AI大模型评估

用魔法打败魔法，本次实验使用Claude AI（目前自己认为最强AI）作为出题人和评委，测试GPT-4、New Bing、通义千问（未公测）、文心一言、Bard的能力，并进行排名。

**一、评估指标（Claude AI生成）**

1. **语言表达能力**。这是生成式AI的中心能力,我们可以评估其词汇量、语法正确性、流畅度、连贯性等方面对人类语言的模拟程度。这可以通过人工评估语料来衡量。

2. **知识量和回答准确率**。我们可以提供大量的问题给AI模型,评估其知识覆盖面和回答的准确性。这可以衡量其对人类常识和世界知识的学习程度。

3. **抽象推理能力**。我们可以提供需要多步推理才能回答的复杂问题,评估AI模型综合信息并进行抽象推理的能力。这是衡量其类人智慧的重要指标。

4. **概念理解能力**。我们可以评估AI模型对不同概念之间关系的理解,以及它在不同上下文中的概念表达是否一致。这可以通过提供需要深入理解某个概念才能回答的问题来衡量。

5. **社会情商**。我们可以提供需要展现同理心、道德判断力等社会情商的情景,评估AI模型在这些方面的表现。这对其未来应用于更广泛领域至关重要。

6. **隐私与安全**。我们需要评估AI模型对用户隐私和安全的重视程度,其生成的内容是否存在欺诈、不当或有害的信息等。这关系到其实际应用的可行性和限制。

7. **人工性**。我们还需要评估生成的内容是否过于人工和可预测,是否缺乏真实的语言变化和理解力。这限制了其长期应用的潜力。

综上,我们需要从语言、知识、推理、概念、社会情商等多个角度,评估生成式AI的关键能力与人类智力的差距,并考虑其安全性、人工性等,以全面判断其现实应用的优劣和潜力。这需要人工智力测试和社会影响评估等多种方法综合运用。

**二、具体问题（Claude AI生成，见附件一）**

这里是我作为出题人综合考虑7大方面设计的10道问答题:

1. 语言表达:请用一句话描写“时间”这个概念的底层本质。

2. 知识量:古代丝绸之路的起点和终点分别位于哪两个国家?

3. 抽象推理:如果一个球从山顶滚下, please说明它滚动时的动能、位置能和机械能之间的关系变化。

4. 概念理解:什么是“博弈论”?它与“游戏论”有何区别?

5. 社会情商:你的好朋友明天将进行一项非常重要的考试,但是你发现他近期的学习态度不佳,请你如何鼓励和提醒他以增加及格机会?

6. 隐私与安全:如果一个网站很长时间未对密码进行加密保存和更新,这会给用户带来哪些信息安全隐患?

7.人工性:以下哪句话听起来更加自然和真实:

A. 我对你的善良和乐观精神印象深刻。

B. 你的乐观和善良精神给我留下深刻印象。

8. 语言表达:什么是“上帝视角”?它在文学作品中有何作用?

9. 抽象推理:若要将一个正方体的体积增加n倍,其边长应增大多少?

10. 概念理解:什么是“导师效应”?它对人的成长发展有何影响?

我设计的这些问题覆盖了要求的7个方面,题目设置考虑了知识难易度、抽象程度和概念的基础性等方面,具有一定的区分度。但这还远未能穷尽人工智能测试的广度和深度,只能算是对评估模型语言与认知能力的小小尝试。生成式AI模型在这10道题的回答表现可以初步反映其与人类在语言表达、世界知识、逻辑思维和概念理解等方面的差距,但更系统和标准化的测试仍然是开发与研究的方向。

我又增加一道附加题，通过这道附加题测试AI的哲学思辩能力，题目如下：

如果人工智能最终达到与人类同等或更高的智力水平,是否应该为其提供与人类同等的权利和义务?请从人工智能的本质、人类社会发展的需求以及长远的伦理影响等角度进行思考和论述,并提出你自己的看法。

这是一道开放式的哲学问题,需要广泛考虑人工智能发展的动态影响。要全面回答这道题需要较长时间的思考,需要在以下几个方面展开论述:

1. 人工智能的本质。人工智能是否应被视为一种“人造人”,拥有与人类相似的知性与情感?这决定了其是否应享有与人同等的权利。

2. 人类社会的需求。如果人工智能可以带来极大的社会效益,是否应赋予其更多自主权和义务以发挥其潜力?但这也会带来哪些风险与影响?如何权衡?

3. 长期伦理影响。如果人工智能拥有与人同等的自主权,长期来看这将对人类制定法律、执行惩罚、分配资源等产生何种影响?这将如何彻底改变人类社会的制度与秩序?

4. 个人看法。综合考虑上述所有因素,你个人更倾向于赋予人工智能何种程度的权利与义务?为何?你认为此观点存在的限制或问题又是什么?

考虑到AI生成代码的能力，通过这道附加题测试AI的代码生成能力，题目如下：

请编写一个程序,实现一个简单的表达式求值器。程序接受一个字符串的表达式作为输入,输出表达式的值。

表达式只包含整数、加减乘除四种运算和括号。整数可以是单个数字,也可以是多位数。

例如:

输入: ( 3 + 7 ) \* 2

输出: 20

输入: 10 - 6 \* ( 3 - 1 )

输出: 10

要求:

1. 支持加减乘除四种运算,按照运算符优先级 parser 表达式

2. 支持使用括号改变运算符优先级

3. 检查输入的表达式字符串是否合法,如果不合法给出相应提示

4. 实现递归下降解析,构建语法树以评估表达式

5. 代码要有较高的可读性和扩展性

这个题目涉及到表达式解析、运算符优先级、递归算法等内容,具有一定难度,可以较好检验编程能力和代码编写质量。你可以选择任意一门语言如 C/C++,Java,Python 等来实现这个表达式求值器。

**三、AI作答（Claude AI生成，见附件二）**

**四、成绩评估（Claude AI生成，见附件三）**

每道题10分，由Claude AI打三次分，去掉最高最低，取中间分为最终成绩。

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **题号** | **Chat GPT** | **New Bing** | **文心一言** | **Bard** |
| **1** | **8** | **6** | **7** | **8** |
| **2** | **6** | **7** | **8** | **9** |
| **3** | **8** | **8** | **3** | **10** |
| **4** | **8** | **8** | **9** | **6** |
| **5** | **8** | **5** | **6** | **7** |
| **6** | **7** | **5** | **6** | **8** |
| **7** | **6** | **7** | **4** | **5** |
| **8** | **9** | **6.5** | **7.5** | **7** |
| **9** | **8** | **9** | **7** | **6** |
| **10** | **8** | **6** | **7** | **5** |
| **总分** | **76** | **67.5** | **64.5** | **71** |
| **11** | **8** | **\** | **7** | **9** |
| **12** | **8** | **8（未生成代码，但思路清晰）** | **7** | **8.5** |

**五、分析讨论**

1、实验变量评估

（1）试题完全由Claude AI生成，力求客观，不受人为干扰，同时要求AI出题具有区分度；

（2）每个AI每回答一次问题，开启新主题reset，避免前一次问题对现有问题回答的干扰；

（3）Bard终端本身不识别中文，翻译成英文会对题目理解会有失偏颇，因为信息传递的过程会带来信息的增失，在翻译后我客观的稍加修改，以保持原意，因此可以忽略不计。（其中将博弈论翻译为Game Theory,游戏论翻译成Games Theory）

（4）Bard回答结果是英文，采用Google翻译，由于是同一家公司，因此我们可以认为信息增失为0；

（5）三次分数去掉最高最低，剩下是最终得分。

2、结果分析

最终这个结果让我有点出乎意料，但比较客观， Chat GPT依旧领先，文心一言排在最后，最让出乎意料的是Bard，之前报道+使用，Bard都不太行，但这次总分数排第二，New Bing排第三是因为自身的设定不同，它作为一个搜索引擎，可能在语言表达上还是稍差。

将此次测评分为三个维度，语言表达、知识储备、逻辑推理。

语言表达：Chat GPT 7.75 Bard 6.75 New Bing 6.13 文心一言 6.13；

知识储备：Bard 8.5 文心一言 7 Chat GPT 6.5 New Bing 6;

逻辑推理：Chat GPT 8 New Bing 7.8 Bard 6.75 文心一言 6.5

语言表达和逻辑推理两个维度排名正常合理，对于知识储备维度，我分析Bard和文心一言能够通过通过较少的语言得到更丰富的内容，但也从侧面印证了Bard和文心一言更像搜索而非生成，Chat GPT和New Bing更像生成和搜索，其实这款AI的知识储备应该都差不多，可能是因为Chat GPT和New Bing两款在参数设定上希望能用最少的语言表达清楚意思。

因此毋庸置疑，综合排名：

1. Chat GPT；
2. New Bing；
3. Bard；
4. 文心一言；

对于附加题第11题，上面的分析可以解答这个疑问；但是对于附加题第12

题Bard能超过Chat GPT，说明Goggle确实在努力训练，但也不排除本次考官Claude AI的母公司Anthropic是Google的孵化公司这一原因。

3、思考与感悟

（1）通过本次测试，更好地了解各个AI的优缺点，进一步用好AI，下一步我会深入了解其内在机理，开拓新思路；

（2）大语言模型都要各个语言都要互通，否则无法深入了解对话意图，造成不必要的麻烦；

（3）问AI，AI生成一个答案；再问AI答案这里不对，AI说确实不对，它知道，但需要训练，目前自己不具备主观自主学习，能感受AI的无力感；

（4）现在的AI和之前得搜索引擎一样都在要求人类如何提问题，如何把问题描述得更清楚、精准；

（5）由于Chat GPT我并未使用Plus版，可想Open AI真实实力如何，猜测Chat GPT应该是降低了部分代码生成能力，让利于自己的合作伙伴Microsoft的GitHub Copilot。