分类号_____ 编 号_____

U D C_____ 密 级_____

南方科技大学
SOUTHERN UNIVERSITY OF SCIENCE AND TECHNOLOGY

本科生毕业设计（论文）

题 目： 基于大语言模型的自动数学建模研究

姓 名： 夏祎杨

学 号： 12012921

系 别： 计算机科学与工程系

专 业： 计算机科学与技术

指导教师： 史玉回 讲席教授

2024 年 6 月 7 日

诚信承诺书

1.本人郑重承诺所呈交的毕业设计（论文），是在导师的指导下，独立进行研究工作所取得的成果，所有数据、图片资料均真实可靠。

2.除文中已经注明引用的内容外，本论文不包含任何其他人或集体已经发表或撰写过的作品或成果。对本论文的研究作出重要贡献的个人和集体，均已在文中以明确的方式标明。

3.本人承诺在毕业论文（设计）选题和研究内容过程中没有抄袭他人研究成果和伪造相关数据等行为。

4.在毕业论文（设计）中对侵犯任何方面知识产权的行为，由本人承担相应的法律责任。

作者签名：夏祎杨

2024 年 6 月 7 日

基于大语言模型的自动数学建模研究

夏祎杨

（计算机科学与工程系　指导教师：史玉回）

[摘要]： 本论文主要研究了基于大语言模型的自动数学建模，特别是在使用 GPT-4 进行数学建模时的应用。本文首先介绍了自动建模的基本概念和方法，然后详细阐述了如何利用大语言模型进行自动建模。本文提出了一种新的建模方法，该方法能够自动地根据问题描述生成相应的数学模型。该方法不仅能够处理传统的线性规划数学建模问题，还能够处理更复杂的实际问题。本文在 LLaMA 和 GPT3.5/4 上进行了大量的实验。在这个过程中，本研究发现提示词的选择对模型输出的质量有着重要的影响。因此，本文设计了一系列的实验，旨在找到最优的提示词，以提高模型响应的性能。实验结果表明，通过合理的提示词选择，可以显著提高模型的性能。本文希望相关的研究成果能够为其他研究者提供参考，帮助他们更好地利用大语言模型进行自动建模。

[关键词]：大语言模型;数学建模;自动建模;GPT4

[ABSTRACT]: This thesis primarily investigates automatic modeling based on large language models, particularly the application of GPT-4 in mathematical modeling. We first introduce the basic concepts and methods of automatic modeling, then elaborate on how to utilize large language models for automatic modeling. We propose a new modeling method that can automatically generate corresponding mathematical models based on problem descriptions. Our method can handle not only traditional mathematical modeling problems but also more complex practical problems. We conducted extensive experiments on LLaMA, GPT-3.5 and GPT-4. In this process, we found that the choice of prompt words has a significant impact on the quality of model output. Therefore, we designed a series of experiments aimed at finding the optimal prompt words to improve model performance. Our experimental results show that through reasonable selection of prompt words, we can significantly improve the performance of the model. We hope that our research findings can provide a reference for other researchers, helping them to better utilize large language models for automatic modeling.

[Keywords]: Large language model; Mathematical modeling; Automatic modeling; GPT4

目录

1. 引言

1.1 研究背景

随着计算机科学和人工智能的飞速发展，大语言模型（Large Language Models，LLMs）已经成为了一个重要的研究领域。这些模型，如 LLaMA 和 GPT-4，已经在各种任务中表现出了显著的性能，包括但不限于自然语言处理、机器翻译、文本生成等。然而，尽管这些模型在处理语言任务方面表现出色，但它们在处理数学问题，特别是自动数学建模方面的能力，仍然有待提高。

自动数学建模是一个复杂的过程，需要理解问题的上下文，选择合适的数学工具，构建模型，然后解决模型。这个过程需要深厚的数学知识和强大的计算能力。然而，大语言模型的潜力使绝大多数人相信，它们可以被训练来执行这样的任务，从而极大地推动自动数学建模的发展。

1.2 研究意义

本研究旨在探索和理解大语言模型在自动数学建模中的应用。本文将研究如何训练这些模型来理解和解决数学建模问题，以及如何评估它们的性能。

通过这项研究，本文希望能够推动自动数学建模的发展，使其更加智能和高效。这将对许多领域产生深远影响，包括工程、物理、经济、生物等，因为这些领域都依赖于数学建模来理解和解决复杂问题。

此外，本文的研究希望能够为自然语言处理、机器学习、人工智能等领域的研究者提供新的工具和方法，从而推动这些领域的发展。

总的来说，本研究将有助于推动大语言模型的发展，提高其在自动数学建模中的性能，从而为各种领域带来实质性的改进。

1.3 研究挑战

在尝试将大语言模型应用于自动数学建模的过程中，本文发现了现在面临的一系列重要的技术挑战和现有工作的不足。

首先，大语言模型需要能够理解和表示复杂的数学概念，包括函数、方程、矩阵、向量等。这需要模型具有强大的抽象思维能力和深厚的数学知识。然而，现有的大语言模型在这方面的能力还有待提高。此外，现有的大语言模型主要依赖于统计学习和模式匹配，而缺乏对数学概念和原理的深度理解。这使得它在处理复杂的数学建模问题时，可能无法建立正确的模型。

其次，自动数学建模不仅需要理解数学概念，还需要在此基础之上能够构建出来数学模型。这是一个相当复杂的过程，需要模型具有强大的逻辑推理能力。然而现在的大语言模型对此的处理并不够优秀。

再者，建模后通常需要处理大规模的数据和计算。这需要模型具有高效的数据处理能力和计算能力。然而，现有的大语言模型在这方面的能力还有待提高。同时，训练和运行大语言模型也需要大量的计算资源。这使得这些模型的应用受到了限制，特别是在资源有限的环境中。当然，因为本文只局限于对自动数学建模的研究，所以并未强求运算结果的准确性，只对其进行初步分析。

此外，大语言模型的性能在很大程度上取决于训练数据的质量和数量。然而，高质量的数学建模训练数据是非常稀缺的，这限制了模型的性能和泛化能力，尤其是对于数学模型的专精。

最后，评估大语言模型在自动数学建模中的性能是一个重要的挑战。这需要开发新的评估指标和方法，以准确地衡量模型的性能。然而，如何找到直观可见的评估方式也值得在本文中深思和探索。

总的来说，本研究面临着一系列重要的技术挑战。仍需深入研究和理解这些挑战，以便开发出更强大、更有效的大语言模型，从而推动自动数学建模的发展。

同时本文发现解决这些问题和不足对于推动大语言模型在自动数学建模方面的也应用具有十分重要的意义。

1.4 解决方式

在解决大语言模型自动数学建模的问题时，本文采取了一种多层次的策略。首先，本文对现有的数学建模相关的大语言模型进行了深入的研究，以了解它们在自动建模领域的应用效果。本文分析了这些模型的优点和缺点，并尝试找出可以改进的地方。再然后，本文参考了 26 条提示词设计原则，从提示词层面尝试如何使用更有效的提示词得到更加精确的建模结果。本研究设计了一系列的实验，通过改变提示词的形式和内容，观察模型的反应，以找出最有效的提示词策略。本研究在 Llama3、GPT3.5/4、通义千问等多个语言模型上进行多组测试实验，最终得到想要的优秀提示词策略。

1.5 主要贡献

本研究的主要贡献在于：本文对大语言模型在自动数学建模中的应用进行了深入的研究和实践。本文不仅分析了现有模型的性能和限制，对于大量不同类型的建模问题进行实验分析，发现了现有自动建模工具的不足之处。还尝试优化提示词设计，找到最优秀的提示词工程策略，得出更有效的建模方式。以供其他研究者研究。除此之外，本文的相关研究不仅推动了自动数学建模的发展，也为其他领域的大语言模型的研究和应用提供了新的视角。

2 相关工作

2.1 大语言模型

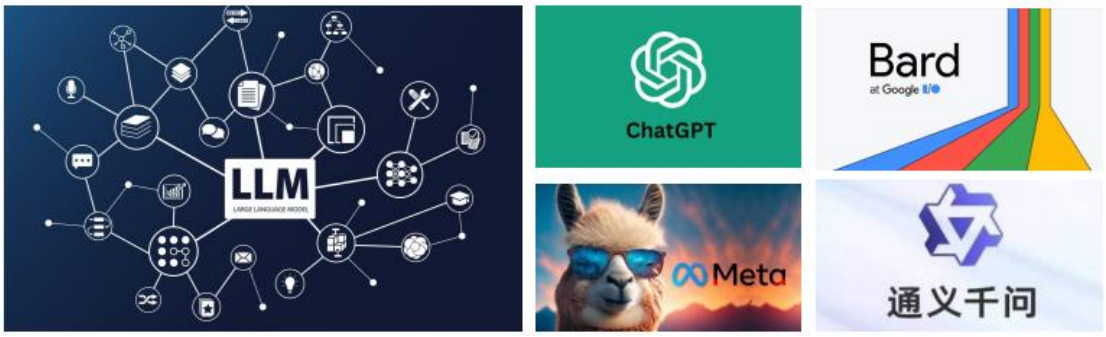
大型语言模型的发展对于推动自然语言处理的进步起着关键作用。本节将回顾大型语言模型的主要发展历程，为本文的研究奠定基础。从 Google 的 BERT 开始，它通过双向训练方法彻底改变了人们对上下文理解的方式。而 T5 则通过将各种自然语言处理相关任务统一到一个框架中，进一步推动了该领域的发展。同时，GPT-1 引入了一个创新性的模型，利用 Transformer 架构进行无监督学习。其继任者 GPT-2 将参数数量显著扩展到 15 亿个，展示出了在文本生成方面的强大能力。再接着，GPT-3 已经拥有近 1750 亿个参数，并展示出了对各种语言任务的高熟练度，标志着大型语言模型规模和能力的重大飞跃。

在其他最近提出的大型语言模型中，Gopher 不仅通过其 2800 亿参数模型提供了先进的语言处理能力，而且还将伦理考虑放在了首位。Meta 的 LLaMA 系列强调了效率的重要性，表明可以用更少的资源获得更强大的性能，同时 Chinchilla 也倡导这一概念，它提出经过优化训练的更小的模型也可以获得卓越的结果。这一系列创新中的最新成果是 Mistral，它在效率和性能方面表现出色，优于其它大型模型。在大型语言模型发展这一轨迹的最新里程碑是 OpenAI 的 GPT-4 和谷歌的 Gemini 系列。它们代表了该领域的另一项重大进步，指增强了文字理解和生成能力，这为大型语言模型在各个领域的应用设定了新的基准。

GPT-4，作为 OpenAI 的最新大型语言模型，继续了这一发展轨迹，将参数数量扩展到了惊人的数目，并在理解和生成能力上取得了新的突破。GPT-4 不仅在文本生成方面展示了卓越的能力，而且在理解复杂的语义信息、处理非结构化的文本数据以及生成高质量的文本输出方面也有显著的提升。此外，GPT-4 还在伦理考虑方面做出了重要的贡献，将人性化的理解和生成能力与对公平性、透明性和解释性的关注相结合。这使得 GPT-4 不仅在技术上达到了新的高度，并且在伦理和社会影响方面也提高到了的基准。

现代还有很多的常见大语言模型，如图一所示，在此就不逐一介绍。由此

可知现在研究对于大语言模型已经有较为深入的探索。

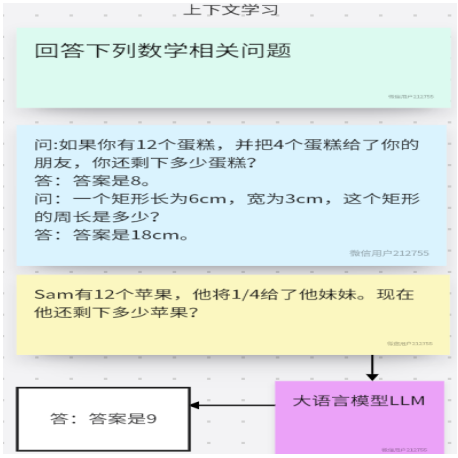


图一：常见的大语言模型

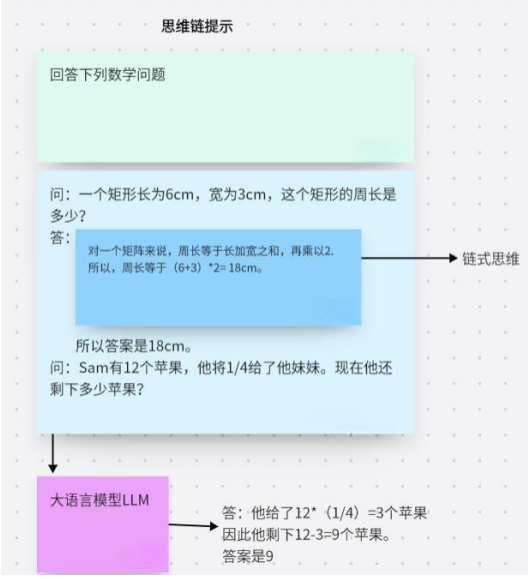
关于大语言模型有三种主要应用方法：上下文学习、思维链提示以及规划。

1. 上下文学习：这种方法以自然语言文本的形式制定任务描述和演示。首先，研究用户给出任务描述，并在任务数据集中选择几个例子作为示范。然后，形成具有特殊设计模板的自然语言提示，最终将测试用例按输入的方式添加到示范中。其中实际答案被留作空白，由大语言模型预测，从而得到结果。下图二中给出了上下文学习的留白公式和流程图。
2. 思维链提示[1]：这是一种改进的提示策略，旨在提高大型语言模型在复杂推理任务中的表现。与上下文学习中简单地使用输入-输出对构造提示不同，思维链提示在提示中加入了中间推理步骤，引导出最终输出。例如，实验中可以将每个演示增加为一个包含输入、中间推理步骤和输出的序列。下图三中给出了思维链提示的流程图。
3. 规划[2]：这是一种将复杂任务分解成更小的子任务，并生成完成任务的行动计划的方法。对于解决一个复杂的任务，任务规划者首先需要清楚地了解任务目标，并根据大语言模型的推理生成合理的计划。然后，计划执行者在环境中根据计划行动，环境将为任务规划者产生反馈。任务规划器可以进一步结合从环境中获得的反馈来完善其初始计划。下图四中给出了规划的流程图。

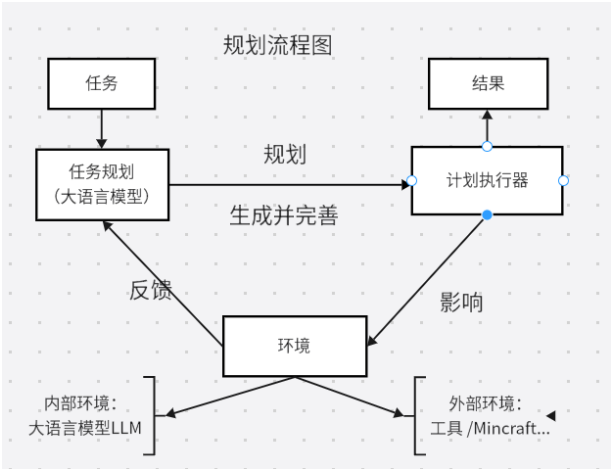
LLM(I，(x1,y1),..f(xk ,yk),f(x(k+1),__))→yk+1



图二：上下文学习，左为答案留白公式，右为流程图



图三：思维链提示流程图



图四：规划流程图

2.2 自动建模

自动建模是一种使用计算机算法自动构建数学模型的过程。这种方法可以处理大量的数据和复杂的问题，而无需人工进行繁琐的建模工作。自动建模的基本步骤包括：定义问题，选择合适的建模方法，生成模型，验证模型的有效性，以及优化模型的性能。在自动建模的过程中，大型语言模型首先会解析问题描述，理解其含义和要求。然后，模型会根据问题描述生成一个初步的数学模型。这个模型可能是一个公式、一个算法，或者一个更复杂的数学结构，取决于问题的具体要求。而这就是本研究想要大语言模型准确做到的。

2.3 提示词

提示词输入是与大型语言模型交互的一个独特方面，它的性能简单，无需微调模型，已经发展成为一个精细的研究领域。这个领域凸显了用户输入和大型语言模型响应之间的复杂关系，并已经深入探索了不同的提示设计如何显著影响大语言模型的性能和输出，这标志着提示工程的诞生。并随着这一领域的迅速扩展，揭示了提示词在少样本（Few-shot）和零样本（Zero-shot）学习场景中的关键作用。例如，论文[3]详细描述了提示词与 GPT-3 的合作，其中制作的提示词使模型能够以最少的先验示例执行任务。最近研究[4]已经转向理解提示词中的语义和上下文细微差别，研究这些细微的变化如何导致大型语言模型产生截然不同的反应。

而在 ChatGPT[5]的研究中，提示工程相关技术得以深入探讨，其中强调了提示工程在增强软件开发和教育中的大语言模型应用的重要性。该研究进一步强调，有效的提示词设计对于提高大型语言模型性能至关重要，特别是在编码实践和学习经验方面。

因此，本研究自然而然地思考到，是否可以通过设计优秀的提示词来提升自动建模的效果。为此，本文参考了论文文献[6]中的 26 项提示词准则进行学习。这些准则为本研究提供了一个框架，帮助理解如何使人们更有效地与大型语言模

型进行交互，以提高其性能和应用效果。

这 26 项准则包括但不限于：明确的指示，避免歧义，使用适当的语言风格和语境，考虑到模型的局限性，以及在可能的情况下，提供具体的例子或模板。这些原则的应用，可以帮助用户更好地设计和优化提示词，从而提高大语言模型在各种任务中的性能。后文中将会列出本文对这些准则的大致分类。

3. 研究方法

3.1 对现有建模工具的分析

在本研究中，选择了 SeedModeler、GPT-4 和 MindOpt 作为实验的对照参考。SeedModeler 和 MindOpt 都是当前市场上领先的团队研发的通用代数建模语言。这两种建模语言的使用步骤大致可以分为以下四步：

1. 自然语言输入需求：用户输入他们希望解决的数学建模问题。
2. 自动数学建模：模型会自动分析用户的需求，并用数学语言表达这个问题。
3. 生成可执行代码：根据数学模型生成对应的代码，并从问题中提取数据或生成测试数据作为模型的数据输入。
4. 计算得出结果：最后，模型会计算并得出结果。

本研究聚焦于第 1 和第 2 步——建模环节。为了验证这些工具在建模方面的准确性，本研究选择了一个简单的线性规划生产问题进行实验（如图五所示）。这个线性规划问题让此研究可以在一个控制的环境中，观察和评估这些工具的性能。

题目

一奶制品加工厂用牛奶生产 A_1, A_2 两种奶制品，1桶牛奶可在甲车间用 12h 加工成 3kg 的 A_1 ，或者在乙车间用 8h 加工成 4kg 的 A_2 ，根据市场需求，生产出的 A_1, A_2 全部都能售出，每千克 A_1 获利 24 元，每千克 A_2 获利 16 元，现在加工厂每天能得到 50 桶牛奶的供应，每天正式工人总的劳动时间为 480h ，且甲车间设备每天至多能加工 100kg 的 A_1 ，乙车间的设备加工能力可以认为没有上限限制(加工能力足够大)，试为该厂指定一个生产计划，使得每天的获利最大。

图五：测试用简单的线性规划生产问题

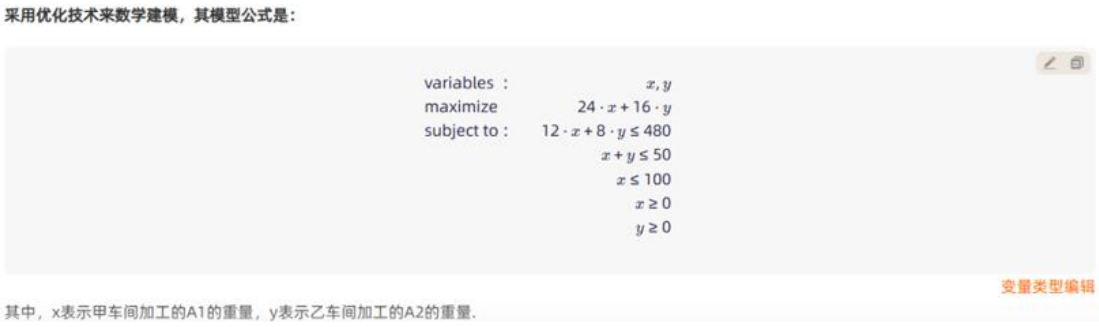
以 SeedModeler 为例，在这项实验中，采用的是上下文学习的方法。因为该语言模型的数据库本来存在相关例题，实验用户只要在提示词输入框中输入问题原题，SeedModeler 分析用户的需求并自动生成了图六左图的数学模型，右图则为它生成的代码以及计算结果。根据实验结果可以发现 SeedModeler 和 MindOpt 能够正确地建立模型，但却无法正确地计算出结果。GPT-4 却表现出了不同的情况，通过与 GPT-4 的文本交互，实验中可以很好地得出正确的答案，提供答案的

方法为提供可计算答案的 python 代码。不过本文中运算结果并不在本文的具体探讨范围内。其中图六所示为 SeedModeler 结果，图七为 MindOpt 生成的数学模型，图八为 GPT4 生成的数学模型。

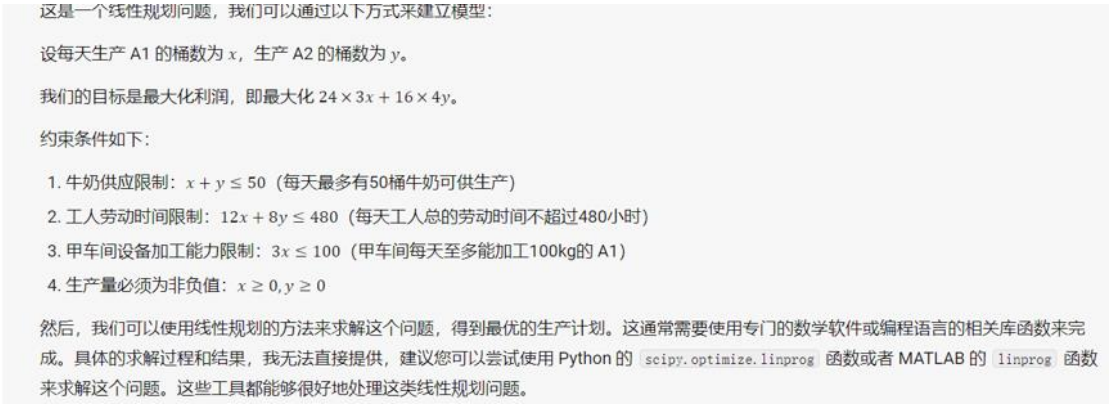
这项实验研究探究了目前现有自动数学建模工具，并将其与 GPT-4 进行比较。实验结果证明了 GPT-4 在目前关于自动数学建模中有着优秀的表现，也因此后文中提示词优化选取 GPT-4 作为自动建模研究对象。

，

图六：SeedModeler 自动建模所得的模型和运算结果



图七：MindOpt 生成的自动数学建模



图八：GPT4 生成的数学模型

3.2 基于大型语言模型的问题自动建模提示词优化

3.2.1 动机

由于预训练好的大语言模型生成的响应质量直接取决于用户提供的提示词或指令的质量，因此，制定出大语言模型能够理解并有效响应的提示词至关重要。传递给大语言模型的提示是一种使用户与大语言模型之间的交互编程方式，这也同时增强了大语言模型处理各种任务的能力。本研究的主要焦点在于如何制定提示以提高输出质量的方法。这需要本研究全面理解大语言模型的功能和行为，还有其底层机制以及控制其响应的原则。

在本研究中，通过参考论文《Principled Instructions Are All You Need for Questioning LLaMA-1/2, GPT-3.5/4》中的 26 条原则 [6]。本文设计了五大类提示词原则，以便在不同的场景和情况下进行全面的提示。示例如图九所示。

3.2.2 提示词优化

对于 26 项提示词优化原则，根据其特性，本文将其分为五类，如图九所示：（1）提示词的结构和清晰：例如，将目标受众整合到提示中，使得模型明白其受众是该领域的专家。（2）特异性信息：例如，在提示中添加特定的短语，如“确保您的答案是公正的，不依赖于刻板印象”。（3）用户互动和参与：例如，允许模型通过向用户提问来获取精确的细节和要求，直到它有足够的信息来提供所需的输出，如“从现在开始，我希望你问我问题…”。（4）内容和语言风格：例如，对 LLM 无需客气，因此无需添加“请”、“如果你不介意”、“谢谢”、“我愿意”等短语，而应直截了当地提出问题。（5）复杂任务和编码提示：例如，在交互式对话中，将复杂任务分解为一系列更简单的提示提问。

类别	准则	对应编号
提示词的结构和清晰	在prompt中整合目标的受众，例如：受众是该领域的专家 使用肯定的指示do，同时避开否定的语言don’t 使用引导词，比如‘一步一步地进行思考’ 使用输出引语，像用所需输出的开头来结束提示词 使用分隔符 用“Instruction”开头，然后接“Example”，并用#分隔，使用多个换行符来分割指令，示例，问题，上下文和输入。	2 4 12 20 17 8
特异性信息	使用few-shot prompting 让它用简单语言解释：把我当11岁小孩来解释/把我当初学者来解释 在提示词中加入以下这句话：“确保你的答案是公正的没有刻板印象的” 若要编写与所提供示例相似的文本，请包含具体说明 提供结构：我为你提供开头，根据所提供的单词完成它，保持流程一致。 以关键词，规则，提示的形式。清楚的说明模型为了生成内容必须最训导要求 询问特定话题或想法并测试你的理解：教我 [任何定理/主题/规则名称]，并在最后进行测试，在我回答以后让我知道我的回答是否正确，而不是事先提供答案。 通过添加所有必要信息，为我详细描写一篇关于 [主题] 的详细 [文章/文本/段落]。	7 5 13 26 24 25 15 21
使用者互动与参与	允许模型通过想你提出问题来从你那里引出精确的细节和需求，直到他有足够的信息来提供所需的输出 为详尽的得到文本内容：“通过添加所有必要的信息，为我详细写一篇关于[主题]的详细[文章/文本/段落]”	14 21
内容与语言风格	在不更改文本风格的情况下纠正/更改特定文本 结合一下文本：“你的任务是”，“你必须” 结合一下文本：“你会受到惩罚如果” 为语言模型分配一个角色 使用短语“用自然语言回答一个问题” 不要对大语言模型使用礼貌用语 在提示符中多次重复一个特定的单词或语句 提示词中加上“我会付多少美元来得到个更好的答案”	22 9 10 16 11 1 18 6
复杂任务和编码提示	在交互式对话中将复杂的任务分解成一系列更简单的提示词 在不同文件中的复杂代码：可以命令模型生成跨多文件的代码自动生成可运行脚本。 结合链式法则使用few-shot提示词	3 23 19

图九：五大类提示词原则

然而，在增加实验次数的过程中，本研究发现对于同一个问题，大型语言模型有时会得出错误的结论。为了提高 GPT-4 的改进率和准确性，本文采用提示设计原则。在本项研究中确立了一套指导原则，旨在通过精心设计的提示和指令，从预训练好的大型语言模型中引导出高质量的响应,具体指导如下：

(1) 简洁明了：提示应避免冗长和模糊不清，以免引起模型的混淆或产生无关的响应。提示词应简洁、明确，排除对任务无益的信息，同时具备足够具体的信息以引导模型。这是提示工程的基本原则。

(2) 上下文相关性：提示词需要提供相关的上下文信息，帮助模型理解任务的背景和领域。其中包含关键词、特定领域的术语或情境描述等。这一设计理念在本文的原则中得到了强调。

(3) 任务对齐：提示词应与当前的任务紧密对齐，使用明确指示任务性质的语言和结构。这可能涉及将提示表述为问题、命令或填空语句，以满足任务的预期输入和输出格式。

(4) 示例演示：对于更复杂的任务，提示中需要包含示例可以展示所需的响应格式或类型。这通常涉及显示输入-输出对，尤其是在“少样本”（Few-shot）或“零样本”（Zero-shot）学习场景中。

(5) 避免偏差：提示词的设计应尽量减少模型中由于其训练数据而带有的固有的偏差。使用中性语言，并注意潜在的道德影响，尤其是对于敏感话题。

(6) 增量提示：对于需要进行一系列步骤才能完成的任务，可以将任务分解为一系列相互依赖的提示，逐步指导模型。此外，提示应根据模型的性能和迭代反馈进行调整，也就是说，它需要做好充分的准备，根据初始输出和模型行为来优化提示。此外，提示词还应根据迭代的人类反馈和偏好进行调整。

4 实验与结果分析

4.1 实验设置

本文采用 LLaMA、GPT4 和通义千问对四类数学问题（如表 1 所示）的自动建模进行了实验研究，这四类数学问题是一个手动选择的评判基准，用于原则性的评估。这其中包含了跨各个领域的问题。对于每一大类原则，本文挑出了 10 项人工选择的问题，进行有无原则提示的实验对照。本研究比较每一对来自使用原则和没有使用原则对于相同问题指令的响应，并通过人工评估来评估大语言模型输出的各种指标。

表 1　四类数学问题表

四类数学问题	部分具体子模型
1 优化模型	1.1 数学规划模型
	1.2 微分方程组模型
	1.3 概率模型
2 分类模型	2.1 判别分析模型
	2.2 聚类分析模型
3 评价模型	3.1 层次分析法模型
	3.2 灰度关联度分析模型
	3.3 组合评价模型
4 预测模型	4.1 回归分析法模型
	4.2 时间序列分析模型
	4.3 BP 神经网络模型
	4.4 组合预测模型

数据来源：电子文献四大模型总结[7]

本文采用了现有的 LLaMA、GPT4（ChatGPT）和通义千问语言模型作为本研究的基本模型。本文设置了两个关键指标以供评估这些模型：改进率和准确率。本文将它们一起进行评估，以便对模型性能得到全面了解。

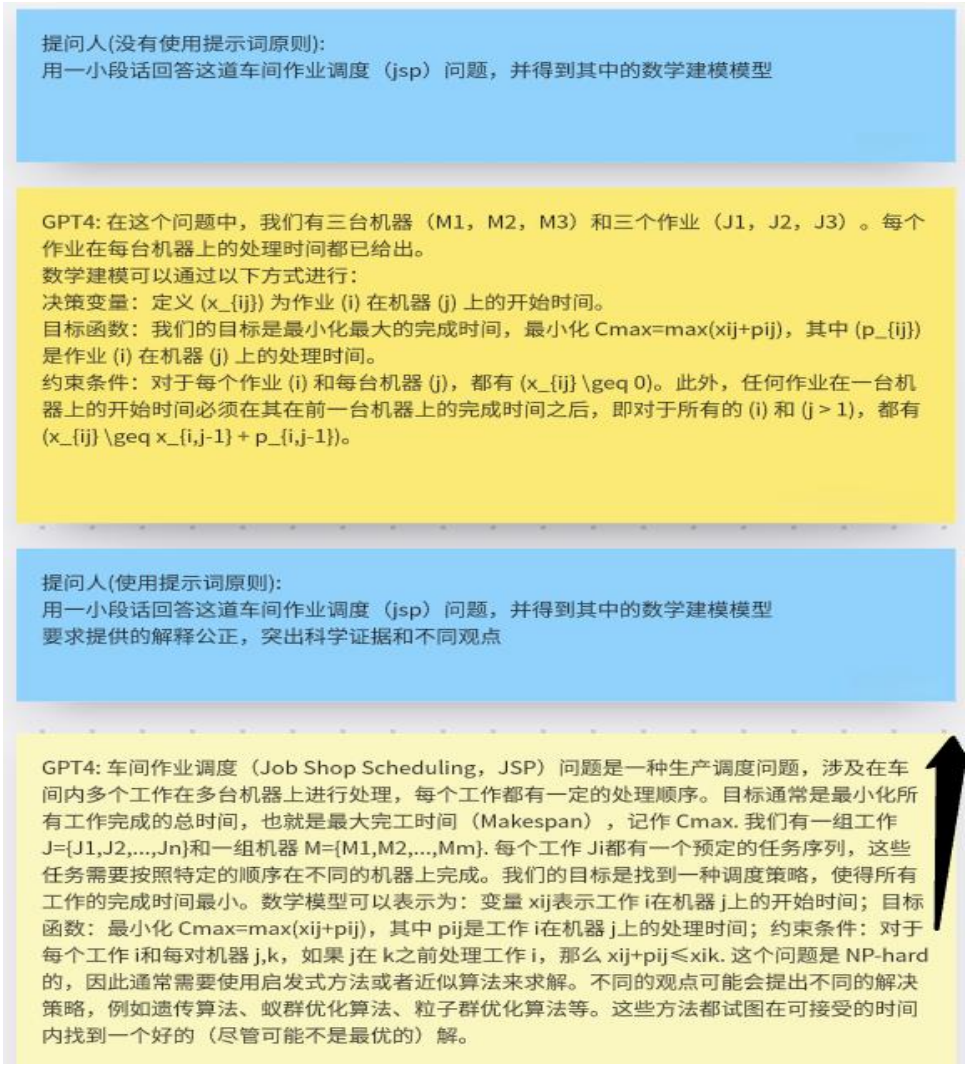
由于本研究使用的问题通常涉及推理任务，从而得到正确的答案，因此参考的 26 项准则中某些原则并不适用。例如原则 14、15、21、22、23 在以下问题：“假设 a 和 b 是正实数，a > b 和 ab = 8。求出 a^2+b^2 除以 a-b 的最小值，并列出它的数学模型”中并不适用，因此本文在实验过程中只会采用五大类提示词原则中的部分编号对应的原则。

对于本研究的性能评估，本文选取了改进率和准确率两个评估指标，定义如

下：

• 改进率：改进率是指在应用了的提示词原则时，一组问题的输出响应质量提高的百分比。提供的准则被使用以后，本研究通过人工评估来衡量不同大语言模型的响应质量的改进率。原始的提示词充当衡量此增强幅度的基线。通过响应质量的提升得以证实，由于使用了结构化的、有原则的指令，模型展现出来的性能得到了提高，如图十所示。

• 准确率：准确率是指模型输出响应的精确度，以供确保它们的回答足够准确且与原题相关联。本文既考虑了绝对准确率，也考虑了相对准确率。绝对准确率是一个直接的度量，用于衡量模型预测的结果与实际答案之间完全匹配的频率。相对准确率则是一个更为复杂的度量，它考虑了模型预测的结果与实际答案之间的接近程度。此研究过程同样采用人工评估的方式来衡量这一指标。

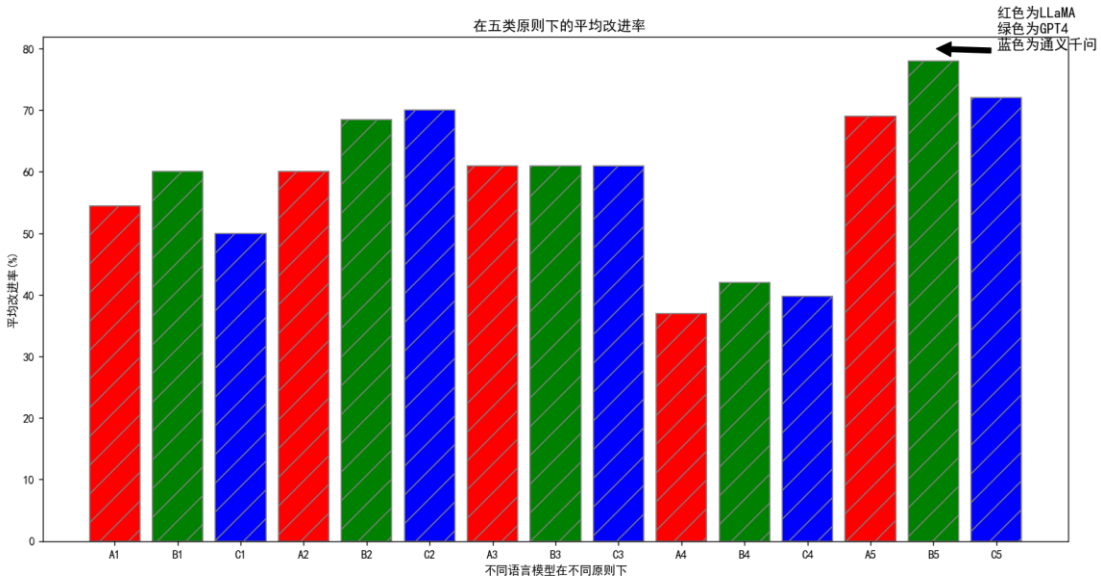


图十：在提示词上使用第二类原则后大语言模型响应的改进示例。

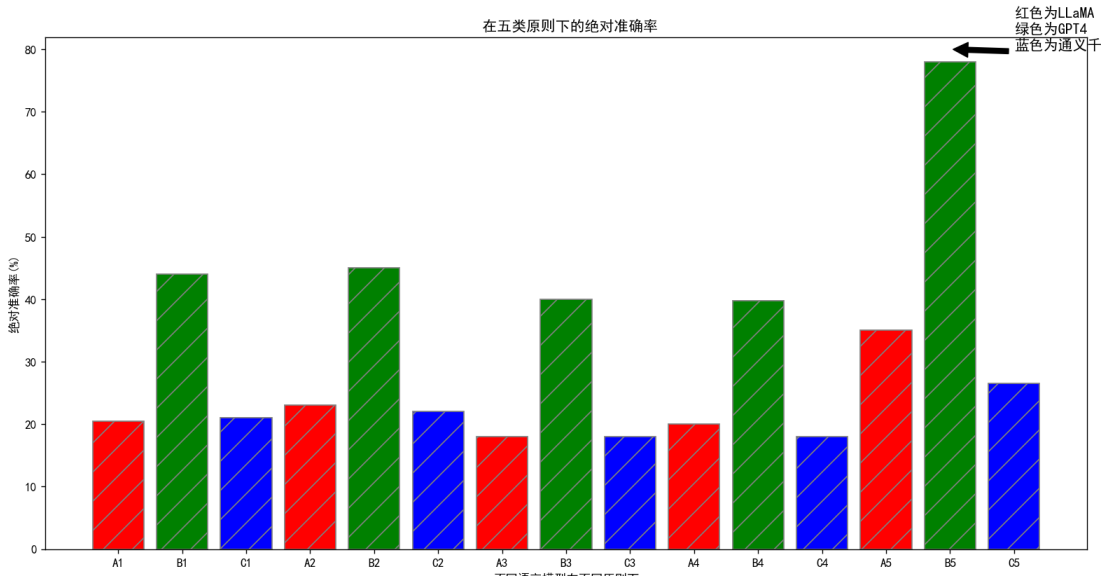
4．2 实验结果

4.2.1．在 LLaMA、GPT4（ChatGPT）和通义千问上的实验结果

改进率：引入的这五类原则带来的改进结果如图十一所示。其中红色柱代表对于五类原则问题使用 LLaMA 的平均改进率，绿色柱代表对于五类原则问题使用 GPT4 的平均改进率，蓝色柱代表对于五类原则问题使用通义千问的平均改进率。横坐标按每三个不同颜色柱体对应一类提示词原则（总共分为五组）的相关实验。纵坐标代表改进率。总的来说，所有的原则都可以为 LLaMA、GPT4（ChatGPT）和通义千问上的改进率上带来显著的提升。从下图中可知，对于 LLaMA 和通义千问而言，使用提示词原则的改进率在 35%~65%，而对于 GPT4，改进率的范围在 40%~80%。在使用第五类提示词原则的情况下，大语言模型通过这一类原则化的提示使得响应结果得到了最大的改进。尤其是对于使用 GPT4 的情况下提升幅度最大，平均改进率到达了 75%。

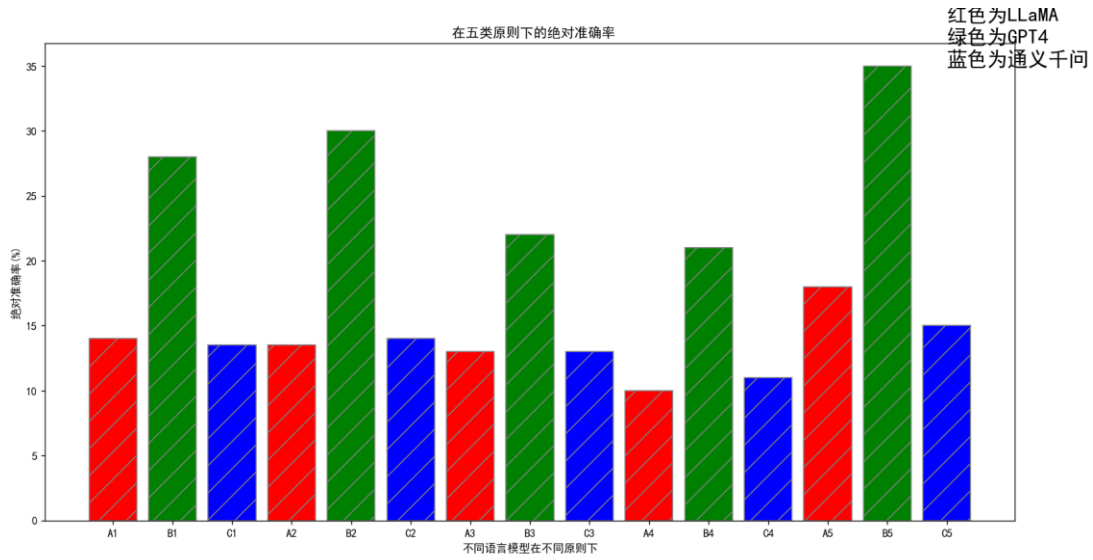


图十一：五类提示词原则分为五组的平均改进率，其中红色为 LLaMA 下的指标，绿色为 GPT4 下的指标，蓝色为通义千问下的指标。



图十二：五类提示词原则分为五组的绝对准确率，其中红色为 LLaMA 下的指标，绿色为 GPT4 下的指标，蓝色为通义千问下的指标。

准确率：（1）绝对准确率：在 LLaMA、GPT4 和通义千问上采用这五类原则时，本研究检查响应结果的准确率。红绿蓝三色柱体仍分别代表相应的大语言模型。只有纵坐标改为了绝对准确率。通常，这三类模型的绝对准确率在 20%~40%，如图十二所示。其中对于 LLaMA 和通义千问，绝对准确率基本可以在于 15%到 40%之间，而对于 GPT4，绝对准确率可以达到 40%以上。（2）相对准确率：图十三说明，平均对比不同模型，应用这五类提示词原则一般使相对准确度增加 10%左右。而对于 GPT4，这个相对准确度的增强可以超过 20%。



图十三：五类提示词原则分为五组的相对准确率，其中红色为 LLaMA 下的指标，绿色为 GPT4 下的指标，蓝色为通义千问下的指标。

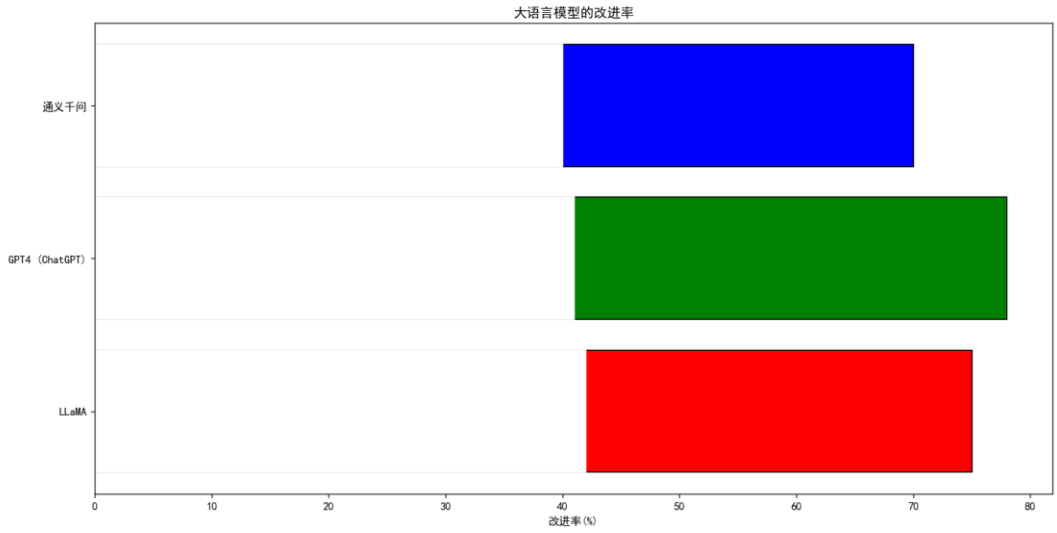
4.2.2 对于语言模型的单独结果

改进率：如图十四所示，本文在不同原则下使用单一模型进行实验，结果显示响应质量有了显著的改进。在三种大语言模型中，本研究观察到平均改进率接近 50%。

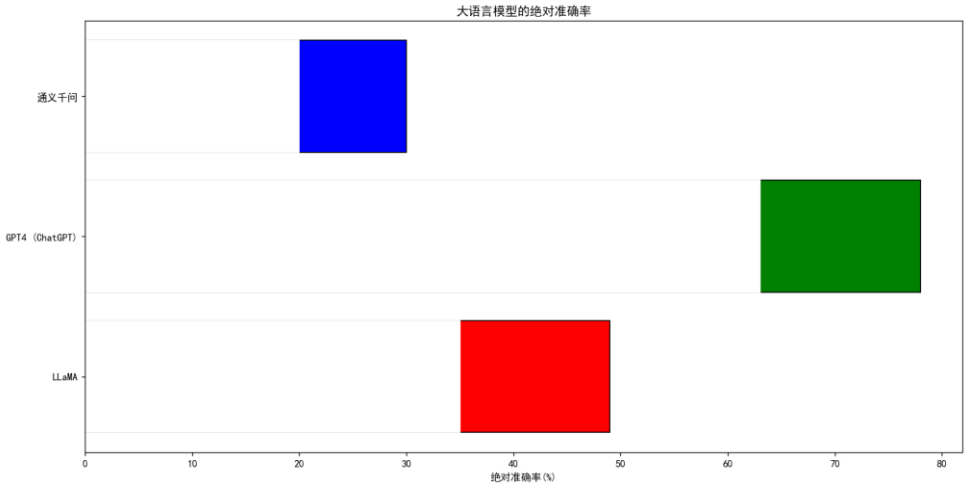
准确率：图十五详细展示了各模型的绝对准确率，而图十六则揭示了不同大小的模型在相对准确率上的增幅程度。从大到小，模型排序为 GPT4，LLaMA，以及通义千问。这些数据揭示了一个明显的趋势：模型的规模越大，其准确率改进的幅度就越大。

4.2.3 更多原则使用具体实例

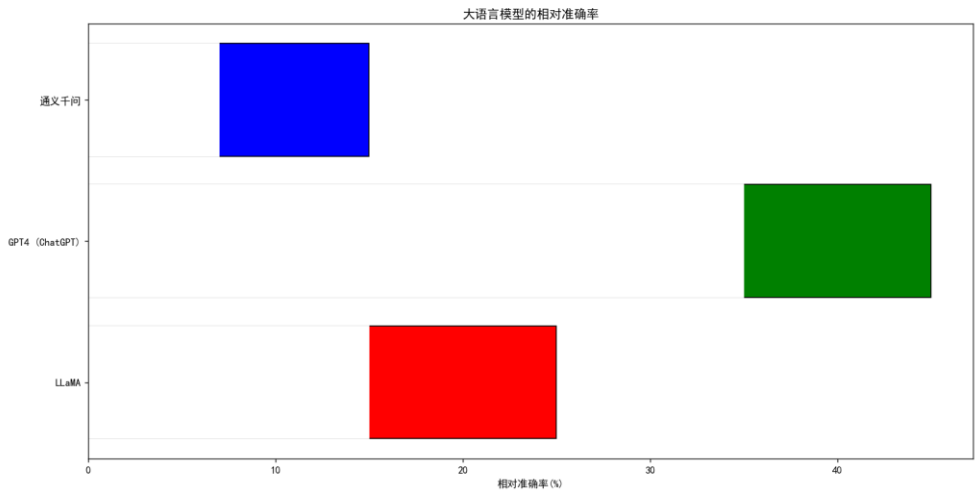
本研究在图十七，十八和十九中展示了更多不同提示词原则的在四类数学问题中的实验实例。从结果上来看，对提示词原则的使用显著提高了这些模型生成的响应的准确性和改进率。



图十四：改进率在三个模型下的区间柱状图



图十五：绝对准确率在三个模型下的区间柱状图



图十六：相对准确率在三个模型下的区间柱状图

In [4]:
```
review = f"""
这是一个关于学生作弊概率的问题。在这个问题中，教师给学生提出两个问题，一个是问题A：“你在考试中有过作弊的行为，是吗？”，另一个是问题B：

教师通过抽取一副共有13张的纸牌（A,2,3,…,Q,K）来决定学生需要回答哪个问题。如果抽取的是A至10的牌，那么学生需要回答问题A；如果抽取的是J,Q,
"""
```

In [5]:
```
prompt = f"""
得出review中P（A)的概率\
得出P（B)的概率\
得出P（C）的全概率公式\
得出学生作弊概率的数学模型公式

review: ```{review}```
"""

response = get_completion(prompt)
print(response)
```

根据review中的描述，我们可以得出以下概率：

P(A) = 10/13，即学生需要回答问题A的概率为10/13。

P(B) = 3/13，即学生需要回答问题B的概率为3/13。

根据全概率公式，我们可以得出：

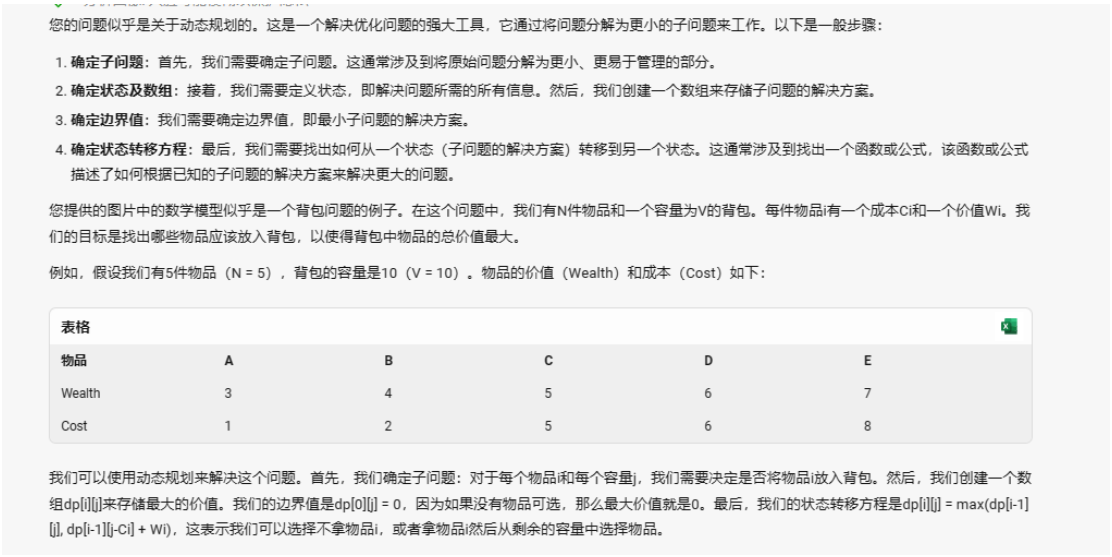
P(C) = P(A) * P(作弊|A) + P(B) * P(作弊|B)，其中P(作弊|A)表示在回答问题A的情况下作弊的概率，P(作弊|B)表示在回答问题B的情况下作弊的概率。

因此，学生作弊的概率的数学模型公式为：

P(作弊) = P(A) * P(作弊|A) + P(B) * P(作弊|B) = (10/13) * P(作弊|A) + (3/13) * P(作弊|B)。

图十七：LLaMA 在使用第五类原则的时候采用的分解成子问题的提问结果

图十八 使用所介绍的提示词原则以后在 GPT4 上的准确度的提升



您的问题似乎是关于动态规划的。这是一个解决优化问题的强大工具，它通过将问题分解为更小的子问题来工作。以下是一般步骤：

1. **确定子问题**：首先，我们需要确定子问题。这通常涉及到将原始问题分解为更小、更易于管理的部分。
2. **确定状态及数组**：接着，我们需要定义状态，即解决问题所需的所有信息。然后，我们创建一个数组来存储子问题的解决方案。
3. **确定边界值**：我们需要确定边界值，即最小子问题的解决方案。
4. **确定状态转移方程**：最后，我们需要找出如何从一个状态（子问题的解决方案）转移到另一个状态。这通常涉及到找出一个函数或公式，该函数或公式描述了如何根据已知的子问题的解决方案来解决更大的问题。

您提供的图片中的数学模型似乎是一个背包问题的例子。在这个问题中，我们有N件物品和一个容量为V的背包。每件物品i有一个成本Ci和一个价值Wi。我们的目标是找出哪些物品应该放入背包，以使得背包中物品的总价值最大。

例如，假设我们有5件物品（N = 5），背包的容量是10（V = 10）。物品的价值（Wealth）和成本（Cost）如下：

表格

物品	A	B	C	D	E
Wealth	3	4	5	6	7
Cost	1	2	5	6	8

我们可以使用动态规划来解决这个问题。首先，我们确定子问题：对于每个物品i和每个容量j，我们需要决定是否将物品i放入背包。然后，我们创建一个数组dp[i][j]来存储最大的价值。我们的边界值是dp[0][j] = 0，因为如果没有物品可选，那么最大价值就是0。最后，我们的状态转移方程是dp[i][j] = max(dp[i-1][j], dp[i-1][j-Ci] + Wi)，这表示我们可以选择不拿物品i，或者拿物品i然后从剩余的容量中选择物品。

图十九：通义千问在使用提示词原则原则解决多维背包问题（MKP）

4．3 实验结论

上述实验结果，可得知以下结论：本研究发现在所有的提示词原则中，第五类原则在大语言模型中表现最佳，尤其是在 GPT4 模型上，其改进率和准确率的提升幅度最为显著。这表明第五类原则，即"复杂任务分解成一系列简单子问题进行提问"的原则，非常适合用于处理自动数学建模任务。

这种原则的主要思想是将复杂的问题分解为一系列更简单的子问题，然后分步提出提示词解决这些子问题。基于这一种提示词策略的提示词如图十七所示。这种方法对解决数学建模任务特别有效，因为数学建模通常涉及到复杂的问题，通常需要通过分解和分步来理解和解决。

上述柱状图的对比（图十一，十二和十三）也直观的反应了对于大语言模型，第五类提示词原则对于自动建模的提升有着显著效果。本文希望这部分的实验内容能给想用大语言模型自动建模的人提供一些参考和帮助。

5 总结

本毕业论文通过详尽的分析，提出了五大类提示词原则，这些原则增强了大语言模型专注于输入上下文的关键要素的能力，从而生成高质量的响应。在处理输入之前，本研究通过这些精心设计的原则来指导大语言模型，以鼓励模型产生更好的响应。实验结果表明，这种策略可以有效地制定出可能影响输出质量的提示词，从而增强响应的相关性、简洁性和客观性。

未来研究的研究方向是多元化的。能够为读者提供一种全面而深入的见解，关于如何有效地使用大型语言模型以及如何通过精心计的提示词以提高其使用效率和准确性，也是本文的目的之一。本文相信，通过这样的研究，不仅可以提升人们的工作效率，还可以为未来的人工智能研究开辟新的道路。期待在未来的研究中，探究者能够发现更多的可能性以推动人工智能的发展。也希望本研究能够对此做出贡献以及让读者能够从这篇论文中获得启发。

参考文献

[1] Wei J, Wang X, Schuurmans D, et al. Chain-of-thought prompting elicits reasoning in large language models[J]. Advances in neural information processing systems, 2022, 35: 11.

[2] Zhou D, Schärli N, Hou L, et al. Least-to-most prompting enables complex reasoning in large language models[J]. arxiv preprint arxiv:2205.10625, 2022:11.

[3] Brown T, Mann B, Ryder N, et al. Language models are few-shot learners[J]. Advances in neural information processing systems, 2020, 33: 13.

[4] Shin T, Razeghi Y, Logan IV R L, et al. Autoprompt: Eliciting knowledge from language models with automatically generated prompts[J]. arxiv preprint arxiv:2010.15980, 2020 :13.

[5] White J, Fu Q, Hays S, et al. A Prompt Pattern Catalog to Enhance Prompt Engineering with ChatGPT[J]. arxiv preprint arxiv:2302.11382, 2023 :13.

[6] Bsharat S M, Myrzakhan A, Shen Z. Principled Instructions Are All You Need for Questioning LLaMA-1/2, GPT-3.5/4[J]. arxiv preprint arxiv:2312.16171, 2023:14.

[7] 蔡勇全. 例谈高中数学建模中的几种常见类型[J]. 中学数学杂志, 2016, (03) :20.

[8] Arora S, Narayan A, Chen M F, et al. Ask me anything: A simple strategy for prompting language models[C]//The Eleventh International Conference on Learning Representations. 2022: 20.

[9] Devlin J, Chang M W, Lee K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding[J]. arxiv preprint arxiv:1810.04805, 2018: 21.

[10] Hoffmann J, Borgeaud S, Mensch A, et al. Training compute-optimal large language models[J]. arxiv preprint arxiv:2203.15556, 2022:21.

[11] Imani S, Du L, Shrivastava H. Mathprompter: Mathematical reasoning using large language models[J]. arxiv preprint arxiv:2303.05398, 2023:21.

[12] Kamalloo E, Dziri N, Clarke C L A, et al. Evaluating open-domain question answering in the era of large language models[J]. arxiv preprint arxiv:2305.06984, 2023:22.

[13] Moens M F, Huang X J, Specia L, et al. Findings of the association for computational linguistics: Emnlp 2021[C]//Findings of the Association for Computational Linguistics: EMNLP 2021. 2021:23.

[14] Li R, Allal L B, Zi Y, et al. Starcoder: may the source be with you![J]. arxiv preprint arxiv:2305.06161, 2023:23.

[15] Regian J W, Shute V J, Shute V. Cognitive approaches to automated instruction[M]. Routledge, 2013:24.

[16] Li Y, Choi D, Chung J, et al. Competition-level code generation with alphacode[J]. Science, 2022, 378(6624):25.

[17] Li Z, Peng B, He P, et al. Guiding large language models via directional stimulus prompting[J]. Advances in Neural Information Processing Systems, 2024, 36:26.

[18] Pan R, **ng S, Diao S, et al. Plum: Prompt learning using metaheuristic[J]. arxiv preprint arxiv:2311.08364, 2023:26.

[19] Team G, Anil R, Borgeaud S, et al. Gemini: a family of highly capable multimodal models[J]. arxiv preprint arxiv:2312.11805, 2023:27.

[20]田军. 图书自动分类的数学建模及实现[J]. 图书情报工作, 2001 (09) :27.

致谢

经过几个月的努力，本论文在史玉回和赵琪老师的指导下完成，从开始论文选题到系统的实现，再到论文文章的实现，每走一步都是对自我的考验，从一无所知到一步步地探索再到完成论文，老师对我的论文指导是严谨认真负责的态度，提供科学合理的建议，让我在迷茫中看到希望。掌握了基本研究方向，在真正实践的过程中，发现并没有那么简单，在写作中遇到了很多的困难和障碍，思绪万千，也发现自身的不足之处并为之改正。同时，我也要感谢在完成论文的过程中，同学们给予的学习方法，资料等。

最后，由于我的学术水平有限，所写论文难免有不足之处，恳请各位老师和同学提出批评和指正。