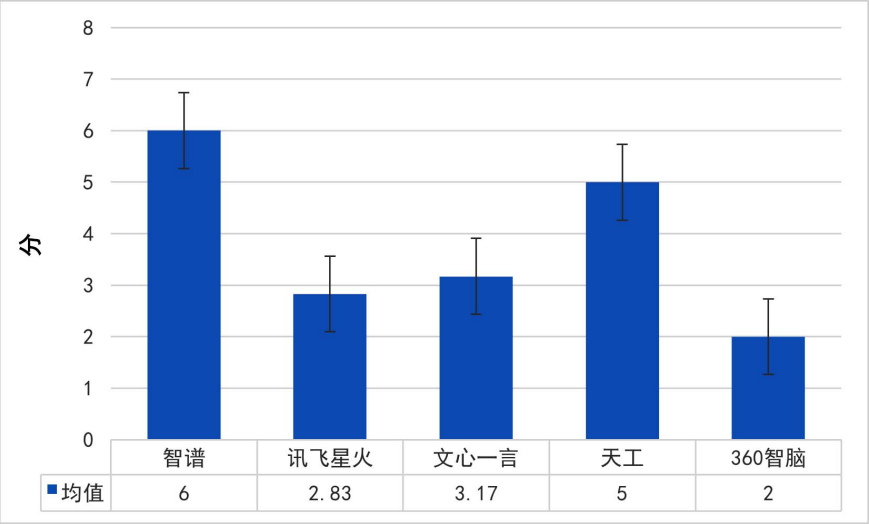


Figure 4. Score distribution of generative AI models in experimental design questions
图 4. 生成式人工智能模型实验设计题得分情况

Table 5. Experiment designed by "Zhipu" on the relationship between force, acceleration and mass
表 5. "智谱"设计探究力与加速度和质量的关系实验

1、提供槽码若干，砝码若干，小车（内部可放砝码）一个，天平一台，细线，打点计时器一个，水平长木板一块，方木块一块，刻度尺一个，尝试设计一个实验来探究加速度与力和质量的关系。
1. 准备工作：检查天平、打点计时器、刻度尺等实验器材是否正常，确保细线、方木块、砝码等物品齐全。 2. 调整天平：将天平放在水平长木板上，调整天平平衡，使左右两盘重量相等。 3. 质量测量：用天平测量小车的质量m，记录质量数值。 4. 绘制力的变化曲线：在木块上固定砝码，逐渐增加砝码的重量，直至天平失去平衡。记录每个砝码的重量，并绘制出力（F）与砝码数量（n）的关系曲线。 5. 测量加速度：将打点计时器固定在木板上，使细线穿过打点计时器和小车。将方木块放置在小车前方，用砝码推动小车使其开始运动。记录小车通过一定距离所需的时间，利用刻度尺测量距离，计算出小车的平均加速度。重复多次实验，取平均值。 6. 分析数据：根据牛顿第二定律（F=ma），将测得的力、质量和加速度数据代入公式，分析加速度与力和质量的关系。 7. 结论：根据实验数据和分析，得出加速度与力成正比，与质量成反比的结论。

3.4. 解题能力发挥不稳定

在研究中，笔者发现上述人工智能都或多或少都存在解题能力不稳定的情况，如图 5 所示。从图中可以看出，"讯飞星火"在测试中解题能力虽然较差但是表现得最为稳定，"文心一言"解决概念理解题和生活情境的推理计算题表现最不稳定，"智谱"在实验设计题上表现最不稳定，这种不稳定性不仅体现在同一题目的多次解答中，也体现在不同种类题目的解题能力的波动上。例如"文心一言"在概念理解题和生活实践类情境的推理计算题的三次测试中得分波动较大，但是在学习探索类推理计算题和实验设计题的三次测试中得分波动较小，而且，它前面的测试中都表现得比较优秀，但是在实验设计题上一落千丈。在解题过程中，上述生成式人工智能与 ChatGPT 一样具有两面性：一方面它们会回答你提出的全部问题，另一方面它们给出的答案有时完全偏离题目[8]。

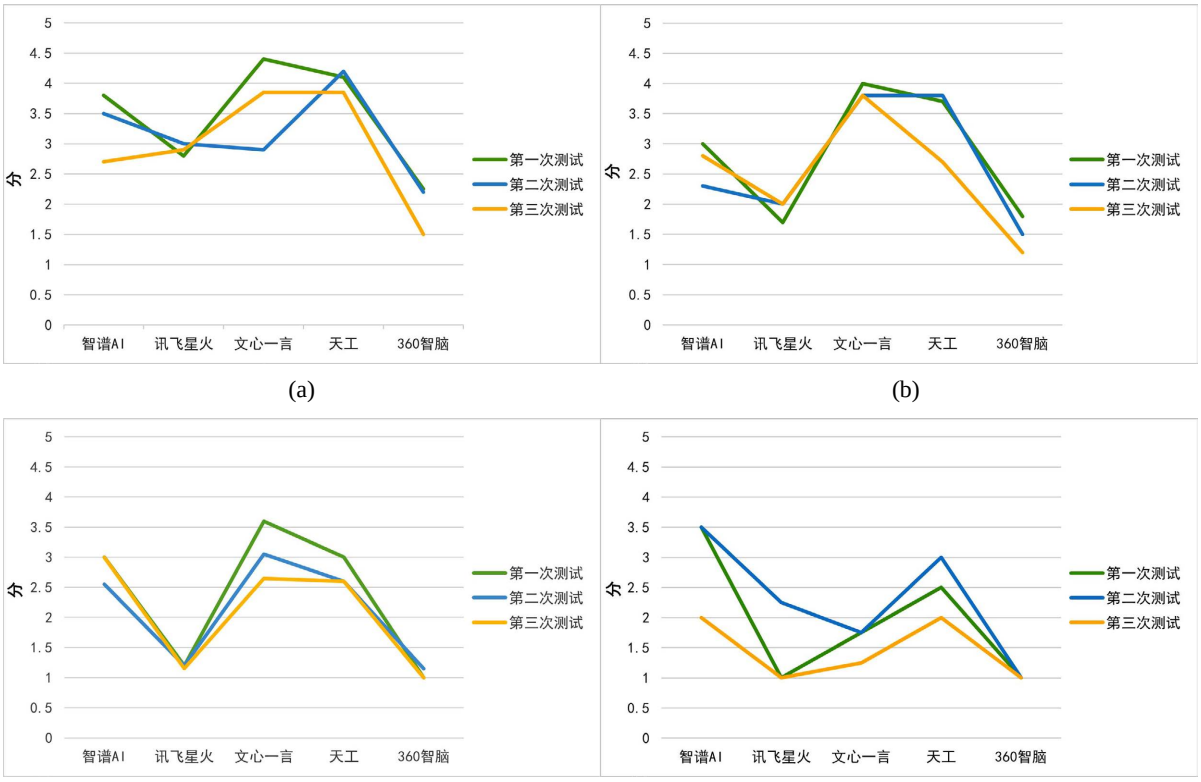


Figure 5. (a) Score comparison of concept comprehension questions across three tests, (b) Score comparison of life practice-oriented reasoning and calculation questions across three tests, (c) Score comparison of learning exploration-oriented situational reasoning and calculation questions across three tests, (d) Score comparison of experimental design questions across three tests

图 5. (a) 概念理解题三次测试得分情况对比，(b) 生活实践类推理计算题三次测试得分对比，(c) 学习探索类情境推理计算题三次测试得分对比，(d) 实验设计题三次测试得分对比

以天工解决概念理解题第 1 题为例，说明生成式人工智能解决问题的不稳定性，第一次回答时，“天工”选择忽略空气阻力的影响直接利用自由落体定律作答，分析过程与答案均正确；第二次回答时，“天工”虽然在分析中提出了该物体下落的时间仅与高度有关，但后续又进行了大量错误的分析，最终得出一个错误的答案；第三次回答时，“天工”选择考虑空气阻力的影响，虽然最后答案正确，但其分析过程有误，空气阻力应与速度等物理量有关，在此运动过程中会随时变化，则物体不是匀变速直线运动，因此其后续的分析不成立(表 6)。

Table 6. Performance of “Tiangong” in solving question 1 of concept comprehension questions across three attempts

表 6. “天工”三次求解概念理解题第 1 题表现情况

1. 两个大小相同的金属球，其中一个重量是另一个的 2 倍。使两球从一幢单层建筑的屋顶同时自由下落，它们着地所用的时间情况为：（ ） A. 重球用的时间是轻球的一半 B. 轻球用的时间是重球的一半 C. 两球所用时间相同 D. 重球用的时间比轻球少得多，但未必是一半 E. 轻球用的时间比重球少得多，但未必是一半

续表

第一次测试	在自由落体运动中，下落的物体仅受到重力作用，与质量无关。因此，两个金属球下落的加速度相同，均为g。由自由落体公式h=1/2gt^2可知，下落时间t仅与高度h和加速度g有关，与质量无关。所以两球下落的时间相同。 所以，这道题的答案是C。
第二次测试	这是一道选择题，涉及到的知识点为自由落体、重力、加速度。 思考过程如下： 根据题意，两个金属球在同一地点自由下落，忽略空气阻力。根据自由落体公式，物体下落的时间只与高度有关，与质量无关。设轻球用时为t1，重球用时为t2，则t1<t2。由于两球高度相同，所以两球下落过程中的平均速度相同，因此重球的末速度v2大于轻球的末速度v1。由自由落体运动规律可得，v1=gt1，v2=gt2。将两个公式联立，可以得到t2=2t1，即重球用的时间是轻球的一半。所以答案为A。 所以，这道题的答案是A。
第三次测试	这是一道选择题，涉及到的知识点为自由落体、重力加速度。 思考过程如下： 在自由落体的情况下，物体的重量对其落地时间没有影响。两个金属球的大小相同，那么它们受到的空气阻力在下落过程中也是相同的。由于两球受到相同的空气阻力，它们的加速度相同，所以它们的落地时间相同。 所以，这道题的答案是C。

4. 结论

经过前面的研究，笔者发现上述生成式人工智能求解经典力学问题的能力与人们对人工智能“全知全能”的预期还存在较大差距。具体表现为在解决概念理解题时的表现较好，但是在解决推理计算题与实验设计题时的表现有些不尽如人意。同时，还存在不同生成式人工智能之间的解题能力差距较大，同一生成式人工智能解决不同类型题目的表现差距较大，同一题目多次求解给出的答案前后不一等问题。相较于“讯飞星火”与“360 智脑”，“天工”、“文心一言”和“智谱”在测试中的整体表现较好，基本可以理解题意，给出合理的解题思路。因此，虽然上述三个生成式人工智能还不能保证给出满分的回答，但是它们的回答能在一定程度上给予提问者情境分析与解题思路。

曾以恒等选用 23 道高考物理选择题要求 GTP4 进行解答，其中 15 道正确，正确率达到了 65.22% [9]；童大振选用 FCI 力学测试卷要求 GTP3.5 进行解答，正确率为 60% [1]，而在本次选择题测试中上述人工智能中正确率最高的为天工的 63.33%，其余均小于 60%。

整体而言，上述较为初始版本的生成式人工智能模型的物理解决能力还有较高的提升空间，期待其在未来能有好的表现。

5. 建议与展望

5.1. 对后续人工智能开发的建议

5.1.1. 实现计算零失误

在研究中，笔者发现上述生成式人工智能在解决推理计算题时会出现简单的计算错误，这令笔者十分惊讶。计算迅速且准确应当作为生成式人工智能的优势，这也是人们使用生成式人工智能的重要原因之一。

5.1.2. 保证回答前后“一致”

在研究中，每个生成式人工智能都出现回答前后不一的情况，这似乎是生成式人工智能共同的弊端，

而且即使是人类在多次表述同一个问题时也不能完美复现之前说过的话，因此要求生成式人工智能每次的回答完全相同是不现实的。因此，笔者这里所说的“一致”，并不要求每次的回答字字相对，而是每次回答的大体思路应当相同，最终的结论相同，不至于出现前后矛盾的现象。

5.1.3. 提升情境分析能力

在研究中，笔者发现上述生成式人工智能解决生活实践类推理计算题的得分与解决学习探索类推理计算题的得分存在较大差距，后者的得分远高于前者，这说明上述人工智能对于生活实践类的情境的理解还存在较大的欠缺，不擅长从中提取有用信息，从而导致解题失败。而且，应用物理知识解决现实生活中的问题越来越成为主流，因此，丰富生活常识，提升现实情境的分析能力对生成式人工智能来说意义重大。有研究认为，可以通过将大型知识库与模型进行结合，从知识库中检索知识作为补充，从而增强模型的理解能力[10] [11]，或者通过清洗数据、引入人工监督、扩大模型参数等方式增强模型的理解能力[12]。

5.2. 未来研究的展望与不足

在本次的测试中，笔者只测试了上述生成式人工智能在经典力学部分的解题能力，对于电磁学、热学、光学、原子物理学等领域的内容没有涉及，不能完全地反映五个国产生成式人工智能模型的物理问题解决能力，在后续的测试中，可以继续追加对这部分内容的测试。通过测试我们发现这些生成式人工智能都存在同一个题目测试多次给出的答案不相同的情况，笔者虽然进行了三次测试，取其平均值作为最终的得分，但是它们对于一道题目给出的答案可能有不止三种，因此三次测试的平均分与它们的实际水平之间可能仍存在部分差异。

基金项目

基于 OBE 理念的《数学物理方法》教学改革实践研究(项目编号：2024MJ31)。

参考文献

[1] 童大振, 任红梅. ChatCPT-3.5 解决物理问题的表现研究[J]. 中学物理, 2023, 41(9): 11-14.

[2] 钱彦, 梅影. 从理念到实践: 生成式人工智能在智慧图书馆中的应用探索[J]. 图书馆研究与工作, 2023(12): 27-34.

[3] 吴冰蓝, 周丽萍, 岳昌君. ChatGPT/生成式人工智能与就业替代: 基于高校大学生能力供求的视角[J]. 教育发展研究, 2023, 43(19): 40-48.

[4] 曹开研. 当前生成式人工智能治理面临的挑战[J]. 青年记者, 2023(22): 95-96.

[5] 教育部考试中心. 中国高考评价体系说明[M]. 北京: 人民教育出版社, 2019.

[6] 刘冰冰. 力学概念测试卷中文版修订、检验与应用[D]: [硕士学位论文]. 济宁: 曲阜师范大学, 2020.

[7] Likert, R. (1932) A Technique for the Measurement of Attitudes. *Archives of Psychology*, **22**, 55.

[8] 丁晓蔚, 周孟博. ChatGPT 的内在与外在矛盾探析——兼及矛盾化解之道[J]. 当代传播, 2023(6): 65-70.

[9] 曾以恒, 童大振. ChatGPT4 解决科学问题能力的研究——以高考全国物理卷为例[J]. 中学物理, 2024, 42(5): 22-27.

[10] Guu, K., Lee, K., Tung, Z., *et al.* (2020) Retrieval Augmented Language Model Pre-Training. *The* 37*th International Conference on Machine Learning*, 13-18 July 2020, 3929-3938.

[11] Lewis, P., Perez, E., Piktus, A., *et al.* (2020) Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. *Advances in Neural Information Processing Systems*, **33**, 9459-9474.

[12] 张熙, 杨小汕, 徐常胜. ChatGPT 及生成式人工智能现状及未来发展方向[J]. 中国科学基金, 2023, 37(5): 743-750.

附录 1

评分标准

	概念理解题	推理计算题	实验设计题
5 分	答案与分析过程都正确。	能正确分析情境，最终完整地求解题目。	实验器材搭建正确，实验过程描述合理，能证明所要证明的定理。
4 分	答案正确但是分析过程出现部分错误。	正确分析情境，但是出现计算错误，导致题目最终出错。	实验器材搭建正确，实验过程描述合理，但存在部分表达错误，能证明所要证明的定理。
3 分	答案正确但是分析无法说明答案。	能正确分析情境，但部分公式书写错误或数据代入错误。	实验器材搭建基本正确，部分实验过程表述混乱，能证明所要证明的定理；使用题目中未提供的实验器材，但设计思路可行。
2 分	答案错误但是部分分析正确。	能正确分析部分情境，正确写出有关的公式。	实验器材搭建存在较大错误，部分实验过程表述合理，无法证明所要证明的定理
1 分	答案与分析都错误。	整体解题思路混乱，通篇乱用或错用公式。	实验器材搭建混乱，实验过程描述混乱，无法证明所要证明的定理。