

# 作业5 前沿语言大模型性能测试

19376463 员佳程

## Abstract

任意选取3~4个目前前沿语言大模型（不限中文或者英文），通过提示工程的方法来检验和对比不同模型下游任务上的性能，可选择3~5个不同的自然语言下游任务来进行测试。

## Introduction

自然语言处理大模型是指利用深度学习技术训练出的能够理解和生成自然语言的模型。这些模型通常基于预训练的语言模型，如GPT（Generative Pre-trained Transformer），BERT（Bidirectional Encoder Representations from Transformers）等。其中，ChatGPT是一个基于GPT的聊天机器人，它能够根据用户输入的文本生成相应的回复，并具有一定的上下文感知能力，可以进行流畅的对话。另外，文心一言是一款由百度公司开发的中文文本生成模型，它采用了类似GPT的架构，并在大规模中文语料上进行了预训练，能够生成符合语法和语义规则的高质量中文文本。此外，GPT3.5是ChatGPT的基础模型框架，我们也用它与其他的模型进行比较。这些大模型的出现，为自然语言处理领域的发展提供了强有力的支持，为各种语言应用的实现提供了更加高效、准确的解决方案。

## Methodology

### 一、前沿语言大模型介绍

#### 1. ChatGPT

ChatGPT，全名为Chat Generative Pre-trained Transformer，是一种基于Transformer架构的深度学习模型，用于处理序列数据，主要用于自然语言处理任务，如文本分类、问答、机器翻译等等。

ChatGPT是由美国OpenAI公司开发的，该公司是由特斯拉CEO埃隆·马斯克、美国创业孵化器Y Combinator总裁阿尔特·芒格等人于2015年创立，旨在研究人工智能技术，并推动人工智能技术的安全、负责任发展。

ChatGPT是一种大型的语言模型，它通过对大量文本数据的学习，能够生成人类可读的文本或理解人类输入的文本。它能够处理各种自然语言任务，如文本分类、问答、对话等等。由于其预训练的模型和强大的生成能力，ChatGPT能够生成高质量的文本，从而在各种自然语言处理任务中表现出色。

ChatGPT的应用非常广泛，如在智能客服、推荐系统、机器翻译、对话系统等领域都有着广泛的应用。它能够帮助人们更快速、更准确地理解语言，并进行有效的沟通。在教育领域，ChatGPT还可以帮助学生更好地理解知识、提高学习效率。在医疗领域，ChatGPT可以用于智能问诊、健康建议等方面，为医生提供辅助诊断的参考意见。

尽管ChatGPT在自然语言处理领域取得了巨大的成功，但它仍然存在一些局限性。例如，它可能会出现语义理解错误、回答不准确等问题。此外，由于其需要大量的计算资源和数据存储，其训练和使用成本也相对较高。

总之，ChatGPT是一种强大的人工智能技术，它在自然语言处理领域有着广泛的应用和前景。它可以帮助人们更好地理解和交流，从而提高生活质量和工作效率。然而，我们也需要认识到其局限性和潜在的风险，并采取相应的措施来确保其安全和负责任的使用。

## 2. 文心一言

文心一言是百度研发的知识增强大语言模型，自2023年3月16日起，在百度内部进行广泛的测试，并在3月16日进行首批内测。它能够与人对话互动，回答问题，协助创作，高效便捷地帮助人们获取信息、知识和灵感。

文心一言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。

在产品体验方面，文心一言生成文本的领域涵盖了编程语言、自然语言处理、机器视觉等AI核心领域，场景覆盖了人工智能日常研发、教育、医疗、金融、商业创意、文娱等场景。除了文字，文心一言还支持图片、视频、音频、表格、知识的输出。在人机交互过程中，用户可以随时提出需求，文心一言会根据用户的需求提供相应的信息、答案或执行任务的结果。

在应用方面，文心一言已经在百度内部进行了广泛的测试，并在多个领域展现了强大的功能。例如，在文学创作领域，文心一言可以帮助作家进行创作，提高创作效率，减少创作成本；在商业创意领域，文心一言可以根据用户的需求，快速生成多种形式的创意内容。

总之，文心一言是一款强大的人工智能大模型，具有丰富的知识储备和强大的对话能力，能够为人类提供各种便利和服务。

### 3. GPT3.5

GPT-3.5是OpenAI训练的一种预训练语言模型，是GPT-3的升级版。GPT-3.5在文本分类、问答、翻译等领域有广泛应用，例如可以用于自动翻译、文本生成等任务。CHATGPT是GPT-3.5的改进版，通过引入人类反馈的强化学习来提高对话生成能力和上下文理解能力。与GPT-3.5相比，CHATGPT具有更好的对话生成能力和更丰富的上下文理解，能够更准确地理解和回答用户的问题。

## 二、自然语言处理的下游任务

自然语言处理还涉及多个下游任务，包括：

1. 机器翻译：将一种语言的文本翻译成另一种语言的文本。
2. 命名实体识别：从文本中识别出人名、地名、组织机构名等命名实体。
3. 情感分析：分析文本中的情感倾向，如正面、负面或中性。
4. 文本分类：将文本分为不同的类别，如新闻、评论、广告等。
5. 问答系统：根据用户提出的问题，回答相应的问题。
6. 关键词提取：从文本中提取出关键词。
7. 文本生成：给出提示，完成一段文本的创作。

这些下游任务都是自然语言处理的重要应用方向，涵盖了自然语言处理的多个方面。在实际应用中，我们可以根据具体需求选择相应的任务和模型，进行相应的训练和优化，实现更加高效、准确的自然语言处理应用。

# Experimental Studies

## 一、实验内容

本文选择了机器翻译、情感分析、问答系统、文本生成四个自然语言处理的下游任务对三种大模型进行了测试，采用提示工程的方法来检验它们的性能。

### 1. 机器翻译

在机器翻译任务中，我们将一些中国古典诗词或名句翻译成英文，要求语言模型将他们翻译回中文。以下是所用的中英文预料，其中第（1）句作为提示给到模型。

（1）但愿人长久，千里共婵娟。We wish each other a long life so as to share the beauty of this graceful moonlight, even though miles apart.

（2）独在异乡为异客，每逢佳节倍思亲。A lonely stranger in a strange land I am cast, I miss my family all the more on every festive day.

（3）大江东去，浪淘尽，千古风流人物。The endless river eastward flows; with its huge waves are gone all those gallant heroes of bygone years.

（4）二人同心，其利断金。If two people are of the same mind, their sharpness can cut through metal.

（5）富贵不能淫，贫贱不能移，威武不能屈，此之谓大丈夫。It is a true great man whom no money and rank can confuse, no poverty and hardship can shake, and no power and force can suffocate.

文心一言：



chatGPT（ChatBot）：

1. 异乡独客心孤独，佳节思亲倍切肺。
2. 东流无尽水潺潺，昔日英雄尽随波。
3. 心意相通，两人一心，利刃可断金属。
4. 真正的伟人，不受金钱和地位的迷惑，不被贫穷和艰辛所动摇，不被权力和暴力所扼杀。

复制 播放

GPT3.5:



1. 我是一个在异乡漂泊的孤独陌生人，每逢佳节倍思亲。
2. 东流无尽水，去岁红颜老。
3. 志同道合者，金石为开。
4. 真正的伟人，不被金钱地位所迷惑，不被贫困艰辛所动摇，不被权势武力所压抑。

## 2. 情感分析

在情感分析任务中，我们会给出一段文本，要求语言模型分析出这段文本所表达的感情是积极的或消极的。消极语料如下：

(1) 如果你在外资公司工作过，尤其你在外资公司担任人力资源的话，就会觉得作者写的东西不是新野，拉拉口中讲的大道理，都是书中有的。作者对人物的描写也太肤浅，真的很失望。

(2) 小熊宝宝我觉得孩子不喜欢，能换别的吗

(3) 宝宝一岁七个月，不是太喜欢，也许是因为色彩不够鲜艳. 而且对他而言，可能太简单了.

(4) 很后悔, 不怎么样的一本书, 千万别买

(5) 虽然是老乡的作品，但是从客观上来讲的话，还是有点牵强，虽然作者的经历很丰富文采也不错，但不太喜欢结构和风格，这一次跟风跟错了。

(6) 收到的书有断页现象，从12页就跳45页了

(7) 此书我觉得我个人不是特别喜欢，所有的内容太过虚幻了。

积极语料如下：

(8) 环境很好，地点很方便，服务也很好，下回还会住的！

(9) 服务非常好, 希望不是正好赶上奥运才这样. 去哪里都有人带路. 宾至如归啊! 早餐还算丰富, 门口也有不少餐厅. 出门就是海边. 风景美. 值得一住!

(10) 入住大连富丽华东楼，新装修的，比西楼条件好，以后还会选择！

(11) 酒店不错，服务人员都很好，比较有礼貌。

(12) 干净，清静，交通也算方便。



判断下面每句话是积极的还是消极的

2

- (1) 如果你在外资公司工作过,尤其你在外资公司担任人力资源的话,就会觉得作者写的东西不是新野,拉拉口中讲的大道理,都是书中有的。作者对人物的描写也太肤浅,真的很失望。
- (2) 小熊宝宝我觉得孩子不喜欢,能换别的吗
- (3) 宝宝一岁七个月,不是太喜欢,也许是因为色彩不够鲜艳,而且对他而言,可能太简单了。
- (4) 很后悔,不怎么样的一本书,千万别买
- (5) 虽然是老乡的作品,但是从客观上来讲的话,还是有点牵强,虽然作者的经历很丰富文采也不错,但不太喜欢结构和风格,这一次跟风跟错了。
- (6) 收到的书有断页现象,从12页就跳45页了
- (7) 此书我觉得我个人不是特别喜欢,所有的内容太过虚幻了。
- (8) 环境很好,地点很方便,服务也很好,下回还会住的!
- (9) 服务非常好,希望不是正好赶上奥运才这样,去哪里都有人带路,宾至如归啊!早餐还算丰富,门口也有不少餐厅,出门就是海边,风景美,值得一住!
- (10) 入住大连富丽华东楼,新装修的,比西楼条件好,以后还会选择!
- (11) 酒店不错,服务人员都很好,比较有礼貌。
- (12) 干净,清静,交通也算方便。

文心一言:

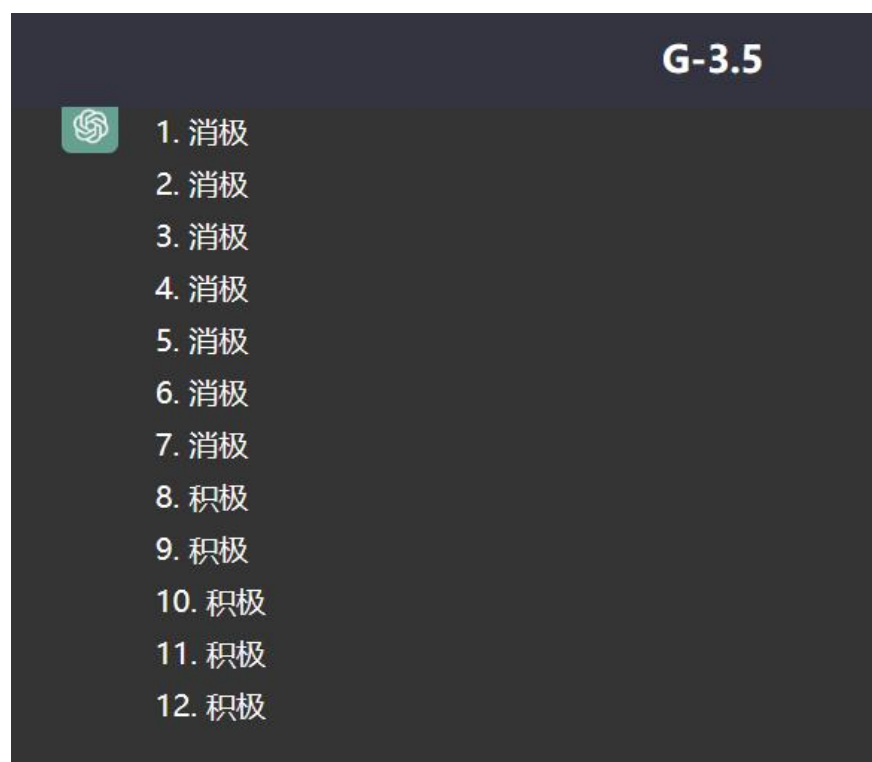


- (1) 消极。该评论者认为书中内容没有新意,对人物的描写肤浅,表达了消极的情感态度。
- (2) 中性。这句话并没有表达出明显的情感倾向,只是简单地询问是否能换别的东西。
- (3) 中性。这句话并没有表达出明显的情感倾向,只是简单地描述了孩子对这本书的反应。
- (4) 消极。该评论者对这本书持否定态度,认为这本书不好,不建议购买。
- (5) 消极。虽然这句话中提到了作者的经历和文采,但是整体上表达了对于作品的不太喜欢和对于跟风行为的后悔,情感态度为消极。
- (6) 中性。这句话描述了书的质量问题,没有表达出明显的情感倾向。
- (7) 消极。该评论者认为这本书的内容过于虚幻,表达了不太满意和失望的情感态度。
- (8) 中性。该评论并没有表达出明显的情感倾向,只是简单地描述了住宿环境和感受,因此可以判断为中性情感态度。
- (9) 积极。评论者对住宿体验表示满意,认为环境好、服务好、早餐丰富、地理位置方便,表达了积极的情感态度。
- (10) 积极。该评论者对大连富丽华东楼表示满意,认为装修比西楼好,以后还会选择,表达了积极的情感态度。
- (11) 中性。这句话没有表达出明显的情感倾向,只是简单地描述了酒店服务和人员的礼貌。
- (12) 中性。这句话没有表达出明显的情感倾向,只是简单地描述了酒店的干净、清静和交通便利的特点。

ChatGPT (ChatBot):



GPT3.5:

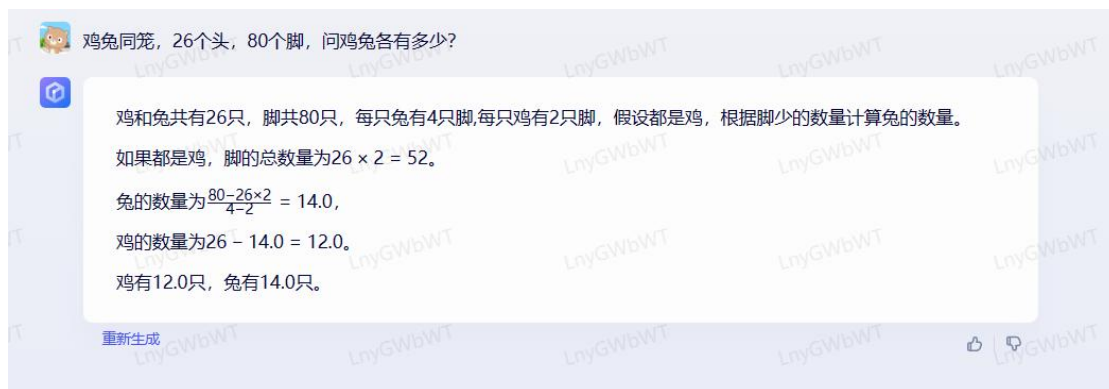


### 3. 问答系统

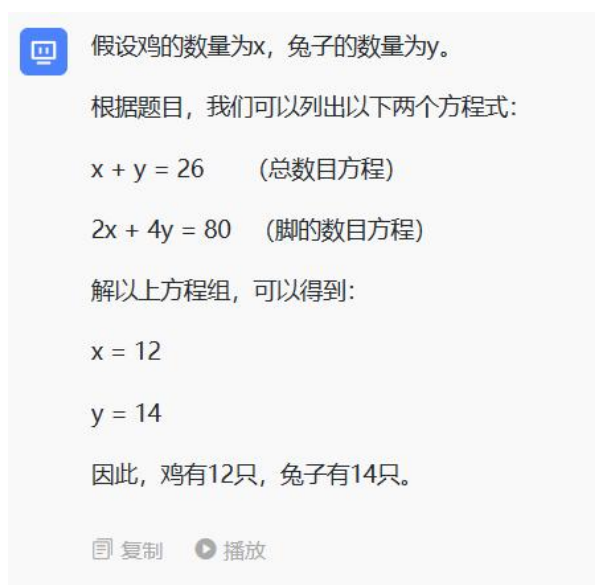
在问答任务中，我们将一些数学问题交给语言模型进行回答。

文心一言：

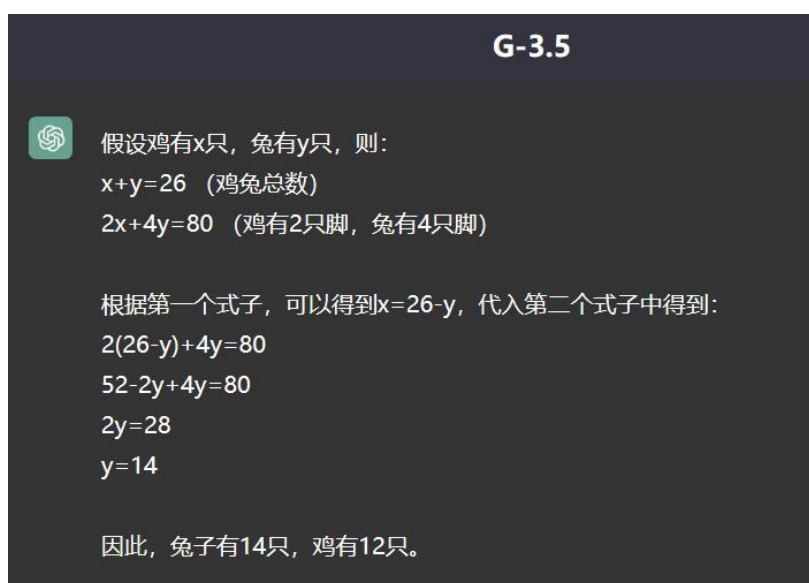




ChatGPT:



GPT3.5:



#### 4. 文本生成

在文本生成任务中，我们要求语言模型完成2023年高考新高考1卷的语文作文题目。





请你完成一篇作文：好的故事，可以帮我们更好地表达和沟通，可以触动心灵、启迪智慧；好的故事，可以改变一个人的命运，可以展现一个民族的形象……故事是有力量的。

以上材料引发了你怎样的联想和思考？请写一篇文章。

要求：选准角度，确定立意，明确文体，自拟标题；不要套作，不得抄袭；不得泄露个人信息；不少于800字。

文心一言：

故事的力量：心灵之火与智慧之源

作为一个古老的民族，我们中国人自古以来就有着丰富的故事，从神话传说到历史故事，从《山海经》到《三国演义》，这些故事不仅是我们的民族文化的重要组成部分，更是我们民族精神的重要载体。这些故事不仅让我们了解了历史和文化，更重要的是，它们拥有着强大的力量，能够启迪我们的智慧，触动我们的心灵。

好的故事可以更好地表达和沟通。无论是《红楼梦》中的爱情故事，还是《西游记》中的冒险故事，它们都以生动形象的方式展现出了人物内心的情感和思想。这些故事通过情节的展开和人物的塑造，让我们更好地理解人性的复杂和多样，让我们更好地了解了人与人之间的情感和思想。正是因为这些故事能够更好地表达和沟通，所以它们才能够成为经典的文化符号，让我们在今天依然能够感受到它们的魅力和价值。

好的故事可以触动心灵、启迪智慧。从孔子的《论语》到荀子的《劝学》，从庄子的《逍遥游》到孟子的《尽心篇》，这些经典故事不仅是儒家、道家、墨家等诸子百家的代表性作品，更是中华文化的重要组成部分。这些故事不仅让我们了解了中国古代的思想和智慧，更重要的是，它们通过启迪我们的智慧，让我们更好地理解人生和社会，让我们更好地面对生活中的挑战和困难。正是因为这些故事能够触动心灵、启迪智慧，所以它们才能够成为我们民族的精神支柱和文化瑰宝。

好的故事可以改变一个人的命运。每一个人都有自己的故事，而这些故事往往会影响到我们的人生轨迹。有时候，一个简单的小故事就能够激发出我们内心深处的力量，让我们变得更加勇敢、坚强、自信。有时候，一个感人肺腑的故事能够让我们重新审视自己的人生和价值观，让我们变得更加成熟和理性。正是因为这些故事能够改变一个人的命运，所以它们才

好的故事可以展现一个民族的形象。在现代社会中，一个民族的形象和声誉往往会影响到国家的国际地位和形象。有时候，一个简单的小故事就能够展现出一个民族的友好和热情，让外国友人更加了解和喜欢我们的文化。有时候，一个感人肺腑的故事能够让我们重新审视自己的人生和价值观，让我们变得更加成熟和理性。正是因为这些故事能够展现一个民族的形象，所以它们才能够成为国家文化软实力的重要体现。

在今天这个信息爆炸的时代，我们需要更多的好故事来启迪我们的智慧、触动我们的心灵、改变我们的人生轨迹、展现我们民族的形象。我们需要通过好故事来传递

正能量、弘扬真善美、促进文化交流与融合。因此，让我们一起来发掘和传播好故事吧！让好故事成为我们生活中的一抹亮色，让好故事成为我们心灵之火与智慧之源！

### ChatGPT:

标题：故事的力量

故事，是人类文化中最为基础、古老的形式之一。从古至今，无论是神话、传说，还是小说、电影，故事一直是人们表达思想、传递价值观的重要手段。好的故事，不仅可以帮我们更好地表达和沟通，还可以触动心灵、启迪智慧，改变一个人的命运，展现一个民族的形象。故事，是有力量的。

首先，故事是人们表达思想、传递价值观的重要手段。故事可以通过情节、人物、语言等多种方式，将作者的思想和价值观传递给读者。例如，《红楼梦》中的贾宝玉、林黛玉等人物形象，以及其中所反映的封建礼教、家族制度等社会问题，都是作者曹雪芹通过故事来表达自己的思想和价值观。同样，现代的电影、小说也同样具备这种表达思想、传递价值观的功能。比如《肖申克的救赎》中的主人公安迪，通过坚定不移的信念、勇气和智慧，最终逃脱了监狱的囚禁，得到了自由和尊严。这个故事不仅是一部优秀的电影，更是一种对人性、自由、正义等价值观的传递和表达。

其次，好的故事可以触动心灵、启迪智慧。故事中所展现的情节、人物形象往往能够深深地打动我们的内心，引发共鸣，让我们从中获得情感上的满足和启示。例如，安徒生童话中的《小美人鱼》，讲述了一个爱上人类王子的美人鱼，最终为了爱情牺牲自己的故事。这个故事虽然是一个童话，但其中所展现的爱情、牺牲等主题却深深地触动了读者的心灵。同样，中国古代的《孟子》、《庄子》等哲学经典中，也包含了许多富有启发性的故事，这些故事通过讲述人物的行为和思想，向读者展示了智慧和道德的重要性。

最后，好的故事可以改变一个人的命运，可以展现一个民族的形象。故事中所包含的人生哲理、价值观念，往往能够影响一个人的思想、行为和命运。例如，中国古代的《聊斋志异》中的许多故事，都包含了对道德、人性、命运等方面的深刻思考，这些故事不仅可以启迪读者的智慧，还可以影响他们的行为和命运。同样，好的故事也可以展现一个民族的形象。比如，美国的电影、文学作品中所展现的自由、民主、人权等价值观，已经成为了美国文化的重要组成部分，也成为了美国形象的重要标志。

总之，故事是有力量的。好的故事可以帮我们更好地表达和沟通，可以触动心灵、启迪智慧；好的故事可以改变一个人的命运，可以展现一个民族的形象。因此，我们应该珍视故事，通过阅读、传递、创作优秀的故事，来丰富自己的生活、提高自己的素质，同时也为社会的进步、文明的发展做出自己的贡献。

### GPT3.5:

好的故事：启迪智慧，改变命运，展示形象

故事是人类传承和交流文化的重要载体，是人类智慧和情感的结晶。好的故事不仅能够帮助我们更好地表达和沟通，更能够触动我们的心灵，启迪我们的智慧。在这篇文章中，我将从三个方面来探讨好的故事的力量：启迪智慧，改变命运，展示形象。

启迪智慧，是好的故事的重要功能之一。在古代，故事是人们学习智慧和道德的重要途径，如孔子的《论语》、荀子的《劝学》、庄子的《逍遥游》等都是经典的故事，它们不仅传承了古代智慧，更通过情节和人物的塑造，启迪了人们的智慧。在现代社会，好的故事也同样能够启迪我们的智慧，比如莫言的《丰乳肥臀》、余华的《活着》等，它们通过生动的故事情节，让我们更好地了解人性的复杂和多样，更好地理解世界和人生的意义。

改变命运，是好的故事另一个重要功能。故事中的人物往往代表着我们自己或者周围的人，通过他们的成长和经历，我们能够更好地认识自己，更好地追求自己的梦想。比如《小王子》中的小王子，通过与其他星球的人物的交流，最终找到了自己的归宿；比如《三体》中的罗辑，通过自己的智慧和勇气，拯救了人类的未来。这些故事告诉我们，命运不是注定的，而是由我们自己创造的，只要有勇气和智慧，就能够改变自己的命运。

展示形象，是好的故事的另一个重要功能。故事是一个民族文化的重要组成部分，也是展示一个民族形象和精神的重要方式。好的故事能够让外国友人更好地了解和喜欢我们的文化，从而促进文化交流和融合。比如《红楼梦》中的爱情故事，不仅让我们了解了清代的文化和风俗，更让我们感受到了中国古代文人的情感和思想；比如《西

游记》中的冒险故事，不仅展现了中国古代的神话和传说，更让我们看到了中国人勇敢、智慧和善良的一面。

好的故事是有力量的，它们能够帮助我们更好地表达和沟通，触动我们的心灵，启迪我们的智慧，改变我们的命运，展示我们民族的形象。因此，让我们珍视好的故事，让它们成为我们生活中的重要组成部分，共同创造美好的未来！

## 二、结果分析

### 1. 机器翻译

在将英文句子翻译为中文文本的任务中，文心一言的总体表现要更优，这可能是因为文心一言使用更多的中文语料库进行训练的原因。但chatGPT和GPT3.5也有一些令人眼前一亮的翻译，如“东流无尽水潺潺，昔日英雄尽随波”很有诗意且押韵、“志同道合者，金石为开”甚至比文心一言的翻译更加准确。这些都表明了chatGPT、GPT3.5在处理中文语言任务上已经有了相当的能力。

### 2. 情感分析

在我们所给出的12段语料中，ChatGPT和GPT3.5全部判断正确，而文心一言审题不清，将一些句子判为中性，并且在提示一次后未加改正，可见文心一言的理解能力有待提高。然而文心一言给出了每句话的解释，这是很好的。

### 3. 问答系统

在我们给出的鸡兔同笼小学数学题中，三种语言模型全部回答正确并给出了解题过程。然而，在面对初中或高中的数理问题时，这些语言模型的表现常常惨不忍睹。

### 4. 文本生成

面对我们给出的全国1卷语文作文题“故事的力量”，三个语言模型都完成了自己的创作。这些文章都引经据典，比较符合题意，没有出现偏题跑题现象。然而我们发现文心一言的引用名篇名句较少，且在最后两段论述中出现大量白话套话，而chatGPT和GPT3.5引用的中外名篇反而更多一些。然而从写文采上看，三篇文章都有待提高，没有很多优美的语言和辞藻，略显白话。

## Conclusions

本次作业在自然语言处理的四个下游任务上对三种语言模型进行了测试，并在机器翻译任务中使用了提示工程的方法。总的来看，三种语言模型都能够完成一些简单的任务，然而在比较依赖创作能力的机器翻译和文本生成任务上，三种模型都显得有所不足，可见这些语言大模型能够帮助人类完成一些简单的任务，而人类的创造能力是它们不容易取代的。相比而言，chatGPT和GPT3.5的总体性能要优于文心一言，文心一言在处理中文信息方面没有展现出很强的统治力。