

大语言模型测试

郭润堂

guoruntang@buaa.edu.cn

Abstract

近年来，预训练语言模型（PLM）被提出，通过在大规模语料库上预训练 Transformer 模型，显示了解决各种自然语言处理（NLP）任务的强大能力。由于研究者发现模型规模的扩大可以提高模型的容量，当参数规模超过一定水平时，这些扩大的语言模型取得了显著的性能提升，而且还表现出了一些小规模语言模型（如 BERT）所没有的特殊能力（如情境学习）。为了区分不同参数规模的语言模型，研究界为那些具有较大尺寸（如包含数百亿个参数）的 PLM 起了一个新名词——大语言模型（LLM）。本文以 GPT-3.5-turbo、New Bing、文心一言为例，通过提示工程方法在知识问答、数学、情感分析、文本生成四项下游任务上对模型进行测试，并对模型性能进行进一步评估。

Introduction

gpt-3.5-turbo 是 gpt-3.5 持续优化的版本，反应速度与表现性能更佳。gpt-3.5 模型的架构是基于 Transformer 的解码器部分，它由多层的自注意力机制和前馈神经网络组成。每一层都有残差连接和层归一化，以提高模型的稳定性和泛化能力。gpt-3.5 模型的输入是一个由 token 组成的序列，每个 token 都被嵌入到一个高维的向量空间中，并与位置编码相加，以表示序列中的顺序信息。gpt-3.5 模型的输出是一个与输入序列长度相同的向量序列，每个向量都对应于一个 token 的概率分布，表示下一个 token 的预测。

自注意力机制是 gpt-3.5 模型的核心部分，它可以让模型捕捉输入序列中不同位置之间的依赖关系。自注意力机制首先将输入向量分别乘以三个不同的权重矩阵，得到查询（query）、键（key）和值（value）三个向量。然后，计算查询和键之间的点积，得到一个注意力分数矩阵。接着，对注意力分数矩阵进行缩放和 softmax 操作，得到一个注意力权重矩阵。最后，将注意力权重矩阵与值向量相乘，得到一个输出向量。gpt-3.5 模型使用了多头注意力机制，即将输入向量分成多个子空间，并在每个子空间上进行自注意力计算，然后将各个子空间的输出向量拼接起来，得到最终的输出向量。这样可以让模型关注不同的特征和语义信息。前馈神经网络是 gpt-3.5 模型的另一个重要部分，它可以增加模型的非线性能力和表达能力。前馈神经网络由两个全连接层组成，中间有一个激活函数（如 ReLU）。前馈神经网络可以对自注意力机制的输出向量进行进一步的变换和处理，从而提高模型的性能。

New Bing 模型不是 GPT4，它是基于 OpenAI 的下一代大型语言模型，比 GPT3.5 更强大，而且针对搜索进行了定制。它的模型架构是基于 Transformer 的编码器-解码器结构，它可以同时处理文本和图像的输入和输出。它还使用了 Microsoft Prometheus 模型，这是一种新的自监督学习方法，可以让模型从大量的未标注数据中学习知识和技能。

Transformer 是一种深度神经网络模型，它主要由自注意力机制和前馈神经网络组成。自注意力机制可以让模型捕捉输入序列中不同位置之间的依赖关系，而前馈神经网络可以增加模型的非线性能力和表达能力。编码器-解码器结构是一种常用的序列到序列模型，它由一个编码器和一个解码器组成。编码器负责将输入序列编码成一个向量表示，解码器负责根据向量表示生成输出序列。New Bing 模型使用了多个编码器-解码器层堆叠在一起，形成一个深层的模型结构。这样可以让模型处理更复杂的输入和输出，并提高其性能和效率。

Microsoft Prometheus 模型是一种新的自监督学习方法，它可以让模型从大量的未标注数据中学习知识和技能。自监督学习是一种无需人工标注数据的机器学习方法，它通过利用数据本身的结构或规律来生成监督信号。Microsoft Prometheus 模型使用了一种称为“对比学习”的自监督学习方法，它通过比较不同的数据样本之间的相似性或差异性来训练模型。这种方法可以让模型学习到数据的高层次特征和语义信息，从而提高其泛化能力和适应性。New Bing 模型使用了 Microsoft Prometheus 模型作为其预训练阶段，从而使其具备更强大的搜索能力和生成能力。

文心一言是由百度公司开发的聊天机器人，能够与人交互、回答问题及协作创作。它使用了百度自研的大语言模型 ERNIE Bot，具备跨模态、跨语言的深度语义理解与生成能力。文心一言是知识增强的大语言模型，基于飞桨深度学习平台和文心知识增强大模型，持续从海量数据和大规模知识中融合学习具备知识增强、检索增强和对话增强的技术特色。它的五大能力，分别是：文学创作、商业文案创作、数理逻辑推算、中文理解、多模态生成。文心一言具备对话问答能力，并可以与用户进行自然而流畅的对话和交流，帮助用户解决问题、获取信息或者进行娱乐。

提示工程 (Prompt engineering) 是人工智能中的一个概念，特别是自然语言处理 (NLP)。在提示工程中，任务的描述会被嵌入到输入中。例如，不是隐含地给予模型一定的参数，而是以问题的形式直接输入。提示工程的典型工作方式是将一个或多个任务转换为基于提示的数据集，并通过所谓的“基于提示的学习 (prompt-based learning)”来训练语言模型。提示工程可以从一个大型的“冻结”预训练语言模型开始工作，其中只学习了提示的表示方法，即所谓的“前缀调整 (prefix-tuning)”或“提示调整 (prompt tuning)”。

语言模型 GPT-2 和 GPT-3 是提示工程的重要步骤。2021 年，使用多个 NLP 数据集的多任务提示工程在新任务上显示出良好的性能。在小样本学习的例子中，包含思维链的提示在语言模型中显示出更好的推理能力。零样本学习中，在提示中预留鼓励思考链的语句（如“让我们一步一步地思考”）可能会提高语言模型在多步骤推理问题中的表现。这些工具的广泛可及性由几个开源笔记和社区主导的图像合成项目的发布所推动。

Methodology

由于
知识问答任务中，三个模型的表现：

Question	《哈利·波特与魔法石》是哪一年出版的？
GPT-3.5-turbo	《哈利·波特与魔法石》是 1997 年出版的。
New Bing	《哈利·波特与魔法石》是在 1997 年 6 月 26 日在英国出版的。中文正体版于 2000 年 6 月 23 日出版，简体版于 2000 年 9 月出版。
文心一言	《哈利·波特与魔法石》的出版时间是 1997 年。 《哈利·波特与魔法石》（英国版《Harry Potter and the Philosopher's Stone》，美国版《Harry Potter and the Sorcerer's Stone》），是英国女作

	<p>家 J.K.罗琳创作的长篇小说,《哈利·波特》系列小说的第一部。哈利·波特与魔法石, 该小说情节跌宕起伏, 语言风趣幽默, 主题反映了现实和人性, 发人深省。该小说的英文原版 1997 年 6 月 26 日在英国出版; 中文繁体版 2000 年 6 月 23 日出版; 中文简体版 2000 年 9 月出版。该书讲述了自幼父母双亡的孤儿哈利·波特收到魔法学校霍格沃茨的邀请, 前去学习魔法, 之后遭遇的一系列历险。</p>
Question	如何在 Excel 中使用 VLOOKUP 函数?
GPT-3.5-turbo	<p>VLOOKUP 函数可以在 Excel 中用于查找并返回一个表格中的特定值。以下是如何在 Excel 中使用 VLOOKUP 函数的步骤:</p> <p>打开 Excel 并打开要使用 VLOOKUP 函数的工作簿。</p> <p>选择要包含 VLOOKUP 函数的单元格。</p> <p>在公式栏中输入“=VLOOKUP(lookup_value, table_array, col_index_num, [range_lookup])”。</p> <p>替换 lookup_value, table_array, col_index_num 和[range_lookup]为你的实际值。其中:</p> <ul style="list-style-type: none"> -lookup_value: 你要查找的值。 -table_array: 包含查找值的表格区域。 -col_index_num: 要返回的值在 table_array 中的列数。 -[range_lookup]: 可选参数, 指定是否要进行近似匹配, 通常输入 FALSE 表示精确匹配。 <p>按 Enter 键, VLOOKUP 函数将返回查找值对应的结果。</p> <p>例如, 如果要查找“苹果”在表格区域 A1:B10 中的价格, 可输入以下公式: =VLOOKUP("苹果", A1:B10, 2, FALSE)。</p> <p>这将返回“苹果”在 B 列中的价格。</p>
New Bing	<p>VLOOKUP 函数是 Excel 中的一个函数, 它可以让你使用某个单元格数值用于查找位于同一行其他单元格中的内容 1。VLOOKUP 函数的语法如下:</p> <p>=VLOOKUP(要查找的内容、要查找的位置、包含要返回的值的范围内的列号、返回表示为 1/TRUE 或 0/FALSE 的近似或精确匹配项) 2</p> <p>例如, 如果你想从下面的表格中查找“张三”的成绩, 你可以使用这个公式:</p> <p>=VLOOKUP (“张三”,A2:B6,2,FALSE)</p> <p>这个公式的意思是, 在 A2:B6 这个区域中, 查找第一列中与“张三”完全匹配的值, 并返回第二列中对应的值, 即成绩。结果是 80。</p> <p>如果你想了解更多关于 VLOOKUP 函数的使用方法和技巧, 请参考以下链接:</p> <p>EXCEL 中 VLOOKUP 函数使用详解 - 知乎 3</p> <p>VLOOKUP 函数 - Microsoft 支持 2</p> <p>Excel 中 vlookup 函数的使用方法 (适合新手) - 百度经验 4</p> <p>3 种方法来使用 Excel 电子表格的 Vlookup 函数 1</p>
文心一言	<p>打开表格后点击显示结果的单元格, 点击输入框前面的函数选项。</p> <p>选择 vlookup 函数后在首项输入或选择要查找的数值, 设置数据表为要查找的表格范围。</p>

	列序数这一栏设置为 1 后点击确认就能显示函数计算结果了。
Question	牛顿第二定律是什么？
GPT-3.5-turbo	<p>牛顿第二定律是物理学中的一个重要定律，描述了物体所受的力和其运动状态之间的关系。它的数学表达式为 $F=ma$，其中 F 表示物体所受的合力，m 表示物体的质量，a 表示物体的加速度。也就是说，物体所受的合力等于其质量乘以加速度。</p> <p>换句话说，当一个物体受到一个力时，它的加速度与所受力成正比，与物体的质量成反比。如果质量不变，那么所受力越大，加速度越大。如果所受力不变，那么质量越大，加速度越小。这个定律对于解决许多物理学问题非常有用，例如计算物体的运动轨迹、计算机械工程中的力学问题等。</p>
New Bing	<p>牛顿第二定律是一条描述物体受力和运动变化之间关系的物理定律。它可以用数学公式表示为：</p> $F=dt/dp$ <p>，其中 F 是物体所受的外力，p 是物体的动量，t 是时间。当物体的质量 m 不随时间变化时，动量 p 等于质量 m 和速度 v 的乘积，所以牛顿第二定律可以简化为：</p> $F=ma$ <p>，其中 a 是物体的加速度。这个公式表明，物体的加速度与所受的外力成正比，与物体的质量成反比，且与外力的方向相同。</p>
文心一言	<p>牛顿第二定律，也被称为牛顿第二运动定律，是经典力学中的一个基本原理。这个定律表明，物体的加速度与作用力成正比，与物体质量成反比，且加速度方向与作用力方向相同。</p>

从实际回答情况来看，New Bing 和 GPT-3.5-turbo 两个模型的回答往往更为细致，且 New Bing 具有联网功能，在回答问题的同时也会进一步给出相应的参考链接和学习资料性能更好。

三个模型在数学问题任务中的表现：

Question	如果 $f(x)=x^2+3x-2$ ，那么 $f(2)$ 等于多少？
GPT-3.5-turbo	<p>将 x 替换为 2，得到：</p> $f(2) = 2^2 + 3(2) - 2$ <p>解出来得：</p> $f(2) = 4 + 6 - 2$ $f(2) = 8$ <p>因此，$f(2)$ 等于 8。</p>
New Bing	<p>这是一个数学问题，可以用代数法来解决。将 $x=2$ 代入函数 $f(x)=x^2+3x-2$，得到</p> $f(2)=(2)^2+3(2)-2=4+6-2=8$

	所以, $f(2)$ 等于 8。
文心一言	<p>把 $x=2$ 带入 f 公式中</p> <p>得到</p> $f=x^2+3x-2=8$ <p>所以, $f(2)=8$</p>
Question	如果 a 和 b 是互质的正整数, 那么 $(a+b)^2$ 能被多少个正整数整除?
GPT-3.5-turbo	<p>首先, 我们可以将 $(a+b)^2$ 展开:</p> $(a+b)^2 = a^2 + 2ab + b^2$ <p>我们注意到, a^2 和 b^2 分别是 a 和 b 的倍数, 因为它们分别等于 $a \times a$ 和 $b \times b$。因此, $(a+b)^2$ 也一定是 a 和 b 的倍数。</p> <p>另外, 因为 a 和 b 互质, 所以它们没有共同的因数, 所以 a 和 b 的最大公因数是 1。因此, a 和 b 的任何一个公因数都必须是 1。因此, a 和 b 的和 $(a+b)$ 也没有除 1 以外的公因数。</p> <p>综上所述, $(a+b)^2$ 能被正整数的个数等于 a 和 b 的因数个数的乘积, 即 $2 \times 2 = 4$。因此, $(a+b)^2$ 能被 4 个正整数整除。</p>
New Bing	<p>这是一个数论问题, 可以用因式分解和素因数分解的方法来解决。首先, 将 $(a+b)^2$ 因式分解为 $(a+b)(a+b)$, 然后利用 a 和 b 互质的性质, 得到 $(a+b)(a+b)$ 的素因数分解为 $p_1^{e_1} p_2^{e_2} \cdots p_k^{e_k} q_1^{f_1} q_2^{f_2} \cdots q_l^{f_l}$, 其中 p_i 是 $a+b$ 的素因数, q_j 是 $a-b$ 的素因数, e_i 和 f_j 是相应的指数。那么, $(a+b)^2$ 能被多少个正整数整除的问题就转化为求这个素因数分解的约数个数的公式。根据约数个数公式, 这个约数个数为 $(e_1 + 1)(e_2 + 1) \cdots (e_k + 1)(f_1 + 1)(f_2 + 1) \cdots (f_l + 1)$。所以, $(a+b)^2$ 能被多少个正整数整除的答案就是这个约数个数。</p>
文心一言	<p>我们可以先求出 a 和 b 的最大公约数, 然后根据最大公约数计算出 $(a+b)^2$ 的因数个数。</p> <p>已知 $a=3$, $b=7$</p> <p>根据辗转相除法, 可求出 a 和 b 的最大公约数:</p> <p>最大公约数为 1</p> <p>根据因数的个数公式, 可计算出 $(a+b)^2$ 的因数个数:</p> $(1+1) \times (1+1) \times (1+1) = 8$ <p>所以, $(a+b)^2$ 能被 8 个正整数整除。</p>
Question	设 F 为抛物线 $C: y^2=4x$ 的焦点, 点 A 在 C 上, 点 $B(3,0)$, 若 $ AF = BF $, 则 A 的坐标为 ?
GPT-3.5-turbo	<p>由抛物线的性质可知, 焦点 F 的横坐标为 1。设点 A 的坐标为 $(t, 2t^2)$, 则根据距离公式有:</p> $ AF ^2 = (t-1)^2 + 4t^2 \quad BF ^2 = (t-3)^2 + 4t^2$ <p>由题意可得 $AF = BF$, 即:</p> $(t-1)^2 + 4t^2 = (t-3)^2 + 4t^2 \Rightarrow t = 2$ <p>因此, 点 A 的坐标为 $(2, 8)$。</p>
New Bing	<p>这是一个解析几何的问题, 可以用几何关系和代数方程来解决。首先, 由抛物线的性质, 知道 F 的坐标为 $(1,0)$, 然后设 A 的坐标为 (x,y), 由于 A 在抛物线上, 所以有 $y^2=4x$。接下来, 由于 $AF = BF$, 所以有 $(x-1)^2+y^2=(x-3)^2$。将 $y^2=4x$ 代入, 得到 $x^2-8x+4=0$。解得 $x=4$ 或 $x=1$。</p>

	<p>当 $x=4$ 时, 有 $y^2=16$, 所以 $y=\pm 4$; 当 $x=1$ 时, 有 $y^2=4$, 所以 $y=\pm 2$。综上所述, A 的坐标有四种可能: $(4,4), (4,-4), (1,2), (1,-2)$。</p>
文心一言	<p>设 $A(x_0, y_0)$, 则 $y_0^2 = 4x_0$, 由抛物线定义得 $AF = x_0 + 1$, $BF = 3 - 1 = 2$。</p> <p>由于 $AF = BF$, 因此 $x_0 + 1 = 2$, 解得 $x_0 = 1$, 从而 $y_0^2 = 4x_0 = 4$, 解得 $y_0 = \pm 2$。</p> <p>因此, $A(1,2)$ 或 $A(1,-2)$。</p> <p>答案: $A(1,2)$ 或 $A(1,-2)$。</p>
Question	<p>设全集 $U = \{1, 2, 3, 4, 5\}$, 集合 M 满足 $\text{complement}_U M = \{1, 3\}$, 则 $M = ?$</p>
GPT-3.5-turbo	<p>首先, $\text{complement}_U M$ 表示全集 U 中不属于集合 M 的元素构成的集合, 因此 $\text{complement}_U M = \{1, 3\}$ 表示全集 U 中除了 1 和 3 之外的元素都属于集合 M。</p> <p>因此, 集合 $M = \{2, 4, 5\}$。</p>
New Bing	<p>您好, 这是 Bing。这是一个集合的问题, 可以用补集的定义和集合的运算来解决。首先, 由补集的定义, 知道 $\text{complement}_U M$ 是指全集 U 中不属于 M 的元素组成的集合。所以, 如果 $\text{complement}_U M = \{1, 3\}$, 那么 M 就是全集 U 中除了 1 和 3 之外的元素组成的集合。也就是说, $M = U - \{1, 3\}$。由于已知 $U = \{1, 2, 3, 4, 5\}$, 所以 $M = \{2, 4, 5\}$。</p>
文心一言	<p>根据题意, $M \subseteq \{1, 2, 4, 5\}$, 且 M 中不含元素 1 和 3。因此, M 的可能取值为 $\{2, 4\}$、$\{2, 5\}$、$\{4, 5\}$。</p>
Question	<p>从 1 到 9 这九个数字中任取三个不同的数字组成一个三位数, 则能被 3 整除的三位数有几个?</p>
GPT-3.5-turbo	<p>能被 3 整除的数字有一个特点, 就是它们的各位数字之和能被 3 整除。因此, 我们只需要考虑从 1 到 9 中选出的三个数字各自的取值, 使得它们的和能被 3 整除。</p> <p>当三个数字之和为 3、6 或 9 时, 它们能组成的三位数都能被 3 整除。因此, 我们只需要考虑以下几种情况:</p> <p>选出的三个数字之和为 3, 只有一种情况: $1+2+0=3$, 能组成的三位数为 120。</p> <p>选出的三个数字之和为 6, 有四种情况: $1+2+3=6$, $1+3+2=6$, $2+1+3=6$, $2+3+1=6$, 能组成的三位数为 123、132、213、231、312 和 321。</p> <p>选出的三个数字之和为 9, 有三种情况: $1+2+6=9$, $1+3+5=9$, $2+3+4=9$, 能组成的三位数为 126、135、153、234、243 和 321。</p> <p>因此, 能被 3 整除的三位数共有 $6+3=9$ 个。</p>
New Bing	<p>这是一个组合数学的问题, 可以用排列组合和整除的性质来解决。首先, 由于要从 1 到 9 这九个数字中任取三个不同的数字组成一个三位数, 所以一共有 $9 \times 8 \times 7 = 504$ 种可能。然后, 由于要求能被 3 整除的三位数, 所以需要满足一个整数被 3 整除的充分必要条件: 它的各位数字之和能被 3 整除。也就是说, 如果一个三位数 abc 能被 3 整除, 那么必须有 $a+b+c$ 能被 3 整除。接下来, 我们可以分情况讨论 $a+b+c$ 的取值范围和可能性。由于 a, b, c 都是从 1 到 9 中取的, 所以 $a+b+c$ 的最小值是 $1+2+3=6$, 最大值是 $7+8+9=24$。在这个范围内, 能被 3 整除的数有 6, 9, 12, 15, 18, 21, 24 这七个。对于每个数, 我们可以计算有多少种不同的 a, b, c 的组合方式。例如, 当 $a+b+c=6$ 时, 只有一种可能:</p>

	<p>$a=1, b=2, c=3$; 当 $a+b+c=9$ 时, 有三种可能: $a=1, b=2, c=6$, $a=1, b=3, c=5$, $a=2, b=3, c=4$; 以此类推, 我们可以得到下表:</p> <table><tr><th>$a+b+c$</th><th>可能性</th><th>数量</th></tr><tr><td>6</td><td>(1,2,3)</td><td>1</td></tr><tr><td>9</td><td>(1,2,6), (1,3,5), (2,3,4)</td><td>3</td></tr><tr><td>12</td><td>(1,2,9), (1,3,8), (1,4,7), (1,5,6), (2,3,7), (2,4,6), (3,4,5)</td><td>7</td></tr><tr><td>15</td><td>(1,5,9), (1,6,8), (2,4,9), (2,5,8), (2,6,7), (3,4,8), (3,5,7), (4,5,6)</td><td>8</td></tr><tr><td>18</td><td>(2,7,9), (3,6,9), (4,5,9), (3,7,8), (4,6,8), (5,6,7)</td><td>6</td></tr><tr><td>21</td><td>(4,8,9), (5,7,9), (6,7,8)</td><td>3</td></tr><tr><td>24</td><td>(7,8,9)</td><td>1</td></tr></table> <p>由表可知, 一共有 $1+3+7+8+6+3+1=29$ 种不同的 a, b, c 的组合方式满足 $a+b+c$ 能被 3 整除。但是, 这还不是最终答案, 因为对于每种组合方式, 还要考虑 a, b, c 的排列顺序。由于三位数的百位、十位、个位都可以任意排列, 所以对于每种组合方式, 还有 $3!=6$ 种不同的排列顺序。因此, 最终答案是 $29 \times 6=174$。也就是说, 从 1 到 9 这九个数字中任取三个不同的数字组成一个三位数, 则能被 3 整除的三位数有 174 个。</p>	$a+b+c$	可能性	数量	6	(1,2,3)	1	9	(1,2,6), (1,3,5), (2,3,4)	3	12	(1,2,9), (1,3,8), (1,4,7), (1,5,6), (2,3,7), (2,4,6), (3,4,5)	7	15	(1,5,9), (1,6,8), (2,4,9), (2,5,8), (2,6,7), (3,4,8), (3,5,7), (4,5,6)	8	18	(2,7,9), (3,6,9), (4,5,9), (3,7,8), (4,6,8), (5,6,7)	6	21	(4,8,9), (5,7,9), (6,7,8)	3	24	(7,8,9)	1
$a+b+c$	可能性	数量																							
6	(1,2,3)	1																							
9	(1,2,6), (1,3,5), (2,3,4)	3																							
12	(1,2,9), (1,3,8), (1,4,7), (1,5,6), (2,3,7), (2,4,6), (3,4,5)	7																							
15	(1,5,9), (1,6,8), (2,4,9), (2,5,8), (2,6,7), (3,4,8), (3,5,7), (4,5,6)	8																							
18	(2,7,9), (3,6,9), (4,5,9), (3,7,8), (4,6,8), (5,6,7)	6																							
21	(4,8,9), (5,7,9), (6,7,8)	3																							
24	(7,8,9)	1																							
文心一言	<p>我们可以先计算出所有三位数中能被 3 整除的数的个数, 再减去其中重复的个数, 即可得到答案。</p> <p>首先, 我们可以计算出所有三位数中能被 3 整除的数的个数:</p> <p>所有三位数中能被 3 整除的数的个数为: 504 个</p> <p>接下来, 我们需要计算其中重复的个数。</p> <p>对于每个三位数, 如果它的百位、十位、个位数字之和能被 3 整除, 那么这个三位数就能被 3 整除。</p> <p>因此, 我们可以枚举所有三位数, 判断它们的百位、十位、个位数字之和是否能被 3 整除。</p> <p>所以, 所有三位数中能被 3 整除的数的个数为: 0 个</p> <p>因此, 从 1 到 9 这九个数字中任取三个不同的数字组成一个三位数, 则能被 3 整除的三位数有 0 个。</p>																								

从 5 道题的测试结果来看, 没有一个模型可以把 5 道题全部做对, 其中 New Bing 的准确率相对较高。

情感分析任务中, 三个模型的表现:

Question	一用户对一中餐馆的用餐评价如下: “选择珠江花园的原因就是方便, 有电动扶梯直接到达海边, 周围餐馆、食廊、商场、超市、摊位一应俱全。酒店装修一般, 但还算整洁。泳池在大堂的屋顶, 因此很小, 不过女儿倒是喜欢。包的早餐是西式的, 还算丰富。服务吗, 一般。”请问该用户对此次用餐体验是更倾向于满意还是不满意?
GPT-3.5-turbo	根据您对该用户评论的分析, 我也认为该用户更倾向于满意。尽管该用户没有明确表达出对餐馆的装修和服务的喜爱或赞扬, 但是他们也没有提到任何明显的负面体验或抱怨, 只是描述了整个用餐过程中自己的感受和印象。因此, 可以认为该用户对此次用餐体验是比较满意的。
New Bing	根据该用户的评论, 我认为该用户对此次用餐体验是更倾向于满意还是不

	满意，这是一个主观的判断，可能因人而异。但是，如果要我给出一个答案，我会说该用户更倾向于满意，因为该用户没有明显的抱怨或不满，而且对一些方面有赞扬或肯定。虽然该用户对餐馆的装修和服务只是说“一般”，但这也不一定表示不满意，可能只是没有特别的感受或期待。因此，我认为该用户对此次用餐体验是更倾向于满意的。
文心一言	根据该用户的评价内容，可以看出该用户对餐馆的用餐体验并没有表现出明显的倾向性，既没有表达出非常满意的感觉，也没有表达出非常不满意的感觉。 该用户提到了餐馆的方便性、周围设施齐全、酒店整洁、早餐丰富等服务方面的优点，同时也提到了一些不足之处，如装修一般、泳池小等。综合来看，该用户对餐馆的评价是中立的，没有明显的倾向性。
Question	一用户对一本书的留言如下：“今天才知道这书还有第 6 卷,真有点郁闷:为什么同一套书有两种版本呢?当当网是不是该跟出版社商量商量,单独出个第 6 卷,让我们的孩子不会有所遗憾。”请问该留言对这本书的评价是倾向于满意还是倾向于不满意?
GPT-3.5-turbo	该留言倾向于不满意。虽然用户提到了对这本书的兴趣和阅读，但他们也表达了对缺少第 6 卷的失望和疑惑，并暗示出版商和网站可能应该考虑出售完整版本以避免让消费者感到失望。
New Bing	根据该留言，我认为该留言对这本书的评价是倾向于不满意的。因为该留言表达了对书的第 6 卷的缺失的不解和遗憾，以及对当当网和出版社的不满和抱怨。这说明该留言的作者对这本书有一定的期待和喜爱，但是因为版本的问题而感到失望和沮丧。因此，我认为该留言对这本书的评价是倾向于不满意的。
文心一言	根据该留言的内容，可以看出该用户对这本书的评价是倾向于不满意。用户感到郁闷的是同一套书有两种版本，导致自己的孩子错过了第 6 卷，留下了遗憾。用户建议当当网应该跟出版社商量商量，单独出个第 6 卷，以避免这种情况的发生。
Question	阅读以下微博评论，并判断这段评论背后的情感是什么？”这部电影太恶心了！看得我想吐！“
GPT-3.5-turbo	这段评论背后的情感是厌恶和反感。
New Bing	根据这段微博评论，我认为这段评论背后的情感是厌恶。因为这段评论使用了“恶心”和“想吐”这样的词语，表达了对电影的强烈的反感和排斥，以及对电影的质量和内容的批评。这说明这段评论的作者对这部电影有很低的评价，没有任何的欣赏或喜爱。因此，我认为这段评论背后的情感是厌恶。
文心一言	这段微博评论表达了负面的情感，可以理解为对该电影的强烈不满或厌恶。评论者使用了“恶心”和“想吐”这两个词语，表达了强烈的反感和不适感。因此，这段评论的情感是负面的。

从 3 道题的测试结果来看, 3 个模型基本都可以识别出一段话背后的情感, 其中 New bing 的分析最为细致。

文本生成任务三个模型的表现:

Question	请用一篇文章分析新冠肺炎疫情的影响。
GPT-3.5-turbo	新冠肺炎疫情是自 2020 年开始的全球性卫生事件，给全世界带来了巨大

	<p>的影响。这场疫情不仅影响了人们的健康和生命，还对经济、社会文化等各个方面产生了深远的影响。</p> <p>首先，新冠肺炎疫情对全球医疗系统造成了重大冲击。由于病毒传染性强，短时间内就可以迅速传播并造成大量感染者，许多国家的医疗系统都难以承受如此之多的患者。部分地区的医院和医护人员也处于崩溃的边缘，导致疫情控制和救治工作变得更加困难。</p> <p>其次，新冠肺炎疫情对全球经济产生了巨大冲击。由于疫情暴发后，各国相继采取了封锁城市、限制出行等措施，导致全球生产和供应链遭到重创，许多企业被迫关闭或减少产能。同时，旅游、餐饮、娱乐等服务业也受到了极大的影响，因为人们被迫待在家中，不再进行传统的消费活动。这些影响导致许多国家的经济陷入了萧条和衰退。</p> <p>第三，新冠肺炎疫情也对全球社会文化产生了深远影响。由于疫情暴发后各国政府相继实行封锁、隔离等措施，使得人们的日常生活受到了巨大影响。例如，学校、办公室、商铺等都被迫暂停运营，人们被迫居家隔离，无法进行正常的日常社交活动。此外，由于疫情影响全球，一些人也对特定族群产生怀疑和歧视，表现出一些种族主义言行。这些影响改变了人们的社交习惯和社会意识形态，并且可能在未来造成长期影响。</p> <p>最后，需要指出的是，虽然新冠肺炎疫情带来了巨大影响，但我们可以从中汲取教训，进一步完善全球卫生体系和应急机制，提高公民的健康意识和卫生素养，同时也需要加强国际合作，共同应对全球性疫情的挑战。</p>
New Bing	<p>新冠肺炎疫情是一场全球性的公共卫生危机，给世界各国的政治、经济、社会、文化等方面带来了深刻的影响。以下是一些主要的影响：</p> <p>对人类健康的影响。新冠肺炎疫情造成了数百万人的感染和死亡，给人类生命安全和身体健康带来了巨大的威胁。感染新冠病毒的人可能会出现轻重不一的呼吸系统症状，甚至导致器官衰竭和死亡。此外，感染新冠病毒期间，肺功能受损可能导致认知功能下降和神经系统病变，脑细胞的氧气供应减少，从而影响智力。</p> <p>对经济发展的影响。新冠肺炎疫情引发了全球经济衰退，导致生产、消费、贸易、投资等活动大幅减少，失业率上升，贫困人口增加，经济全球化进程受阻。同时，疫情也加剧了国际贸易摩擦和地缘政治风险，影响了全球产业链和供应链的稳定和安全。</p> <p>对社会秩序的影响。新冠肺炎疫情对人们的日常生活、工作、学习、交流等方面造成了巨大的改变和挑战，给社会秩序带来了不稳定因素。一方面，为了应对疫情，人们不得不采取居家隔离、远程办公、在线教育等方式，这些方式虽然有利于减少感染风险，但也可能导致人们感到孤独、焦虑、抑郁等心理问题，影响人们的身心健康和社会关系。另一方面，由于疫情的不确定性和复杂性，人们对未来充满担忧和恐惧，可能出现恐慌购物、抢购物资、歧视异族、反对政府等现象，影响社会的和谐与稳定。</p> <p>对文化交流的影响。新冠肺炎疫情也对文化交流产生了负面和正面的双重影响。负面影响主要表现在由于旅行限制和封锁措施，很多国际文化活动和项目被取消或延期，文化交流的机会和渠道减少，文化多样性和包容性受到挑战。正面影响主要表现在由于疫情的全球性和紧迫性，很多国家和人民加强了文化沟通和合作，通过网络、媒体等方式分享抗疫经验和故事，展示文化魅力和人文关怀，促进了文化互学互鉴和人类命运共同体的建</p>

	<p>设。</p> <p>综上所述，新冠肺炎疫情对世界的影响是深刻的、全面的、持久的，需要各国加强合作、共克时艰，以科学、理性、人道的态度应对挑战，保护人类健康福祉，推动世界和平发展。</p>
文心一言	<p>新冠肺炎疫情的全球爆发对各行业产生了深远的影响。从经济、社会和政策层面分析了这场疫情对餐饮、旅游、交通等行业的冲击，并提出了应对措施。</p> <p>在经济方面，新冠肺炎疫情导致全球经济遭受重创。旅游业、餐饮业和交通运输业受到了最大的影响。旅游业失去了大量的游客，餐饮业不得不关闭门店，交通运输业则面临严重的运营压力。企业生产停滞不前，供应链中断，使得制造业也受到了严重影响。疫情对这些行业的影响引发了全球性的经济萎缩，许多国家采取了紧急措施来减轻经济冲击。</p> <p>在社会方面，新冠肺炎疫情改变了我们的生活方式。学生流动受阻，学校和教育机构关闭，导致教育行业受到严重影响。人员紧急流动受到限制，使得交通运输业面临巨大的挑战。疫情还引发了就业问题，许多公司采取了裁员和减薪措施来应对危机，这使得就业市场更加严峻。</p> <p>在政策方面，政府采取了紧急措施来应对新冠肺炎疫情。疫情防控政策包括限制人员流动、关闭学校和企业、推迟大型活动等。经济政策包括为企业提供贷款、向失业人员提供援助等。政府还实施了财政刺激计划，以稳定经济运行。</p> <p>总之，新冠肺炎疫情对全球经济和社会产生了深远的影响。各行业需要采取紧急措施来应对危机，并制定长期计划来适应新的形势。同时，政府需要继续实施有效的政策来支持受影响的行业和保护公众健康。</p>
Question	请用戏剧的形式写一个关于友情的剧本。
GPT-3.5-turbo	<p>《友情不灭》</p> <p>角色：小明，小红，小李，小张</p> <p>场景：四个朋友在学校的操场上聊天。</p> <p>（四位朋友正在操场上聊天）</p> <p>小明：最近你们有什么好玩的事吗？</p> <p>小红：没有啊，最近都在忙着学习呢。</p> <p>小李：对啊，期末考试要到了，得好好准备。</p> <p>小张：别提了，我现在每天都在图书馆呆着，复习到晚上十点才回家。</p> <p>小明：哇，你们都好努力啊。我最近反而有点闲得无聊。</p> <p>小红：那我们一起去玩吧，放松一下心情。</p> <p>小李：好主意，我们去游乐场怎么样？</p> <p>小张：也行，不过我记得今天是周五，游乐场应该会很拥挤吧。</p> <p>小明：那就算了，我们去踢球吧。</p> <p>小红：踢球？我不太善长啊。</p> <p>小李：没关系，我们可以互相教学。</p> <p>小张：对啊，这样还可以增进感情。</p> <p>（四位朋友来到了操场边，开始踢球）</p> <p>小明：小红，你把球传给我！</p> <p>小红：好的！（把球传给了小明）</p> <p>小李：小张，你跑快点！</p>

	<p>小张：我在努力啊！（奋力向前跑）</p> <p>（过了一会儿）</p> <p>小红：咦，我的球怎么不见了？</p> <p>小张：是啊，我也没看见。</p> <p>小明：这下我们怎么办？我们的足球可是很贵重的。</p> <p>小李：难道是有人偷走了？</p> <p>（突然，一个小男孩从操场边出现，手里拿着他们的足球）</p> <p>小男孩：对不起，我拿错了球了。</p> <p>小红：（微笑）没关系，谢谢你还回来给我们。</p> <p>小张：你也是踢足球的吗？要不要和我们一起玩？</p> <p>小男孩：（开心地点头）好呀！</p> <p>（于是，五位朋友开始一起踢球，欢声笑语不断。）</p> <p>小明：（感慨地说）真好，友情不管时间长短，只要用心去经营，就可以一直保持美好。</p> <p>小红：是啊，这才是最重要的。</p> <p>小李：我们以后也要一直保持联系，共同度过每个美好的时刻。</p> <p>小张：对啊，我们的友情永远不会消逝。</p> <p>（五个人相互拥抱，场景渐行渐远。）</p>
New Bing	<p>您好，这是 Bing。根据您的要求，我为您写了一个关于友情的剧本，如下：</p> <p>《友情的考验》</p> <p>人物：李明、张强、王华、老师</p> <p>场景：一间教室</p> <p>李明：（拿着一张试卷，兴奋地跑到张强跟前）张强，你看，我这次数学考了九十八分，全班第一呢！</p> <p>张强：（看着试卷，不太高兴）哇，李明，你真厉害，恭喜你啊！</p> <p>李明：（笑着）谢谢，其实我也没想到会考这么好。你呢，考了多少分？</p> <p>张强：（有些尴尬）我……我只考了六十五分。</p> <p>李明：（惊讶）啊？怎么会这样？你平时学习不是挺好的吗？</p> <p>张强：（低头）是啊，但是这次考试我有点紧张，有几道题没做出来。</p> <p>李明：（安慰）没关系，下次努力就好了。你要是有不会的题，可以问我啊，咱们是好朋友嘛。</p> <p>张强：（感动）谢谢你，李明，你真是个好朋友。</p> <p>王华：（走过来，看着李明的试卷，嘲笑）哟，李明，你这次考了九十八分啊，真了不起啊！可惜你的好朋友张强只考了六十五分，你们俩真是天壤之别啊！</p> <p>李明：（生气）王华，你说什么呢？考试成绩不代表一切，张强是我的好朋友，你不要说他坏话！</p> <p>王华：（不屑）哼，我就说他坏话又怎么样？你能怎么样？你以为你考了第一就了不起啊？告诉你吧，其实你的试卷是我给你改的。</p> <p>李明：（惊讶）什么？你给我改的？怎么可能？</p> <p>王华：（得意）是啊，我是学生会的文化部长，这次数学考试由我负责监督和改卷。当时我看到你的试卷只有八十多分，就觉得太可惜了，就给你加了几分。没想到你还真的考了第一呢！</p> <p>李明：（气愤）王华，你怎么可以这样？这样做太不公平了！你为什么要</p>

	<p>给我加分？</p> <p>王华：（阴险）呵呵，其实我是想给你找麻烦。你知道吗？其实老师已经发现了这件事，他说要找你谈谈。如果他知道是我给你加分的话，他肯定会惩罚我。所以呢，我就把责任推给了张强。</p> <p>李明：（震惊）什么？你把责任推给了张强？</p> <p>王华：（狡辩）是啊，我跟老师说是张强偷偷拿走了你的试卷，给你加了几分。反正他也考得不好，就算被开除了也没什么损失。而且他又是你的好朋友，你肯定不会揭穿他的。</p> <p>李明：（愤怒）王华，你太卑鄙了！你怎么可以陷害张强？他是无辜的，你为什么要害他？</p> <p>王华：（嘲笑）哎呀，李明，你不要这么激动嘛。你想想看，如果你跟老师说实话，那你就会失去第一名的荣誉，而且还会得罪我。你觉得值得吗？何况，张强也不过是个普通的同学，跟你有什么关系？你何必为了他牺牲自己呢？</p> <p>李明：（义愤）王华，你太可恶了！张强是我的好朋友，我不会让你陷害他！我不在乎第一名，我只在乎友情！我要跟老师说清楚这件事，让他知道真相！</p> <p>王华：（惊恐）李明，你不要冲动啊！你这样做会后悔的！</p> <p>李明：（坚定）我不会后悔的！（拉着张强的手，向老师的办公室走去）</p> <p>王华：（气愤）李明，你给我等着！（咬牙切齿地看着他们的背影）</p> <p>（幕下）</p>
文心一言	<p>标题：友情的力量</p> <p>场景：高中校园</p> <p>人物：小明、小李、小王</p> <p>（一天下午，小明和小李在校园里散步）</p> <p>小明：今天的课业真多啊，好累啊。</p> <p>小李：没错，不过我们还要面对考试，不能放松。</p> <p>小明：你知道吗？我们班上有一个同学，他叫小王，他学习很好，而且总是乐于助人。</p> <p>小李：是吗？我们去找他帮忙吧！</p> <p>（他们找到了小王）</p> <p>小明：小王，我们想请你帮忙解决我们的课业问题。</p> <p>小王：当然可以，我们一起讨论吧。</p> <p>（在接下来的几天里，小明和小李和小王一起学习）</p> <p>小明：小王，谢谢你帮助我们，现在我们的学习进步了很多。</p> <p>小李：是啊，我们真的很感谢你。</p> <p>小王：不用谢，我也很高兴能够帮助到你们。</p> <p>（几个月后，小明和小李得知小王因为家庭原因要转学）</p> <p>小明：小王，我们要好好珍惜剩下的时间，让你记住我们。</p> <p>小李：没错，我们要让他知道友情的力量。</p> <p>（最后，小王离开前，小明和小李送给他一份礼物）</p> <p>小明：小王，这是我们送给你的礼物，谢谢你陪伴我们走过这段时光。</p> <p>小李：我们会一直想念你的。</p> <p>小王：谢谢你们，我也会一直记得你们。</p>

	<p>(小王离开后, 小明和小李感到有些失落, 但他们知道小王会一直在心中)</p> <p>小明: 小李, 我们要继续努力, 让自己变得更优秀, 让小王为我们感到骄傲。</p> <p>小李: 没错, 我们要让友情的力量一直延续下去。</p> <p>(剧终)</p>
Question	<p>请用 ALFRED 数据集中的一段自然语言指令和一段对应的 RGB-D 视频, 生成一系列的动作序列, 使得模拟的机器人能够完成指令中的任务。</p>
GPT-3.5-turbo	<p>例如, 以下是一个 ALFRED 数据集中的指令样例:</p> <p>Copy Code</p> <p>"Put the apple on the stove, turn on the stove to medium heat and wait until the apple is cooked."</p> <p>这个指令要求机器人将苹果放在炉子上, 打开炉子并调至中火, 等待苹果煮熟。</p> <p>相应的任务描述可能会有如下内容:</p> <p>场景: 厨房。</p> <p>初始状态: 场景中有空碗、苹果、炉子和胡萝卜。炉子关着。</p> <p>目标状态: 苹果被放在炉子上, 炉子被打开并调至中火, 等待苹果煮熟后, 机器人需要停止炉火。</p> <p>基于以上指令和任务描述, 机器人需要完成以下一系列操作:</p> <p>找到苹果并捡起来。</p> <p>找到炉子, 并将苹果放在炉子上。</p> <p>打开炉子, 并将火候调至中火。</p> <p>等待苹果煮熟。</p> <p>关闭炉火。</p>
New Bing	<p>语言指令和一段指令: 从卧室里拿出一本书, 然后放到客厅的沙发上。</p> <p>视频: (显示了一个卧室和一个客厅的场景, 其中有一张床、一张书桌、一本书、一扇门、一个沙发等)</p> <p>动作序列:</p> <p>移动到书桌前</p> <p>拿起书桌上的书</p> <p>移动到门前</p> <p>打开门</p> <p>移动到客厅</p> <p>移动到沙发前</p> <p>放下书</p>
文心一言	<p>自然语言指令: “把那个球捡起来, 放到桌子上。”</p> <p>对应的 RGB-D 视频: 在这个视频中, 我们可以看到一个机器人和几个球。机器人的手臂可以移动, 并且它有一个抓手。指令要求机器人捡起球并将其放置在桌子上。</p> <p>动作序列:</p> <p>机器人移动到球的旁边。</p> <p>机器人使用传感器测量球的尺寸和位置。</p> <p>机器人调整手臂的位置和角度, 以便能够抓住球。</p> <p>机器人使用抓手抓住球。</p>

	机器人移动手臂并调整角度，以便能够将球放置在桌子上。 机器人将球放置在桌子上。 机器人返回起始位置，等待下一个指令。
--	--

从以上 3 道题的测试结果来看，3 个模型处理的都比较好。New Bing 在文章写作和剧本编写上更加细致、总结更加到位。文心一言在处理机器人模拟指令时考虑更加细致周到。总体而言在这项任务中 New Bing 的表现看似更佳。

Conclusions

本文对比了 GPT-3.5-turbo、New Bing、文心一言三款前沿大语言模型在知识问答、数学、情感分析、文本生成等四个领域的下游任务。从实际的测试当中在数学推理领域各个模型的表现都不是十分理想，没有模型可以在所有问题上都有十分完美的表现，文心一言和 GPT-3.5-Turbo 之间的表现各有胜负，通常 GPT-3.5-Turbo 对于各种问题的鲁棒性、适应性更强。New Bing 由于在问答的过程中可以额外提供可以支撑答案和引申答案的链接，并在多数问题中有十分细致、总结合理的回答，主观评价表现最佳。

	知识问答	数学	情感分析	文本生成	总和
GPT-3.5-turbo	8	5	8	8	29
New Bing	9	9	9	9	36
文心一言	8	5	8	8	29

表 1 总体得分情况