

生成式AI的大模型提示工程：方法、现状与展望

黄峻^{1,2}, 林飞¹, 杨静^{3,4,5}, 王兴霞^{3,4,5}, 倪清桦^{1,2}, 王雨桐^{3,4}, 田永林^{3,4}, 李娟娟^{3,4}, 王飞跃^{1,3,5}

(1. 澳门科技大学创新工程学院工程科学系, 澳门 999078;

2. 青岛智能产业技术研究院, 山东 青岛 266109;

3. 中国科学院自动化研究所复杂系统管理与控制国家重点实验室, 北京 100190;

4. 中国科学院自动化研究所多模态人工智能系统全国重点实验室, 北京 100190;

5. 中国科学院大学人工智能学院, 北京 100049)

摘要: 大语言模型和视觉语言模型在各领域的应用中展示出巨大潜力, 成为研究热点。然而, 幻觉、知识迁移、与人类意图对齐等问题仍然影响着大模型的性能。首先, 探讨了提示工程与对齐技术基本原理, 提出基于提示优化、专家反馈机制及实时调整机制的引导概念, 提升了大语言模型在跨领域应用中的性能; 其次, 深入分析提示工程的核心技术, 如多步推理处理复杂任务的原理; 然后, 针对各领域的实际应用, 讨论提示工程的发展现状; 最后, 总结提示工程面临的挑战并展望其未来发展方向。提示工程在理论与应用方面的发展, 为提升大模型在实际应用中的性能提供了全面的解决方案。

关键词: 提示工程; 对齐技术; 生成式AI; 大语言模型; 视觉语言模型

中图分类号: TP18, V323.19

文献标志码: A

doi: 10.11959/j.issn.2096-6652.202424

From prompt engineering to generative artificial intelligence for large models: the state of the art and perspective

HUANG Jun^{1,2}, LIN Fei¹, YANG Jing^{3,4,5}, WANG Xingxia^{3,4,5}, NI Qinghua^{1,2}, WANG Yutong^{3,4},
TIAN Yonglin^{3,4}, LI Juanjuan^{3,4}, WANG Fei-Yue^{1,3,5}

1. Department of Engineering Science, Faculty of Innovation Engineering, Macau University of Science and Technology, Macao, 999078, China

2. Qingdao Academy of Intelligent Industries, Qingdao 266109, China

3. State Key Laboratory for Management and Control of Complex Systems, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China

4. State Key Laboratory of Multimodal Artificial Intelligence Systems, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China

5. School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing 100049, China

Abstract: Large language models and vision-language models have demonstrated significant potential in various downstream applications, making it become a research hotspot. However, the issues such as hallucinations and knowledge transfer impact the performance of these models. Firstly, this paper explores the fundamental principles of prompt engineering and alignment techniques, and proposes the concept of “prescriptive”, which is based on optimizing prompts and expert feedback verification and can be adjusted in real-time. This aims to further enhance the performance of large language models in cross-domain applications. Secondly, the core technologies of prompt engineering, such as the principles of multi-step reasoning for handling complex tasks, are analyzed in depth. Additionally, the development status of prompt

收稿日期: 2024-04-07; 修回日期: 2024-05-26

通信作者: 王飞跃, feiyue.wang@ia.ac.cn

基金项目: 澳门特别行政区科学技术发展基金项目 (No.0145/2023/RIA3); 北京市自然科学基金-怀柔创新联合基金项目 (No.L245025)

Foundation Items: The Science and Technology Development Fund of Macao Special Administrative Region (No.0145/2023/RIA3), Beijing Natural Science Foundation Huairou Innovation Joint Fund Project (No.L245025)

engineering is discussed based on practical applications in various fields. Finally, this paper summarizes the challenges faced by prompt engineering and looks into its future development directions. The development of prompt engineering in theory and application, provide comprehensive solutions for improving the performance of large models in practical applications.

Key words: prompt engineering, alignment technology, generative artificial intelligence, large language models, vision-language models

0 引言

从20世纪50年代发展至今,自然语言处理(nature language processing, NLP)已从解决文本翻译任务扩展至解决多模态复杂任务^[1]。大语言模型(large language models, LLMs)经历了从早期用于计算词语出现概率和统计关系的统计语言模型(statistical language models, SLMs)^[2-3],到捕捉词语语义关系的神经语言模型(neural language models, NLMs)^[4-7]的发展,其机器翻译和文本处理等NLP能力显著提高。随着Transformer架构的引入,预训练语言模型(pre-trained language models, PLMs),如OpenAI的GPT^[8]和Google的BERT^[9],通过在大规模文本数据上进行预训练,彻底改变了NLP的研究和应用。2020年开始,LLMs(如GPT-3^[10]和LLAMA^[11])通过更大规模的预训练数据和模型参数,使PLMs实现更强的生成和理解能力^[12-13],进一步推动了从PLMs到LLMs的发展^[14]。

视觉-语言模型(vision-language models, VLMs)^[15-16]结合图像和文本数据,提升生成式AI(generative AI, GenAI)对复杂场景的理解和生成能力。早期的VLMs大多采用单独训练的图像模型(如卷积神经网络)和文本模型(如递归神经网络),在特征层进行融合,这种方法虽然能够处理

简单图像标注及文本生成任务,但在复杂场景下能力有限^[17]。随着Transformer架构的引入,VLMs被扩展至复杂多模态融合任务。Flamingo^[18]结合视觉和语言输入生成高质量的文本描述;CLIP(contrastive language-image pre-training)^[8,19]对比学习大量图像和文本对,通过自然语言描述高效地检索和分类图像,在多模态理解中表现出强大能力;DALL-E不仅能够根据文本描述生成高质量的图像,还能够理解和创造复杂的多模态内容^[20-21]。随着GenAI技术的发展,这些大模型在创意内容生成、自动化编写和复杂任务处理方面展现出前所未有的能力。

然而,幻觉问题和知识迁移问题严重影响了大模型在下游任务中的性能^[22]。现有研究通过微调技术来解决以上问题,该过程通常包括数据选择、超参数调整、训练和优化(如Dropout、正则化、交叉验证等),能够提高LLMs生成内容的准确性。但是庞大的模型的微调需要消耗大量的计算资源和时间,此外,过拟合也可能导致模型在通用任务上表现较差。如何解决大量数据的标记耗费大量资源、调整模型参数复杂度高的问题,已成为提升传统机器学习模型和预训练模型扩展性的关键。为解决以上问题并提升大模型性能,提示工程成为了一个研究热点。图1展示了提示和提示工程的一个示例。

提示:	提示工程:
用户输入: 生成一段关于人工智能发展的文章。 LLMs输出: 在21世纪的技术浪潮中,人工智能(AI)的快速发展成为了一个标志性的现象。 ... 然而, AI技术的发展也带来了伦理和社会问题,如就业置换、数据隐私泄露和算法偏见等问题。 ... 综上所述,人工智能作为一种强大的技术工具,其发展前景广阔,将继续在全球科技进步和社会发展中扮演重要角色。同时,我们也需要警惕其潜在的负面影响,确保这项技术的发展符合人类的长远利益和伦理标准。	步骤1: 设计一个初始提示: 生成一段关于人工智能发展的文章, 包括其历史和未来趋势。 步骤2: 评估模型生成内容是否符合要求。 步骤3: 根据评估结果优化提示, 例如增加具体要求: 生成一段关于人工智能发展的文章, 涵盖1950年至今的重要里程碑, 并预测未来十年的发展趋势。

图1 提示与提示工程示例

提示工程和大模型同时出现，能够在不通过任何梯度更新或调整模型核心参数的情况下，通过精心设计的少量示例或零示例来适应新任务，如手动创建的自然语言指令、自动生成的自然语言指令或向量表示，通过提供与任务相关的上下文指令来提高模型的任务适应性和通用性。最初OpenAI使用自然语言作为提示来测试GPT-2的翻译功能，随着大模型技术与提示工程的发展，提示逐渐扩展为图片、音频等数据^[23-24]。提示作为人类与大模型交互时的输入，能够指导大模型生成符合任务要求的输出结果。LLMs和VLMs的3类主要应用如图2所示，其中，多模态-文本生成模型（如LLAMA^[11]，Flamingo^[18]），通常以自然语言及图片作为输入，生成文本描述；图像-文本匹配模型（如CLIP^[19]），通常用于匹配文本与相应图片；文本-图像生成模型（如DALL-E^[20-21]），通常以自然语言作为输入，生成相应图片。在自然语言处理中，提示可以是一段描述性文字，用来引导LLMs回答问题、生成文本或执行其他任务；在图像生成任务中，VLMs可以根据用户文本描述生成对应的图像。这种方法减少了大规模数据的标注需求，仅通过提供与任务相关的上下文和指令来引导模型输出，提升了LLMs和VLMs在多种任务中的适应性，并有效解决了GenAI的幻觉问题。

此外，对齐作为LLMs调优的另一关键技术，与提示工程密切相关且相辅相成，两者共同作用，能够显著提升模型的可靠性和安全性^[25-26]。对齐技术使模型的输出符合人类的期望、价值观和社会规范，避免生成有害或有偏见的信息。为有效结合提

示工程与对齐技术，本文提出引导（prescriptive）概念，旨在提升大模型在跨领域应用中的性能，首先通过优化提示寻找有效提示方案，然后通过专家反馈机制帮助识别和修正模型在任务执行中的不足之处，确保输出的准确性和一致性，最后在模型运行过程中，根据实际情况实时调整提示。

提示工程现有的综述性文章涉及多个领域，White等^[27]介绍提示工程在自动化软件开发任务中的应用，并构建自动化提示模式框架，介绍16种应用于改进LLMs对话-回复方法的模式，并展示如何将不同模式应用于不同领域。Zhou等^[28]分析了自动和手动提示生成与选择的方法，提出了自动提示工程师（automatic prompt engineer，APE）算法，用于自动生成和选择最优的提示。APE算法将提示生成视为程序合成问题，利用预训练的大型语言模型生成候选提示，并通过迭代蒙特卡洛搜索方法优化提示选择，显著提高了模型在任务中的性能。Liu等^[24]讨论了提示工程的演变、分类以及其在不同领域的应用。Gu等^[29]专注研究VLMs中的提示工程，探讨了提示工程在多模态-文本生成模型、图像-文本匹配模型和文本-图像生成模型中的应用。Schulhoff等^[30]建立了58种文本提示技术分类法，介绍了包含图像、音频、视频在内的多模态提示技术，还涵盖了提示工程技术（如自动提示（AutoPrompt）生成、提示优化等）的实际应用案例。尽管以上综述对提示工程或对齐技术的概念和方法进行了深入探讨，但缺少对两种技术在大模型中的作用机制的综合性分析。本文旨在分析提示工程与对齐技术的核心概念，并提出引导框架，为科

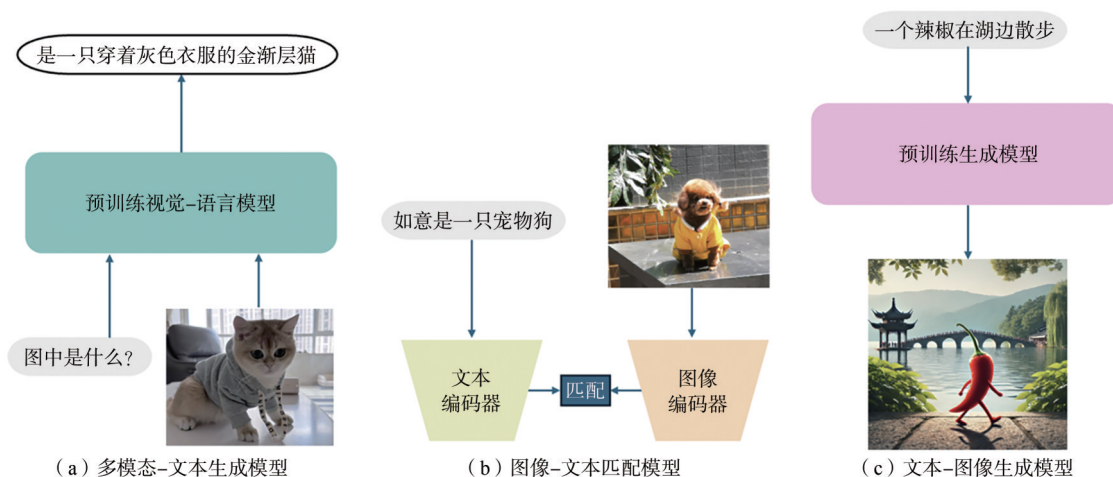


图2 LLMs和VLMs的3类应用

研人员提供构建、评估和执行提示工程的有效方法,从而提升大模型在多种任务中的性能。

本文结构如下:第1节介绍提示工程基本概念与优化方法、外部/内部对齐技术方法及引导概念;第2节梳理零样本/少样本提示、逐步思考与推理、提示自动生成与优化以及检索增强提示等提示工程核心技术;第3节探讨提示工程在自然语言处理、视觉与多模态任务、推理与决策等多模态任务中的实际应用;第4节总结提示工程面临的挑战并展望未来,以推动研究者对这一领域的进一步探索。

1 提示工程基础

1.1 基本概念与优化方法

提示工程是一类通过设计和优化输入提示语,来引导和约束大模型输出的方法,无须调整模型参数。其核心在于通过精心设计的提示,一方面使模型更准确地理解用户需求,另一方面使得模型更加适配任务,从而生成符合预期的内容。提示工程可以分为硬提示和软提示^[29]:硬提示通常是自然语言预定义文本或模板,适用于静态任务^[31];软提示是在提示微调过程中生成的嵌入向量,适用于动态任务^[32]。根据交互方式,提示可以分为在线提示^[33]和离线提示^[34],前者用于实时交互,后者用于批量处理。此外,提示工程还根据应用领域分为信息检索、文本生成、翻译、情感分析、编程与代码生成、对话、特定任务和自定义应用等,优化和定制提示可以提升大模型在各类任务中的效果和适应性^[35]。

在手动提示构建过程中,提示优化的技巧包括:明确问题、提供上下文^[36]、示例导向^[37]、逐步引导^[38]和限定范围^[39]等。明确问题是通过清晰、具体的问题描述,使模型准确理解任务,例如将“解释一下人工智能”改为“用简明的语言解释一下什么是人工智能,尽量控制在两到三句话以内”,这样便于模型生成简洁明了的回答。提供上下文是为模型提供必要的背景信息,模型可根据更丰富的上下文理解和生成内容^[36],例如将“写一封给客户的道歉信”改为“客户因产品配送延误而不满。请写一封道歉信,内容应包括对延误的解释、诚恳的道歉以及补救措施”,这样生成的道歉信会更加贴近实际需求。示例导向是通过提供示例或模板指导模型生成类似风格或结构的输出^[37],例如将“编一个笑话”改为“编一个关于程序员的

笑话。例如:‘为什么程序员喜欢早起?因为他们需要打代码!’”,通过具体的示例引导模型生成类似风格的笑话。逐步引导是将复杂任务分解为一系列简单的步骤,每一步都有明确的提示和预期输出,例如思维链(chain-of-thought, CoT)技术^[38]有助于模型逐步生成复杂的解决方案,例如将“解释如何做一个蛋糕”改为“请按以下步骤解释如何做一个巧克力蛋糕:(1)准备材料;(2)混合干湿材料;(3)烘烤;(4)装饰”,使得模型的解释更加清晰、易懂。限定范围是通过明确限制模型生成内容的范围和风格控制模型的输出^[39],例如将“写一篇关于环境保护的文章”改为“写一篇关于减少塑料使用的环境保护文章,字数控制在300字以内”,这样生成的内容会更加具体。这些方法被广泛应用于问答系统、内容生成、对话系统等自然语言处理任务中,是大语言模型应用的一种重要技巧。提示工程通过这些方法显著提升了大语言模型在各类任务中的表现。

自动化提示优化的主要技术包括:元提示(meta-prompting)^[40],通过设计提示来优化或生成其他提示,从而提升提示工程的整体效果;自动提示^[41],使用包含触发词的提示模板进行反向传播更新;自动提示工程师^[42],生成并迭代优化零样本指令提示;无梯度指令提示搜索(gradient-free instructional prompt search, GrIPS)^[43],通过删除、添加、交换和改写创建提示变体;文本梯度提示优化(prompt optimization with textual gradients, ProTeGi)^[44],通过批评和贪婪算法多步骤改进提示;基于对话的策略梯度离散提示优化(dialogue-comprised policy-gradient-based discrete prompt optimization, DP2O)^[45],结合强化学习和对话构建提示。这些方法通过生成和改进提示模板显著提高了提示的质量和模型的任务完成能力。

1.2 大模型中提示工程、对齐、引导的作用机制

大模型存在以下问题:首先,不良内容的问题尚未得到充分解决,尽管训练数据量显著增加,但其中的偏见^[46]、有害内容和隐私问题仍然存在^[47],未对齐的大模型可能生成带有性别、社会、文化和种族偏见的内容,甚至会涉及隐私泄露;其次,大模型还倾向于生成不真实甚至虚构的内容,包括错误^[48]、幻觉^[49]和不一致性^[50],这些问题不仅影响了大模型在一般领域的可信度,也限制了它们在医学和法律等专业领域的应用;此外,大模型还可能

被用于虚假信息传播、网络攻击，甚至生成致命武器等非法用途，可能对社会产生负面影响。

针对以上问题，提示工程在大模型中通过提供与任务相关的上下文信息、控制输出格式、约束生成内容、提升任务适应性、错误识别与修正以及改进交互方式，提升大模型的任务适配性。用户按需设计提示时，可以向模型提供明确的任务指令和格式要求，确保生成内容的准确性和一致性。同时，提示工程可以通过设置约束条件来限制模型生成的内容范围，减少错误信息的产生。此外，通过少样本学习或零样本，提示工程提高了模型在多任务环境中的适应性，使其能够快速应对新任务。提示工程还可以通过特定提示引导模型进行自我检查和验证，减少错误生成。最终，通过优化人与模型的交互方式，提示工程提升了用户体验和模型响应效率，为用户提供了一种灵活有效的方法，使模型能够在不调整核心参数的情况下处理多样化的任务需求^[35]。

指令微调作为提升大模型性能的另一技术，在实际应用中与提示工程存在一定的差别。指令微调通过调整模型的权重，使其能够更好地满足特定的数据集和任务需求，从而确保数据治理效果和减少LLM响应错误^[51-52]。指令微调能够更深入地优化模型以满足特定需求和上下文，在实际应用中能够达到较高的准确性和可靠性。使用这两种技术，可以在特定应用场景中更有效地利用LLMs，确保模型输出既符合预期，又具备高可靠性和安全性。

对齐使得大模型的行为和输出符合人类用户的期望、价值观和安全规范，其目标是减少或消除模型生成有害、不准确或误导性信息的风险，提升其在实际应用中的可靠性和安全性。AI对齐可以追溯到AI革命的初期，当时的目标是创建能够像人类一样思考和行动的机器，并确保这些机器的行为与人类的最佳利益保持一致。Wiener强调了机器目标与人类意图一致性的重要性^[53]，即“在拥有足够的数据之前需要确保机器的目的符合我们的期望，否则在机器启动后便难以干涉”。2014年，Russell提出了“价值对齐问题”，强调了构建与人类价值观一致的智能系统的必要性^[54]，旨在构建“以人为本”的智能而非纯粹的智能。

尽管AI对齐的概念早已存在，但近几十年里鲜有对其的研究。随着大语言模型的崛起以及AI能力的大幅提升，对齐问题的重要性和紧迫性也随

之增加。对齐技术分为外部对齐和内部对齐，用于解决LLMs在不同层面上的对齐问题^[27]，如图3所示。其中，外部对齐关注的是模型的训练目标与人类的价值观的一致性，主要涉及如何将复杂和多样的人类价值观转化为模型的优化目标。而内部对齐关注的是在模型的实际优化过程中，模型内部推导或优化的目标与设计者的初始目标的一致性，防止模型在处理未见过的情况时表现出不可预测或不符合预期的行为。总之，外部对齐确保模型目标的设定符合人类价值，内部对齐确保模型在实现这些目标时的推理过程和决策操作与设计者的初始目标一致，两者共同保障模型的可靠性和安全性。

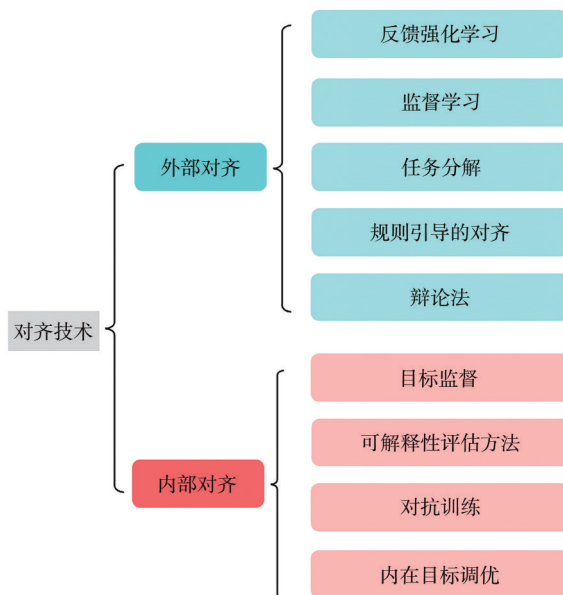


图3 外部对齐与内部对齐技术

外部对齐常用的方法包括人类反馈强化学习（reinforcement learning from human feedback, RLHF）^[55-56]、监督学习^[57]、任务分解^[58]、规则引导的对齐^[59]、辩论法^[60-61]。其中，RLHF根据人类反馈调整模型的行为，通过奖励和惩罚机制使模型的输出更加符合人类的期望和价值观。监督学习通过使用大量标注数据训练模型，使模型在特定任务上表现出色，减少生成不良内容的可能性。任务分解将复杂任务分解为更小的子任务，并对每个子任务进行单独优化，确保整体目标的实现。规则引导的对齐通过明确模型在生成内容时需要遵守的准则和限制，从而引导模型行为。辩论法通过多个模型的辩论确定最符合人类价值观的输出^[61]。外部对齐通过这些方法确保模型行为的伦理性和社会责

示^[72-73]为GenAI分配一个特定的角色，使其在开放任务中生成更贴切情景以及更符合预期的输出，比如提示其扮演“历史学家”或“乔布斯”等；风格提示^[74]通过指定所需的风格、语气或体裁，使GenAI的输出更专注于内容的表现形式，并保证生成的内容满足用户个性化的风格要求；情感提示^[75-76]通过加入具有心理暗示的短语引导模型生成人性化的回应，同时给予用户适当的心理暗示，例如为大模型提出“这件事对于我的职业规划很重要”。

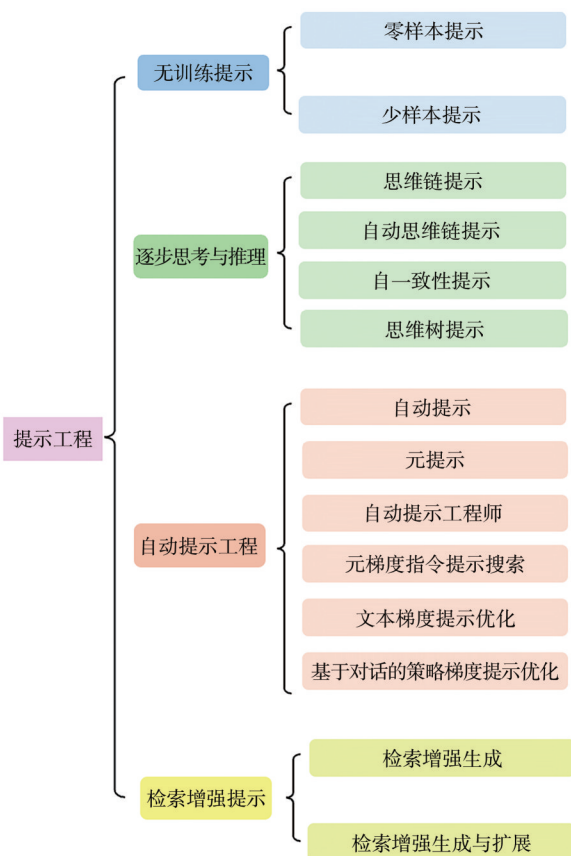


图5 提示工程的前沿技术

在优化生成结果一致性与准确性方面，Wilf等^[77]提出SimToM方法，用于在回答多元问题时消除不相干信息的影响，确保答案基于准确的事实，保证GenAI在处理多实体问题时的一致性和准确性；Weston等^[78]提出的S2A方法、Deng等^[79]提出的RaR方法以及Xu等^[80]提出的RE2方法为GenAI提供重写提示，从而去除与问题无关的信息，并将精简后的新提示传递给模型以获取最终响应，使模型更专注于核心问题，适用于复杂问题的处理；此外，Press等^[81]提出的自问自答方法通过提示大模型决定是否为给定提示生成后续问题，LLMs自己

提出并回答问题，从而解决最初的问题，这一过程使模型能够全面地理解复杂问题。

这些零样本提示方法通过不同的策略优化提示内容，显著提升了LLMs在各种任务中的性能和准确性。这些技术不仅拓宽了模型的应用范围，还为应对复杂任务提供了有效的解决方案。

(2) 少样本提示

少样本提示在零样本提示的基础上，通过提供少量的任务示例进一步优化模型的性能。Kaplan等^[5]提出当模型足够大时，它能够从少量示例或文本指令中学习并执行新的任务，这种特性被称为少样本学习。少样本提示利用上下文学习（in-context learning, ICL）帮助模型更好地理解任务要求和模式，从而生成更准确的答案或执行更有效的任务^[10]。目前，上下文学习已经在自然语言理解^[82-85]、数据生成^[86-88]和推理^[89-91]等各种下游任务中取得良好效果。值得注意的是，少样本学习^[92]是一种更广泛的机器学习范式，通过少量示例调整参数，而少样本提示特定于GenAI设置中的提示，并不涉及更新模型参数。对于少样本提示，模型性能很大程度上受到给定示例的影响，且其输入长度受限，因此，示例的选择及其排列顺序显得尤为重要，必须考虑示例的数量、顺序、标签分布、标签质量、模板格式和相似性等6个因素。对于示例的数量，一些研究^[10,93]表明增加提示中的样本数量能够提高模型的性能，特别是在较大的模型中，但是在某些情况下，当样本数量超过20个时，模型性能的提升可能会减少。对于示例的顺序，Lu等^[94]观察到在某些任务中示例顺序可能引起准确率从低于50%到高于90%的剧烈变化。另外，示例标签分布不均可能导致模型倾向于样本数量更多的标签类别。而对于标签质量对模型性能的影响，目前还没有一个统一的结论，要根据不同的情景具体分析。Min等^[95]认为标签的准确性对于模型输出内容的质量是无关紧要的，但Yoo等^[96]认为在某些设置下，错误的标签会对性能产生显著的负面影响。模型性能对于模板格式的选择也是敏感的。Jiang等^[97]表明使用训练数据中常见的格式能够使模型达到更好的性能。最后，基于示例的相似度计算也能影响模型性能，权衡样本的多样性和相似性也是非常重要的^[93,98-99]。基于这些因素，K近邻^[93]、基于图的无监督选择性注释方法（Vote-K）^[99]、自生成ICL^[100]、提示挖掘^[97]、先过滤后搜索（filter-

then-search, LENS) 方法^[101]、统一演示检索器(unified demonstration retriever, UDR)^[102]、强化学习^[103]等技术被应用于为模型生成最佳示例。特别是, 自生成 ICL^[100]利用 GenAI 自动生成示例, 其在训练数据不可获得的情况下的表现优于零样本场景, 但在有监督场景中, 生成的示例不如实际的示例有效。

2.2 逐步思考与推理

有大量提示工程研究集中在利用 LLMs 将复杂问题分解为更简单的子问题进行推理求解, 这对于人类和 GenAI 都是一种有效的解决策略, 因为明确地分解问题已被证实可以进一步提高 LLMs 的问题解决能力。本节针对算术、常识和符号等多步推理任务, 对目前主流的逐步思考与推理提示技术进行了探讨。

(1) 思维链提示

在面对复杂的逻辑推理任务时, LLMs 常常表现不佳。为了突破这一限制, 文献[89]提出了思维链提示技术, 旨在引导 LLMs 进行连贯、逐步的推理过程。文献[89]的实验结果表明, 相较于传统提示方法, 思维链提示更能激发 LLMs 产生结构化和深入思考的回答, 在算术、常识和符号推理基准测试中取得了显著的性能提升, 只需用 8 个 CoT 示例提示 PaLM 540B^[104]模型即可在 GSM8K 数学推理基准测试集^[105]上达到最先进的准确率, 超过了微调的 GPT-3。

(2) 自动思维链 (Auto-CoT) 提示

人工手动创建高质量的 CoT 示例既耗时又易出错。Zhang 等^[42]引入了 Auto-CoT, 通过提示 LLMs “逐步思考”来生成推理链。Auto-CoT 对任务聚类, 从每一簇中采样代表性问题, 使用启发式规则的零样本 CoT 生成推理链, 形成最终的示例集。多样性问题有效缓解了零样本 CoT 可能产生的错误, 在 10 个公共基准推理数据集上的实验结果表明, 使用 GPT-3 时 Auto-CoT 的性能始终与需要手动设计演示的 CoT 持平或优于 CoT。

(3) 自一致性提示

CoT 提示的本质是采用一种贪心解码方法, 在每个时间步选择概率最高的词作为输出。为了进一步提升推理性能, 文献[106]提出了自一致性解码策略。复杂问题通常被允许采用多种不同的思考方式来得到唯一正确的答案。基于这一思想, 自一致性利用语言模型的解码器生成一组多样化的推理路

径, 每条推理路径可能导致不同的最终答案, 然后采用多数投票方法识别最终答案集中最一致的答案。在各种基准测试中, 自一致性和思维链提示的结合显著提高了推理的准确性, 相比于基线思维链提示, 其在 GSM8K 测试准确率提高了 17.9%, 在 SVAMP^[107]上提高了 11.0%, 在 AQuA^[108]上提高了 12.2%, 在 StrategyQA^[109]上提高了 6.4%。

(4) 思维树 (tree-of-thoughts, ToT) 提示

文献[110]提出了思维树框架, 以增强链式思维在需要探索和前瞻性推理的复杂任务中的提示能力。ToT 通过管理中间推理步骤 (称为“思维”) 的树结构扩展 CoT 提示, 每种思维代表一个向最终解决方案推进的连贯语言序列, 这种结构允许 LLMs 通过判断当前思维步骤向最终解决方案前进了多少以及它继续向前推进以达到目标的可能性进行推理。ToT 将模型生成和评估思维的能力与搜索算法 (如广度优先搜索或深度优先搜索) 结合, 从而实现对推理链的系统性探索, 发现更有效的解决方案, 并在解决方案错误时回溯。在 24 点游戏任务中, ToT 的成功率达到了 74%, 而 CoT 仅为 4%。

2.3 自动提示工程

虽然人工构造的提示显著提升了大模型在通用任务上的性能, 但在特定领域或处理复杂任务时, 这种方法不仅易受到人类经验局限性和主观性的制约, 还高度依赖耗时费力的试错过程。因此, 如何通过自动化的方法实现提示的高效生成与优化, 推动大模型在不同场景下的快速应用, 具有一定的研究价值。

(1) 自动提示 (AutoPrompt)

基于梯度搜索方法, 自动提示^[41]是一种使用包含触发词的提示模板对提示进行反向传播的更新算法。此算法自动生成的提示 x_{prompt} 的模板为:

$$x_{\text{prompt}} = \lambda(x_{\text{inp}}, x_{\text{trig}}) = \{\text{输入}\}[T][T][T][T][P](1)$$

其中, x_{inp} 为输入, $[T] \in x_{\text{trig}}$ 为触发词, $[P]$ 代表 x_{prompt} 的掩码, λ 是将 x_{inp} 映射到 x_{prompt} 的模版。因此, 对于给定的任务输入 x_{inp} , AutoPrompt 可以利用 λ 和梯度搜索方法得到优化后的提示 x_{prompt} 。

(2) 元提示

为了自动生成提示, Reynolds 等^[40]于 2021 年提出了元提示的概念, 希望通过提示自动生成的方式提高大模型在特定任务上的工作效率和精度。其中, 元提示可以是类似“这个问题要求 (This

problem asks us to)”或“把这个问题分成几个步骤来解决(Let’s solve this problem by splitting it into some steps)”的短语，通过对问题意图的陈述逐步解决问题。此外，元提示也可能采用填空模板的形式，这种形式限制了回答的方向，但允许大模型填充与特定问题相关的具体细节。为了解决元提示在复杂推理任务中存在的不足，Ye等^[111]为元提示引入了3个额外的概念，分别是详细的任务描述、上下文规范和逐步推理。实验结果表明，相较于CoT中“逐步思考”的提示，改进后的元提示提高了复杂推理任务的精度。

(3) 自动提示工程师

受到经典程序合成技术和人类提示工程方法的启发，Zhou等^[28]提出了自动提示工程师方法，用于提示的自动生成与筛选。首先，将提示视为需要优化的“程序”，在大模型生成的指令候选集中寻找最大化特定评分函数的指令。然后，为了评估筛选得到的提示的有效性，在另一个大模型接收到该提示后，测试其在零样本情况下的表现。自动提示工程师最优提示筛选过程的数学模型为：

$$\rho^* = \arg \max_{\rho} f(\rho) = \arg \max_{\rho} E_{(Q,A)}[f(\rho, Q, A)] \quad (2)$$

其中， ρ 为提示， ρ^* 为筛选得到的最优提示， $D_{\text{train}} = \{(Q,A)\}$ 为特定任务的输入/输出数据集， $f(\rho)$ 和 $f(\rho, Q, A)$ 分别为以 ρ 和 (ρ, Q, A) 为输入的打分函数， $E_{(Q,A)}$ 为期望。

(4) 无梯度指令提示搜索

现有基于梯度搜索优化提示的方法大幅增加了大模型的计算需求，带来了高昂的计算成本。此外，由于无法直接访问模型的参数信息，一些仅可通过接口调用的大模型不能利用梯度搜索算法对提示进行优化。因此，Prasad等^[43]提出了一种无梯度的指令提示搜索算法GrIPS，通过删除、添加、改写等操作来构建和优化提示。首先，对于某个初始提示，在每轮迭代中随机选取一种编辑操作，生成 m 个新的候选提示。这一过程表明每轮迭代一共产生了 $m \times 1$ 个采样操作。然后，基于模型在测试集上的表现，对这些新生成的候选提示进行评分。若最优候选提示的评分超过当前基础提示的得分，则将其指定为下一轮迭代的基础提示，否则，继续使用当前的基础提示进行后续的搜索。在整个过程中，评分函数Score为：

$$\text{Score} = \text{BalancedAccuracy} + \alpha H,$$

$$H = \sum_{y \in Y} -p_y \log(p_y), p_y = \frac{1}{|S|} \sum_{i=1}^{|S|} 1(\hat{y}_i = y) \quad (3)$$

其中， S 为评分集合， Y 为任务的标签空间， y 为某一任务的标签， $\hat{y}_i$ 是标签的预测结果， H 表示熵， α 为缩放因子（通常取10），BalancedAccuracy的取值范围一般为[0,100]。

(5) 文本梯度提示优化

借鉴基于文本的苏格拉底式对话中的梯度下降思想，Pryzant等^[44]提出了ProTeGi，用大模型的反馈代替微分，用大模型的编辑功能代替反向传播过程。首先，采用训练数据的小批量子集生成自然语言描述的“梯度”，即用小批量数据集指出当前提示存在的问题。然后，在与梯度相反的语义方向上编辑当前提示。这些操作使得束搜索（beam search）可以在更大的提示空间中进行，将候选节点的选择问题转化为最佳臂识别问题，显著提升了算法的整体效率。实验结果表明，相较于APE和GRIPS，基于ProTeGi的提示生成和优化的性能得到了提升。

(6) 基于对话的策略梯度离散提示优化

为了解决目前离散提示优化存在的高成本、低效率和强主观性等问题，Li等^[45]提出了一种基于对话的离散提示优化方法DP2O。该算法主要包含提示集构建和提示匹配两个阶段。在提示集构建阶段，该算法基于GPT4提出了一种基于多轮对话对齐策略的提示集生成方法，并引入了一种新的提示质量评估方法——监督和无监督熵度量（supervised and unsupervised entropy metric, SUE）。因此，SUE能够筛选具有线性复杂度的高质量提示。在提示匹配阶段，该算法采用了一个强化学习框架，然后利用策略网络选择合适的输入提示，最后综合上述决策完成特定的下游任务。

2.4 检索增强提示

提示工程的核心任务是为大模型提供与任务相关的信息和引导，以增强大模型针对特定任务的响应能力。但在对信息的专业性、隐私性及实时性要求较高的领域中，依赖手动或自动设计的提示难以提供有效的上下文信息，进而无法支撑大模型进行精准理解和决策。此外，传统提示工程主要依赖预定义的提示模板和有限的上下文信息，在处理复杂和多变的任务时，其表现往往不尽如人意^[112]。

检索增强生成方法（retrieval-augmented genera-

tion, RAG) 及其拓展方法 SAGE (search-augmented generation and extension) 为动态构建包含与查询相关的丰富上下文信息的提示提供了思路。SAGE 的核心原理如下: 首先, 通过查询分析构建主动的内容搜索机制, 实现对相关信息的挖掘和提取; 其次, 通过内/缓存系统的分层管理, 实现外部数据源和 LLMs 有限上下文窗口信息之间的高效交换; 然后, 针对多样的任务形式, 构建不同的提示模板, 包括但不限于检索增强思维、检索增强推理和检索增强行动等, 并将上下文信息融入提示中。

通过这种方法, SAGE 在提示工程中表现出显著的优势。首先, 基于搜索的主动信息挖掘机制能够为提示构建引入更丰富和相关的上下文信息, 显著提升了 LLMs 的上下文学习能力。其次, 借助内容的多级分层管理, SAGE 能够动态调整和优化提示, 在处理繁琐任务时, 依旧能保持提示构建过程的高效性。此外, SAGE 的检索增强机制拓宽了提示工程的应用范围, 能够支持更广泛的任务类型, 提高了 LLMs 的通用性和灵活性。SAGE 的检索增强机制不仅提升了 LLMs 在不同应用场景下的性能, 也为智能体 (尤其是具身智能体的知识系统) 提供了强有力的支撑。

当前, 检索增强型提示已经在自然语言处理^[112]、图像视频分析^[113]以及多模态任务^[50,114-115]等通用任务中得到了广泛应用, 并在法律^[116]、医疗^[117]和自动驾驶^[39,118-122]等垂直类应用中崭露头角。

3 提示工程应用

基于第2节提示工程的前沿技术, 本节将系统介绍提示工程在自然语言处理、视频与多模态任务、推理与决策等领域的应用。

3.1 自然语言处理

3.1.1 文本生成与理解

近年来, 提示工程在文本生成与理解领域取得了显著进展, 特别是在结合预训练语言模型方面。精心设计复杂且多样化的提示, 可以显著提升生成文本的质量和连贯性。Reynolds 等^[40]提出的提示优化技术, 通过探索和优化不同提示的结构和内容, 使得生成的文本在语义一致性和逻辑连贯性上表现更加优异。通过动态提示调整, GPT-3 能够生成风格多样且内容丰富的新闻报道和创意故事^[10]。在文本理解领域, Peng 等^[123]开发了一种基于提示的机器阅读理解架构, 用于临床概念和关系的提取。

该方法通过引入与语境相关的提示, 有效提升了问答系统的准确性和鲁棒性。这些研究表明, 提示工程在文本生成与理解中的应用, 提升了预训练语言模型的性能, 为新闻撰写、内容创作以及阅读理解等领域提供了新的思路。

3.1.2 机器翻译与对话系统

提示工程通过引入上下文提示, 显著提高了翻译质量和对话生成的自然性与连贯性。Vaswani 等^[124]提出的 Transformer 模型利用注意力机制处理长句和复杂结构, 在多种语言序列之间实现高质量翻译。提示在专业术语和多义词的翻译中发挥着关键作用, 帮助模型准确理解原文语境, 从而生成精确的译文^[125]。在对话系统方面, 提示工程通过指导模型关注对话上下文, 提升了对话生成的连贯性和逻辑性。Zhang 等^[126]开发的 DialoGPT 模型基于大规模对话数据进行预训练, 并结合提示引导, 使生成的对话回复更加自然和流畅。这些研究成果表明, 提示工程在提升机器翻译准确性和对话系统连贯性方面, 具有较大的应用潜力和实际价值。

3.1.3 情感分析与观点挖掘

在情感分析与观点挖掘领域, 提示工程通过精细设计的提示显著提高了模型对复杂文本的理解和情感预测能力。使用 ChatGPT 和其他大型语言模型进行情感分析时, 可以通过设计针对性的提示来增强模型的情感识别准确性和适应性。例如, 研究人员利用 ChatGPT 对金融新闻头条进行情感分类, 通过提示工程优化模型性能, 这种方法在零样本学习环境中卓有成效^[127]。Zhang 等^[128]提出了利用 seq2seq-attention 结构获取输入序列语义信息的自适应提示架构, 提高了提示质量。Gu 等^[129]引入了自动生成的 CoT 技术, 显著提高了模型在情感分析中的透明度和可解释性。

3.2 视觉与多模态任务

3.2.1 图像描述与生成

针对图像描述与生成 (text-to-image, T2I) 任务, 提示工程已被证实是提高模型处理图像任务能力的关键技术, 被广泛应用于图像的生成、分类^[130]、编辑以及标题生成等任务^[131-133]。通过对提示语的巧妙设计, 例如在图像生成中利用提示词以及前后变换的图像对照示例, 模型能够进行图像的复杂转换^[134-135]。此外, 结合文本和图像的多模态提示, 不仅能够优化模型对连续问答过程的上下文理解, 还能增强其描述能力^[136-137]。多模态思维链

等技术通过逐步指导提高了模型在处理图像问题的精确性和效率，显著提升了模型的可操作性能^[138-140]。Li等^[119]提出基于LLMs的场景生成模型SeGPT，用于生成复杂且具有挑战性的自动驾驶场景。

3.2.2 视频-音频理解与生成

提示技术已成功扩展至视频领域，支持文本到视频（text-to-video，T2V）、文本到3D（text-to-3D，T2-3D）和文本到音频（text-to-audio，T2A）的生成。在文本到视频生成方面，Brooks等^[141]和Liang等^[142]表明，多样化的文本提示可以驱动视频内容生成。在视频编辑任务中，Wu等^[143]和Cheng等^[144]提出如何利用提示来有效地调整和优化视频内容。同时，Yousaf等^[145]和Mi等^[146]的研究表明，视频到文本的转换可以借助精确的提示来提取视频关键信息并生成文本描述。此外，Kim等^[147]提出的IG-VLM模型通过将采样帧组合并连接到图像网格中来创建视频表示，再将图像网格与各种引导性提示一同输入VLMs，从而提高模型对长视频的理解能力。Zhang等^[148]提出的LLOVi模型采用一种语言引导视频理解的方法，首先使用字幕模型提取短期视频描述，然后通过LLMs总结这些描述并响应给定的提示。Wang等^[149]提出的VideoAgent模型采用了一种基于LLMs的多帧搜索策略，该策略使得模型能够根据输入提示及先前的帧标题进行自我反省，并预测可能的答案，从而有效地探索下一帧内容。在戏剧产业中，倪清桦等^[150]提出平行戏剧框架，通过构建戏剧大模型、搭建计算实验平台来模拟和优化戏剧制作流程，对剧本创作、舞美设计、演出安排及市场营销等环节实施智能化控制与自动化执行，能够在减少人为误差的同时增强艺术作品的表现力和观众互动性，实现人机协同创作与智能管理，从而提升戏剧创作与演出的效率和质量。Agostinelli等^[151]提出的MusicLM模型，根据自然语言描述生成音乐片段，并基于文本和旋律进行条件优化。

3.3 推理与决策

3.3.1 逻辑推理与数学问题

在逻辑推理与数学问题上，应用提示工程能够显著提升模型解决复杂问题的能力。通过精心设计的逻辑推理路径和数学公式提示，大语言模型（如GPT-4^[23]和PaLM^[104]）在多步推理任务中展现出高效性^[146-148,152-154]。Wang等^[155]提出了一种逻辑支架推理规则生成框架，构建了推理规则库ULogic，

从而增强模型的逻辑推理能力。Wan等^[156]提出了LogicAsker框架，其通过自动化方法全面评估和提高大型语言模型的逻辑推理能力。Deng等^[157]提出了集成大型语言模型与符号求解器的方法，以完成各种推理任务。Li等^[158]使用大型语言模型和知识图谱，实现了更可靠的基于知识的推理，有助于推理结果的追踪。Imani等^[159]提出了MathPrompter框架，利用CoT技术生成多个代数表达式或Python函数，以多种方式解决同一数学问题，从而提高结果的置信度。Wu等^[160]提出了一种名为渐进式纠正提示的方法，用于生成推理路径，从而解决数学应用题。Srivastava等^[161]提出了MathDivide提示技术，通过将数学问题分解为更简单的子问题来提高求解效率。Wu等^[162]探讨了如何定义提示格式并在格式中融入数学知识，以提高LLMs在翻译数学应用题时的效率和质量，还总结了翻译数学论文时涉及的数学知识并将其融入提示中，从而显著提升了模型翻译数学论文的效果。

3.3.2 决策规划支持系统

提示工程在增强模型在复杂任务规划与执行方面的能力上发挥着关键作用，主要表现为提升模型的规划能力、执行效率以及可持续性，提升智能体控制的安全性。

在医疗决策领域，通过整合实时患者数据和多模态数据分析，临床决策支持系统能够为医疗专业人员提供及时、具有指导性的见解，从而提高治疗决策的质量并提高患者的就诊体验^[163-165]。这些系统不仅提高了医疗安全和临床管理的效率，还能通过预测患者的状况和治疗结果来改进临床路径和决策。Nazary等^[166]从可解释的机器学习模型中获取领域知识，并将其无缝融入提示设计中，从而提取与特征重要性有关的关键见解，以辅助医务人员进行决策。

在自动驾驶领域，Wang等^[167]设计了基于LLMs的副驾驶Co-pilot框架，驾驶员通过自然语言描述任务，Co-pilot结合记忆模块与专家微调技术，综合考量生成轨迹的安全性、任务完成度、驾驶行为正确性，并完成相应的轨迹规划与控制任务。Wang等^[120]提出基于LLMs的人-车交互框架，通过结合车内驾驶员情绪、驾驶状态与当前驾驶环境、任务描述，完成分布式自动驾驶车轨迹规划与控制。

具体来说，向大模型提供关于任务目标和执行步骤的明确提示，可以显著提高其规划和执行特定操作的能力^[168]。这种方法在自动化生产、智能机

机器人等多个领域表现出强大的应用潜力^[169]。Singh等^[170]提出了一种程序化的LLMs提示结构,在考虑情境环境、机器人功能以及具体任务需求时,该结构有助于生成有效的提示。Chen等^[171]开发的NLMap,旨在解决将LLMs应用于现实世界的机器人任务所面临的环境限制。NLMap作为一个框架,能够收集和集成上下文信息至LLMs规划器中,在生成条件计划前能够查看和查询场景中的可用对象。此外,Kannan等^[172]开发了一个专门为多机器人任务规划设计的SMART-LLM,该框架融合了任务分解、联盟形成、任务分配等多个规划阶段,全程依赖于程序化的LLMs提示。这些提示基于少量样本学习范式,最优化任务规划的效率和准确性。Lin等^[173]面向无人机系统提出了基于提示微调的复杂任务分解策略,提高了无人机在复杂城市环境中的任务执行效率和准确度。Tian等^[174]将提示微调策略应用于物流系统,在无人机、无人船和无人车场景的3D终端交付中具有显著成效。

4 提示工程挑战与展望

生成式AI是一项新兴技术,研究者对大模型的突现能力和局限性的理解仍然有限。在自然语言理解方面,由于自然语言是一种灵活、开放的输入信息,大模型处理自然语言时面临许多语言交流挑战,例如有歧义、上下文的依赖性和需要实时修正等问题,同时大模型对自然语言的理解可能与人类理解无实质关系且存在交流障碍。在多模态数据理解方面,视觉、文本、音频等多种类型的提示的数据格式、尺度和结构各不相同,如何在同一个模型中有效整合和处理这些跨物理、信息和社会空间^[175-176]的异质数据,保持各模态信息之间的一致性和互操作性,是一个巨大的挑战。

提示工程通过提供特定的提示来引导模型,但在多模态输入文本生成中的应用仍然存在一定的局限性。例如,硬提示和软提示在多模态生成式模型中的潜力尚未被完全挖掘,特别是如何有效地设计和优化软提示,以便在图文匹配和其他多模态任务中提升模型性能,仍需进行深入研究。上下文学习和指令调优在多模态模型中已表现出高效性,但其内部工作机制仍缺乏可解释性。这些模型如何从不同的示例中学习,以及不同示例对模型性能的具体贡献,尚未得到充分研究。这些学习主要包括几个方面。首先是任务理解,即模型如何通过提示了解

和解析任务需求。其次是知识获取,即模型如何从提示中获取必要的知识,通过上下文提示激活模型中预先存储的相关知识。第三是模式识别,即模型如何从提示中识别出有用的模式和结构,从而在新的示例中应用这些模式。此外,还包括示例贡献,即不同示例在提示中对模型性能的具体贡献,理解哪些示例最有助于提升模型的任务表现。最后是指令执行,即模型如何执行提示中的指令,根据指令生成相应的输出,例如在图文匹配任务中生成合适的文本描述或在图像生成任务中生成符合描述的图像。深入理解这些机制有助于更好地设计和优化提示,从而提升模型在各种下游任务中的性能。

视觉提示可以在处理密集对象、幻觉等困难场景时发挥重要作用,尽管预训练的编码器通过匹配损失进行提示已被广泛用于下游任务,但对视觉提示的研究仍相对较少。此外,T2V、T2-3D生成模型依赖于T2I模型的性能,然而,T2I模型的输入不一致问题可能导致生成的视频和3D内容出现错误,从而影响模型的整体性能和可靠性。如何解决T2I模型的输入不一致性问题,以提高T2V和T2-3D模型的稳定性和准确性,也是亟待解决的挑战。此外,T2A模型生成的音乐仍存在杂音及不和谐音调问题,特别是输入图像会导致此类模型出现不一致性问题。

大模型通常在网络规模数据集上进行预训练,这些数据集可能保留了偏见知识或敏感隐私信息^[177]。如何在提示工程的部署过程中控制这些潜在风险,确保生成内容的公平性和透明性,是一个重要的研究方向。对抗鲁棒性在各种模型架构中被广泛研究,但在VLMs中,图像-文本匹配与理解仍有待进一步研究,特别是VLMs在对抗攻击下的鲁棒性以及提示大小对模型鲁棒性的影响,还有待进一步研究。

对齐技术虽然是解决以上问题的一种主要技术,但是内部对齐和外部对齐在保持大模型与人类价值观的一致性上仍面临挑战。外部对齐的主要挑战在于如何准确定义和量化复杂、多样且具有文化、社会差异的人类价值观,并将这些定性的价值观转化为模型的优化目标,同时避免大模型找到不符合设计者意图的内容。而内部对齐则面临模型行为不可预测和目标泛化问题,确保模型在实际优化过程不背离设计者设定的目标,即使面对未见过的情况也能表现出符合预期的行为。这些问题的解决

需要综合考虑内部对齐和外部对齐的策略，从而确保模型的可靠性和安全性。

未来研究可以探索结合音频、热成像等更多模态提示的方法，创建能够处理多种模态输入的统一模型，并结合有效的优化策略进一步提升多模态模型的能力，如微软亚洲研究院的 Kosmos^[178]和 Meta AI 的 IMAGEBIND^[179-180]。此外，未来可以探索在多模态生成模型中应用软提示，例如 CoOp (context optimization) 等上下文优化模型^[16]，以及进一步研究上下文学习和指令调优的内部机制。

探索提示工程在大模型预训练编码器中的应用也将是一个重要的研究方向，尤其是在处理密集对象、减少幻觉问题以及提升现代生成式AI的适应性等方面的应用。在 T2I 任务中，结合视觉提示和视觉注释可以提供更多的视觉线索，从而生成个性化和更准确的图像。同样，研究如何在多模态输入-输出的大模型中使用统一的提示方法，进而提升其视觉信息（如图像和视频）和文本信息处理能力，能够实现视觉定位（在图像中找到特定位置或物体）和视觉-文本共引用对话（结合视觉和文本进行互动和交流）。这一研究有望在 T2V、T2-3D、T2A 任务中展现出更大的潜力。此外，加强提示工程中的AI生成内容的伦理性研究，确保模型生成的内容不包含偏见，保护用户隐私，避免不道德的内容生成，也是未来的重要课题。未来还可以继续研究如何提高大模型在对抗攻击下的鲁棒性，以确保模型在各种复杂应用中的稳定性和可靠性，如自动驾驶、艺术、区块链、医疗等领域^[181-194]。

5 结论

尽管 LLMs 和 VLMs 的突现能力使其在生成和理解任务方面表现出色，但是它们在处理特定任务时仍面临一定挑战，仍需要在大量标注数据上进行微调，且模型的表现可能会受到数据质量和任务特异性的限制，导致直接修改模型参数的成本高昂且收效甚微。此外，幻觉、知识迁移、一致性缺失、有害信息等问题，导致大模型在处理实际问题时表现不佳。提示工程不需要调整模型参数，直接通过多模态提示引导大模型执行任务，成为了解决这些问题的一种有效手段。结合提示工程与对齐技术可以进一步确保模型输出符合人类的价值观和预期。本文深入讨论提示工程与对齐技术，提出基于提示优化、专家反馈机制以及实时调整机制的引导概

念，首先通过提示优化帮助大模型更好地理解任务需求，专家反馈机制通过持续的专业评估和反馈进一步验证和改进提示与对齐过程，以确保输出的准确性和一致性，实时调整机制根据输出动态优化模型行为。引导概念为研究者构建、评估、执行提示工程与对齐提供指导，进而提升了大模型在多种任务上的性能。

参考文献：

- [1] NADKARNI P M, OHNO-MACHADO L, CHAPMAN W W. Natural language processing: an introduction[J]. Journal of the American Medical Informatics Association, 2011, 18(5): 544-551.
- [2] GOLDBERG Y. A primer on neural network models for natural language processing[J]. Journal of Artificial Intelligence Research, 2016, 57: 345-420.
- [3] ZHAI C X. Statistical language models for information retrieval a critical review[J]. Foundations and Trends in Information Retrieval, 2007, 2(3): 137-213.
- [4] KIROS R, SALAKHUTDINOV R, ZEMEL R. Multimodal neural language models[C]//Proceedings of the 31st International Conference on International Conference on Machine Learning. New York: ACM, 2014: 595-603.
- [5] KAPLAN J, MCCANDLISH S, HENIGHAN T, et al. Scaling laws for neural language models[EB]. arXiv preprint, 2020, arXiv: 2001.08361.
- [6] CHURCH K W. Word2vec[J]. Natural Language Engineering, 2017, 23(1): 155-162.
- [7] PENNINGTON J, SOCHER R, MANNING C. Glove: global vectors for word representation[C]//Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing(EMNLP). Stroudsburg: Association for Computational Linguistics, 2014: 1532-1543.
- [8] LIU X, ZHENG Y N, DU Z X, et al. GPT understands, too[EB]. arXiv preprint, 2023, arXiv: 2103.10385.
- [9] DEVLIN J, CHANG M W, LEE K, et al. BERT: pre-training of deep bidirectional transformers for language understanding[EB]. arXiv preprint, 2018, arXiv: 1810.04805.
- [10] BROWN T B, MANN B, RYDER N, et al. Language models are few-shot learners[EB]. arXiv preprint, 2020, arXiv: 2005.14165.
- [11] TOUVRON H, MARTIN L, STONE K, et al. Llama 2: open foundation and fine-tuned chat models[EB]. arXiv preprint, 2023, arXiv: 2307.09288.
- [12] WEI J, TAY Y, BOMMASANI R, et al. Emergent abilities of large language models[EB]. arXiv preprint, 2022, arXiv: 2206.07682.
- [13] SCHAEFFER R, MIRANDA B, KOYEJO O. Are emergent abilities of large language models a mirage? [C]//Proceedings of the 37th International Conference on Neural Information Processing Systems. New York: Curran Associates Inc., 2023: 55565- 55581.
- [14] YU H K, LIU X Y, TIAN Y L, et al. Sora-based parallel vision for smart sensing of intelligent vehicles: from foundation models to foundation intelligence[J]. IEEE Transactions on Intelligent Vehicles, 2024, 9(2): 3123-3126.
- [15] ZHOU K Y, YANG J K, LOY C C, et al. Conditional prompt learning for vision-language models[C]//Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2022: 16795-16804.
- [16] ZHOU K Y, YANG J K, LOY C C, et al. Learning to prompt for vision-

- language models[J]. *International Journal of Computer Vision*, 2022, 130(9): 2337-2348.
- [17] ZHANG J Y, HUANG J X, JIN S, et al. Vision-language models for vision tasks: a survey[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024, 46(8): 5625-5644.
- [18] ALAYRAC J B, DONAHUE J, LUC P, et al. Flamingo: a visual language model for few-shot learning[C]//*Proceedings of 2022 Advances In Neural Information Processing Systems*. New York: Curran Associates, Inc., 2022: 23716-23736.
- [19] SHTEDRITSKI A, RUPPRECHT C, VEDALDI A. What does CLIP know about a red circle? Visual prompt engineering for VLMs[C]//*Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision (ICCV)*. Piscataway: IEEE Press, 2023: 11987-11997.
- [20] RAMESH A, PAVLOV M, GOH G, et al. Zero-shot text-to-image generation[C]//*Proceedings of the 38th International Conference on Machine Learning*. New York: PMLR, 2021: 8821-8831.
- [21] RAMESH A, DHARIWAL P, NICHOL A, et al. Hierarchical text-conditional image generation with CLIP latents[EB]. *arXiv preprint*, 2022, arXiv: 2204.06125.
- [22] JI Z W, LEE N, FRIESKE R, et al. Survey of hallucination in natural language generation[J]. *ACM Computing Surveys*, 2023, 55(12): 1-38.
- [23] OPENAI, ACHIAM J, ADLER S, et al. GPT-4 technical report[EB]. *arXiv preprint*, 2023, arXiv: 2303.08774.
- [24] LIU P F, YUAN W Z, FU J L, et al. Pre-train, prompt, and predict: a systematic survey of prompting methods in natural language processing[J]. *ACM Computing Surveys*, 2023, 55(9): 1-35.
- [25] SHEN T H, JIN R R, HUANG Y F, et al. Large language model alignment: a survey[EB]. *arXiv preprint*, 2023, arXiv: 2309.15025.
- [26] WANG Y F, ZHONG W J, LI L Y, et al. Aligning large language models with human: a survey[EB]. *arXiv preprint*, 2023, arXiv: 2307.12966.
- [27] WHITE J, FU Q C, HAYS S, et al. A prompt pattern catalog to enhance prompt engineering with ChatGPT[EB]. *arXiv preprint*, 2023, arXiv: 2302.11382.
- [28] ZHOU Y C, MURESANU A, HAN Z W, et al. Large language models are human-level prompt engineers[EB]. *arXiv preprint*, 2022, arXiv: 2211.01910.
- [29] GU J D, HAN Z, CHEN S, et al. A systematic survey of prompt engineering on vision-language foundation models[EB]. *arXiv preprint*, 2023, arXiv: 2307.12980.
- [30] SCHULHOFF S, ILIE M, BALEPUR N, et al. The prompt report: a systematic survey of prompting techniques[EB]. *arXiv preprint*, 2024, arXiv: 2406.06608.
- [31] DONG Q X, LI L, DAI D M, et al. A survey on In-context learning[EB]. *arXiv preprint*, 2022, arXiv: 2301.00234.
- [32] LESTER B, AL-ROUFI R, CONSTANT N. The power of scale for parameter-efficient prompt tuning[EB]. *arXiv preprint*, 2021, arXiv: 2104.08691.
- [33] YU F Y, PAN K J. The effects of student question-generation with online prompts on learning[J]. *Journal of Educational Technology & Society*, 2014, 17(3): 267 - 279.
- [34] SUN H. Offline prompt evaluation and optimization with inverse reinforcement learning[EB]. *arXiv preprint*, 2023, arXiv: 2309.06553.
- [35] ZIEGLER D M, STIENNON N, WU J, et al. Fine-tuning language models from human preferences[EB]. *arXiv preprint*, 2019, arXiv: 1909.08593.
- [36] GIRAY L. Prompt engineering with ChatGPT: a guide for academic writers[J]. *Annals of Biomedical Engineering*, 2023, 51(12): 2629-2633.
- [37] ZHANG Y H, ZHOU K Y, LIU Z W. What makes good examples for visual in-context learning? [C]//*Proceedings of the 37th International Conference on Neural Information Processing Systems*. New York: Curran Associates Inc., 2023: 17773-17794.
- [38] FENG G H, ZHANG B H, GU Y T, et al. Towards revealing the mystery behind chain of thought: a theoretical perspective[C]//*Proceedings of the 37th International Conference on Neural Information Processing Systems*. New York: Curran Associates Inc., 2023: 70757-70798.
- [39] TIAN Y L, LI X, ZHANG H, et al. VistaGPT: generative parallel transformers for vehicles with intelligent systems for transport automation[J]. *IEEE Transactions on Intelligent Vehicles*, 2023, 8(9): 4198-4207.
- [40] REYNOLDS L, MCDONELL K. Prompt programming for large language models: beyond the few-shot paradigm[C]//*Proceedings of the Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*. New York: ACM, 2021: 1-7.
- [41] SHIN T, RAZEGHI Y, LOGAN IV R L, et al. AutoPrompt: eliciting knowledge from language models with automatically generated prompts[EB]. *arXiv preprint*, 2020, arXiv: 2010.15980.
- [42] ZHANG Z S, ZHANG A, LI M, et al. Automatic chain of thought prompting in large language models[EB]. *arXiv preprint*, 2022, arXiv: 2210.03493.
- [43] PRASAD A, HASE P, ZHOU X, et al. GrIPS: gradient-free, edit-based instruction search for prompting large language models[EB]. *arXiv preprint*, 2022, arXiv: 2203.07281.
- [44] PRYZANT R, ITER D, LI J, et al. Automatic prompt optimization with “gradient descent” and beam search[EB]. *arXiv preprint*, 2023, arXiv: 2305.03495.
- [45] LI C, LIU X M, WANG Y C, et al. Dialogue for prompting: a policy-gradient-based discrete prompt generation for few-shot learning[EB]. *arXiv preprint*, 2023, arXiv: 2308.07272.
- [46] SHAH D, SCHWARTZ H A, HOVY D. Predictive biases in natural language processing models: a conceptual framework and overview[EB]. *arXiv preprint*, 2019, arXiv: 1912.11078.
- [47] GEHMAN S, GURURANGAN S, SAP M, et al. RealToxicityPrompts: evaluating neural toxic degeneration in language models[EB]. *arXiv preprint*, 2020, arXiv: 2009.11462.
- [48] DALE R. GPT-3: what’s it good for?[J]. *Natural Language Engineering*, 2021, 27(1): 113-118.
- [49] AKYÜREK E, BOLUKBASI T, LIU F, et al. Towards tracing factual knowledge in language models back to the training data[EB]. *arXiv preprint*, 2022, arXiv: 2205.11482.
- [50] POESIA G, POLOZOV O, LE V, et al. Synchromesh: reliable code generation from pre-trained language models[EB]. *arXiv preprint*, 2022, arXiv: 2201.11227.
- [51] LIU H T, LI C Y, WU Q Y, et al. Visual instruction tuning[C]//*Proceedings of the 37th International Conference on Neural Information Processing Systems*. New York: Curran Associates Inc., 2023: 34892-34916.
- [52] LI J N, LI D X, SILVIO S, et al. Blip-2: bootstrapping language-image pre-training with frozen image encoders and large language models[C]//*Proceedings of the 40th International Conference on Machine Learning*. New York: PMLR, 2023: 19730-19742.
- [53] WIENER N. Some moral and technical consequences of automation: as machines learn they may develop unforeseen strategies at rates that baffle their programmers[J]. *Science*, 1960, 131(3410): 1355-1358.
- [54] RUSSELL S J, NORVIG P. *Artificial intelligence: a modern approach*[M]. New Jersey: Prentice Hall, 2016.
- [55] OUYANG L, WU JEFF, XU J, et al. Training language models to fol-

- low instructions with human feedback[C]//Proceedings of the 36th International Conference on Neural Information Processing Systems. New York: Curran Associates Inc., 2022: 27730-27744.
- [56] LIU R B, JIA C Y, ZHANG G, et al. Second thoughts are best: learning to re-align with human values from text edits[C]//Proceedings of 2022 Advances In Neural Information Processing Systems. New York: Curran Associates, Inc., 2022: 181-196.
- [57] LIU H, SFERRAZZA C, ABBEEL P. Chain of hindsight aligns language models with feedback[EB]. arXiv preprint, 2023, arXiv: 2302.02676.
- [58] STIENNON N, OUYANG L, WU J, et al., Learning to summarize with human feedback[C]//Proceedings of 2020 Advances In Neural Information Processing Systems. New York: Curran Associates, Inc., 2020: 3008-3021.
- [59] SUN Z Q, SHEN Y K, ZHOU Q H, et al. Principle-driven self-alignment of language models from scratch with minimal human supervision[C]//Proceedings of the 37th International Conference on Neural Information Processing Systems. New York: Curran Associates, Inc., 2023: 2511-2565.
- [60] DU Y L, LI S, TORRALBA A, et al. Improving factuality and reasoning in language models through multiagent debate[EB]. arXiv preprint, 2023, arXiv: 2305.14325.
- [61] KHAN A, HUGHES J, VALENTINE D, et al. Debating with more persuasive LLMs leads to more truthful answers[EB]. arXiv preprint, 2024, arXiv: 2402.06782.
- [62] ELHAGE N, NANDA N, OLSSON C, et al. A mathematical framework for transformer circuits[J]. Transformer Circuits Thread, 2021, 1(1): 12.
- [63] VILONE G, LONGO L. Explainable artificial intelligence: a systematic review[EB]. arXiv preprint, 2020, arXiv: 2006.00093.
- [64] WALLACE E, FENG S, KANDPAL N, et al. Universal adversarial triggers for attacking and analyzing NLP[EB]. arXiv preprint, 2019, arXiv: 1908.07125.
- [65] AKHTAR N, MIAN A. Threat of adversarial attacks on deep learning in computer vision: a survey[J]. IEEE Access, 2018, 6: 14410-14430.
- [66] ZHU B H, JORDAN M, JIAO J T. Principled reinforcement learning with human feedback from pairwise or k-wise comparisons[C]//Proceedings of the 40th International Conference on Machine Learning. New York: PMLR, 2023: 43037-43067.
- [67] LIPTON Z C. The myths of model interpretability: in machine learning, the concept of interpretability is both important and slippery[J]. Queue, 2018, 16(3): 31-57.
- [68] CRITCH A, KRUEGER D. AI research considerations for human existential safety (ARCHES)[EB]. arXiv preprint, 2020, arXiv: 2006.04948.
- [69] CASPER S, DAVIES X, SHI C, et al. Open problems and fundamental limitations of reinforcement learning from human feedback[EB]. arXiv preprint, 2023, arXiv: 2307.15217.
- [70] ZOU A, WANG Z F, CARLINI N, et al. Universal and transferable adversarial attacks on aligned language models[EB]. arXiv preprint, 2023, arXiv: 2307.15043.
- [71] SUZGUN M, MELAS-KYRIAZI L, JURAFSKY D. Prompt-and-rerank: a method for zero-shot and few-shot arbitrary textual style transfer with small language models[EB]. arXiv preprint, 2022, arXiv: 2205.11503.
- [72] WANG Z M, PENG Z Y, QUE H R, et al. RoleLLM: benchmarking, eliciting, and enhancing role-playing abilities of large language models[EB]. arXiv preprint, 2023, arXiv: 2310.00746.
- [73] JIANG H, ZHANG X J, CAO X B, et al. PersonaLLM: investigating the ability of large language models to express personality traits[EB]. arXiv preprint, 2023, arXiv: 2305.02547.
- [74] LU A, ZHANG H X, ZHANG Y Z, et al. Bounding the capabilities of large language models in open text generation with prompt constraints[EB]. arXiv preprint, 2023, arXiv: 2302.09185.
- [75] LI C, WANG J D, ZHANG Y X, et al. Large language models understand and can be enhanced by emotional stimuli[EB]. arXiv preprint, 2023, arXiv: 2307.11760.
- [76] HUANG J T, LAM M H, LI E J, et al. Emotionally numb or empathetic? evaluating how LLMs feel using EmotionBench[EB]. arXiv preprint, 2023, arXiv: 2308.03656.
- [77] WILF A, LEE S S, LIANG P P, et al. Think twice: perspective-taking improves large language models' theory-of-mind capabilities[EB]. arXiv preprint, 2023, arXiv: 2311.10227.
- [78] WESTON J, SUKHBATAR S. System 2 attention (is something you might need too)[EB]. arXiv preprint, 2023, arXiv: 2311.11829.
- [79] DENG Y H, ZHANG W T, CHEN Z X, et al. Rephrase and respond: let large language models ask better questions for themselves[EB]. arXiv preprint, 2023, arXiv: 2311.04205.
- [80] XU X H, TAO C Y, SHEN T, et al. Re-reading improves reasoning in large language models[EB]. arXiv preprint, 2023, arXiv: 2309.06275.
- [81] PRESS O, ZHANG M R, MIN S, et al. Measuring and narrowing the compositionality gap in language models[EB]. arXiv preprint, 2022, arXiv: 2210.03350.
- [82] ZHAO T, WALLACE E, FENG S, et al. Calibrate before use: improving few-shot performance of language models[C]//Proceedings of the 38th International Conference on Machine Learning. New York: PMLR, 2021: 12697-12706.
- [83] LEE D H, KADAKIA A, TAN K M, et al. Good examples make a faster learner: simple demonstration-based learning for low-resource NER[EB]. arXiv preprint, 2021, arXiv: 2110.08454.
- [84] EISENSTEIN J, ANDOR D, BOHNET B, et al. Honest students from untrusted teachers: learning an interpretable question-answering pipeline from a pretrained language model[EB]. arXiv preprint, 2022, arXiv: 2210.02498.
- [85] ZHANG H X, ZHANG Y Z, ZHANG R Y, et al. Robustness of demonstration-based learning under limited data scenario[EB]. arXiv preprint, 2022, arXiv: 2210.10693.
- [86] LI S Y, CHEN J S, SHEN Y L, et al. Explanations from large language models make small reasoners better[EB]. arXiv preprint, 2022, arXiv: 2210.06726.
- [87] DAI Z Y, ZHAO V Y, MA J, et al. Promptagator: few-shot dense retrieval from 8 examples[EB]. arXiv preprint, 2022, arXiv: 2209.11755.
- [88] YU W H, ITER D, WANG S H, et al. Generate rather than retrieve: large language models are strong context generators[EB]. arXiv preprint, 2022, arXiv: 2209.10063.
- [89] WEI J, WANG X Z, SCHUURMANS D, et al. Chain-of-thought prompting elicits reasoning in large language models[C]//Proceedings of the 36th International Conference on Neural Information Processing Systems. New York: Curran Associates Inc., 2022: 24824-24837.
- [90] LAMPINEN A K, DASGUPTA I, CHAN S C Y, et al. Can language models learn from explanations in context?[EB]. arXiv preprint, 2022, arXiv: 2204.02329.
- [91] ZHOU D, SCHÄRLI N, HOU L, et al. Least-to-most prompting enables complex reasoning in large language models[EB]. arXiv preprint, 2022, arXiv: 2205.10625.
- [92] LI F F, FERGUS R, PERONA P. One-shot learning of object categories[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2006, 28(4): 594-611.

- [93] LIU J C, SHEN D H, ZHANG Y Z, et al. What makes good In-context examples for GPT-3? [EB]. arXiv preprint, 2021, arXiv: 2101.06804.
- [94] LU Y, BARTOLO M, MOORE A, et al. Fantastically ordered prompts and where to find them: overcoming few-shot prompt order sensitivity[EB]. arXiv preprint, 2021, arXiv: 2104.08786.
- [95] MIN S, LYU X X, HOLTZMAN A, et al. Rethinking the role of demonstrations: what makes in-context learning work?[C]//Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. Stroudsburg: Association for Computational Linguistics, 2022: 11048-11064.
- [96] YOO K M, KIM J, KIM H J, et al. Ground-truth labels matter: a deeper look into input-label demonstrations[EB]. arXiv preprint, 2022, arXiv: 2205.12685.
- [97] JIANG Z B, XU F F, ARAKI J, et al. How can we know what language models know?[J]. Transactions of the Association for Computational Linguistics, 2020, 8: 423-438.
- [98] MIN S, LYU X X, HOLTZMAN A, et al. Rethinking the role of demonstrations: what makes In-context learning work? [EB]. arXiv preprint, 2022, arXiv: 2202.12837.
- [99] SU H J, KASAI J, WU C H, et al. Selective annotation makes language models better few-shot learners[EB]. arXiv preprint, 2022, arXiv: 2209.01975.
- [100] KIM H J, CHO H, KIM J, et al. Self-generated in-context learning: leveraging auto-regressive language models as a demonstration generator[EB]. arXiv preprint, 2022, arXiv: 2206.08082.
- [101] LI X N, QIU X P. Finding support examples for in-context learning[EB]. arXiv preprint, 2023, arXiv: 2302.13539.
- [102] LI X N, LV K, YAN H, et al. Unified demonstration retriever for in-context learning[EB]. arXiv preprint, 2023, arXiv: 2305.04320.
- [103] ZHANG Y M, FENG S, TAN C H. Active example selection for in-context learning[EB]. arXiv preprint, 2022, arXiv: 2211.04486.
- [104] CHOWDHERY A, NARANG S, DEVLIN J, et al. Palm: scaling language modeling with pathways[J]. Journal of Machine Learning Research, 2023, 24(240): 1-113.
- [105] COBBE K, KOSARAJU V, BAVARIAN M, et al. Training verifiers to solve math word problems[EB]. arXiv preprint, 2021, arXiv: 2110.14168.
- [106] WANG X Z, WEI J, SCHUURMANS D, et al. Self-consistency improves chain of thought reasoning in language models[EB]. arXiv preprint, 2022, arXiv: 2203.11171.
- [107] PATEL A, BHATTAMISHRA S, GOYAL N. Are NLP models really able to solve simple math word problems? [EB]. arXiv preprint, 2021, arXiv: 2103.07191.
- [108] LING W, YOGATAMA D, DYER C, et al. Program induction by rationale generation: learning to solve and explain algebraic word problems[EB]. arXiv preprint, 2017, arXiv: 1705.04146.
- [109] GEVA M, GUPTA A, BERANT J. Injecting numerical reasoning skills into language models[EB]. arXiv preprint, 2020, arXiv: 2004.04487.
- [110] YAO S Y, YU D, ZHAO J, et al. Tree of thoughts: deliberate problem solving with large language models[C]//Proceedings of the 37th International Conference on Neural Information Processing Systems. New York: Curran Associates Inc., 2023: 11809-11822.
- [111] YE Q Y, AXMED M, PRYZANT R, et al. Prompt engineering a prompt engineer[EB]. arXiv preprint, 2023, arXiv: 2311.05661.
- [112] GAO Y F, XIONG Y, GAO X Y, et al. Retrieval-augmented generation for large language models: a survey[EB]. arXiv preprint, 2023, arXiv: 2312.10997.
- [113] CHEN W H, HU H X, SAHARIA C, et al. Re-imagen: retrieval-augmented text-to-image generator[EB]. arXiv preprint, 2022, arXiv: 2209.14491.
- [114] HUANG R J, HUANG J W, YANG D C, et al. Make-an-audio: text-to-audio generation with prompt-enhanced diffusion models[C]//Proceedings of the 40th International Conference on Machine Learning. New York: PMLR, 2023: 13916-13932.
- [115] ZHU Y H, REN C Y, XIE S Y, et al. REALM: rag-driven enhancement of multimodal electronic health records analysis via large language models[EB]. arXiv preprint, 2024, arXiv: 2402.07016.
- [116] RYU C, LEE S, PANG S, et al. Retrieval-based evaluation for LLMs: a case study in Korean legal QA[C]//Proceedings of the Natural Language Processing Workshop 2023. Stroudsburg: Association for Computational Linguistics, 2023: 132-137.
- [117] MA A, A L. A rag chatbot for precision medicine of multiple myeloma[EB]. medRxiv preprint, 2024, medRxiv: 24304293.
- [118] DAI X Y, GUO C, TANG Y, et al. VistaRAG: toward safe and trustworthy autonomous driving through retrieval-augmented generation[J]. IEEE Transactions on Intelligent Vehicles, 2024, 9(4): 4579-4582.
- [119] LI X, LIU E L, SHEN T Y, et al. ChatGPT-based scenario engineer: a new framework on scenario generation for trajectory prediction[J]. IEEE Transactions on Intelligent Vehicles, 2024, 9(3): 4422-4431.
- [120] WANG X, HUANG J, TIAN Y L, et al. Parallel driving with big models and foundation intelligence in cyber-physical-social spaces[J]. Research, 2024, 7(3): 1-16.
- [121] WANG X, HUANG J, TIAN Y L, et al. AGI in metaverse for smart cities and societies: a cyber physical social approach[C]//Proceedings of the 2024 Australian & New Zealand Control Conference (ANZCC). Piscataway: IEEE Press, 2024: 61-66.
- [122] 黄峻, 田永林, 戴星原, 等. 基于深度学习的自动驾驶多模态轨迹预测方法:现状及展望[J]. 智能科学与技术学报, 2023, 5(2): 180-199.
- HUANG J, TIAN Y L, DAI X Y, et al. Deep learning-based multimodal trajectory prediction methods for autonomous driving: state of the art and perspectives[J]. Chinese Journal of Intelligent Science and Technology, 2023, 5(2): 180-199.
- [123] PENG C, YANG X, YU Z H, et al. Clinical concept and relation extraction using prompt-based machine reading comprehension[J]. Journal of the American Medical Informatics Association, 2023, 30(9): 1486-1493.
- [124] VASWANI A, SHAZEER N, NIKI P, et al. Attention is all you need[C]//Proceedings of the 31st International Conference on Neural Information Processing Systems. New York: Curran Associates, Inc., 2017: 6000-6010.
- [125] LI Y F, YIN Y J, LI J, et al. Prompt-driven neural machine translation[C]//Proceedings of the Findings of the Association for Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 2022: 2579-2590.
- [126] ZHANG Y Z, SUN S Q, GALLEY M, et al. DialoGPT: large-scale generative pre-training for conversational response generation[EB]. arXiv preprint, 2019, arXiv: 1911.00536.
- [127] FATOUROS G, SOLDATOS J, KOUROUMALI K, et al. Transforming sentiment analysis in the financial domain with ChatGPT[J]. Machine Learning with Applications, 2023, 14: 100508.
- [128] ZHANG P F, CHAI T T, XU Y D. Adaptive prompt learning-based few-shot sentiment analysis[J]. Neural Processing Letters, 2023, 55(6): 7259-7272.
- [129] GU X, CHEN X L, LU P, et al. AGCVT-prompt for sentiment classification: automatically generating chain of thought and verbalizer in prompt learning[J]. Engineering Applications of Artificial Intelligence,

- 2024, 132: 107907.
- [130] WANG Y T, WANG J G, CAO Y S, et al. Integrated inspection on PCB manufacturing in cyber-physical-social systems[J]. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 2023, 53(4): 2098-2106.
- [131] DING M, YANG Z Y, HONG W Y, et al. Cogview: mastering text-to-image generation via transformers[C]//*Proceedings of the 35th Conference on Neural Information Processing Systems*. New York: Curran Associates, Inc., 2021: 19822-19835.
- [132] HINZ T, HEINRICH S, WERMTER S. Semantic object accuracy for generative text-to-image synthesis[J]. *IEEE Trans Pattern Anal Mach Intell*, 2022, 44(3): 1552-1565.
- [133] TAO M, TANG H, WU F, et al. DF-GAN: a simple and effective baseline for text-to-image synthesis[C]//*Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR)*. Piscataway: IEEE Press, 2022: 16494-16504.
- [134] LI B W, QI X J, THOMAS L, et al. Controllable text-to-image generation[C]//*Proceedings of the 33rd International Conference on Neural Information Processing Systems*. New York: Curran Associates Inc., 2019: 2665-2675.
- [135] ROMBACH R, BLATTMANN A, LORENZ D, et al. High-resolution image synthesis with latent diffusion models[C]//*Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Piscataway: IEEE Press, 2022: 10674-10685.
- [136] LI X J, YIN X, LI C Y, et al. Oscar: object-semantics aligned pre-training for vision-language tasks[C]//*Proceedings of the 2020 European Conference on Computer Vision*. Cham: Springer, 2020: 121-137.
- [137] KHALIL M, KHALIL A, NGOM A. A comprehensive study of vision transformers in image classification tasks[EB]. arXiv preprint, 2023, arXiv: 2312.01232.
- [138] CROWSON K, BIDERMAN S, KORNIS D, et al. VQGAN-CLIP: open domain image generation and editing with natural language guidance[C]//*Proceedings of 2022 European Conference on Computer Vision*. Cham: Springer, 2022: 88-105.
- [139] KWON G, YE J C. CLIPstyler: image style transfer with a single text condition[C]//*Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Piscataway: IEEE Press, 2022: 18041-18050.
- [140] BAR-TAL O, OFRI-AMAR D, FRIDMAN R, et al. Text2LIVE: text-driven layered image and video editing[C]//*Proceedings of 2022 European Conference on Computer Vision*. Cham: Springer, 2022: 707-723.
- [141] LI C X, LIU H Y, LIU Y F, et al. Endora: video generation models as endoscopy simulators[EB]. arXiv preprint, 2024, arXiv: 2403.11050.
- [142] LIANG J Y, FAN Y C, ZHANG K, et al. MoVideo: motion-aware video generation with diffusion models[EB]. arXiv preprint, 2023, arXiv: 2311.11325.
- [143] WU J Z, GE Y X, WANG X T, et al. Tune-A-video: one-shot tuning of image diffusion models for text-to-video generation[C]//*Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision (ICCV)*. Piscataway: IEEE Press, 2023: 7589-7599.
- [144] CHENG J X, XIAO T J, HE T. Consistent video-to-video transfer using synthetic dataset[EB]. arXiv preprint, 2023, arXiv: 2311.00213.
- [145] YOUSAF A, NASEER M, KHAN S, et al. Videoprompter: an ensemble of foundational models for zero-shot video understanding[EB]. arXiv preprint, 2023, arXiv: 2310.15324.
- [146] MI Y C, LI Y, SHU Y, et al. CLiF-VQA: enhancing video quality assessment by incorporating high-level semantic information related to human feelings[EB]. arXiv preprint, 2023, arXiv: 2311.07090.
- [147] KIM W, CHOI C, LEE W, et al. An image grid can be worth a video: zero-shot video question answering using a VLM[EB]. arXiv preprint, 2024, arXiv: 2403.18406.
- [148] ZHANG C, LU T X, ISLAM M M, et al. A simple LLM framework for long-range video question-answering[EB]. arXiv preprint, 2023, arXiv: 2312.17235.
- [149] WANG X H, ZHANG Y H, ZOHAR O, et al. VideoAgent: long-form video understanding with large language model as agent[EB]. arXiv preprint, 2024, arXiv: 2403.10517.
- [150] 倪清桦, 郭超, 王飞跃. 平行戏剧: 新时代戏剧的人机协同创作与智能管理[J]. *智能科学与技术学报*, 2023, 5(4): 436-445.
- NI Q H, GUO C, WANG F Y. Parallel theaters: human-machine collaborative creation and intelligent management for theatrical art[J]. *Chinese Journal of Intelligent Science and Technology*, 2023, 5(4): 436-445.
- [151] AGOSTINELLI A, DENK T I, BORSOS Z, et al. MusicLM: generating music from text[EB]. arXiv preprint, 2023, arXiv: 2301.11325.
- [152] AHN J, VERMA R, LOU R Z, et al. Large language models for mathematical reasoning: progresses and challenges[EB]. arXiv preprint, 2024, arXiv: 2402.00157.
- [153] SAHOO P, SINGH A K, SAHA S, et al. A systematic survey of prompt engineering in large language models: techniques and applications[EB]. arXiv preprint, 2024, arXiv: 2402.07927.
- [154] WANG X Y, AMAYUELAS A, ZHANG K X, et al. Understanding reasoning ability of language models from the perspective of reasoning paths aggregation[EB]. arXiv preprint, 2024, arXiv: 2402.03268.
- [155] WANG S Y, WEI Z Y, CHOI Y, et al. Can LLMs reason with rules? logic scaffolding for stress-testing and improving LLMs[EB]. arXiv preprint, 2024, arXiv: 2402.11442.
- [156] WAN Y X, WANG W X, YANG Y L, et al. A & B == B & A: triggering logical reasoning failures in large language models[EB]. arXiv preprint, 2024, arXiv: 2401.00757.
- [157] DENG S J, DONG H H, SI X J. Enhancing and evaluating logical reasoning abilities of large language models[C]//*Proceedings of ICLR 2024 Workshop on Secure and Trustworthy Large Language Models*. [S.l.:s.n.], 2024.
- [158] LI Y H, ZHANG R, LIU J Y. An enhanced prompt-based LLM reasoning scheme via knowledge graph-integrated collaboration[EB]. arXiv preprint, 2024, arXiv: 2402.04978.
- [159] IMANI S, DU L, SHRIVASTAVA H. MathPrompter: mathematical reasoning using large language models[EB]. arXiv preprint, 2023, arXiv: 2303.05398.
- [160] WU Z Y, JIANG M, SHEN C. Get an A in math: progressive rectification prompting[C]//*Proceedings of the 28th AAAI Conference on Artificial Intelligence*. California: AAAI, 2024: 19288-19296.
- [161] SRIVASTAVA S S, GANDHI A. MathDivide: improved mathematical reasoning by large language models[EB]. arXiv preprint, 2024, arXiv: 2405.13004.
- [162] WU H, YU X G. Prompt incorporates math knowledge to improve efficiency and quality of large language models to translate math word problems[C]//*Proceedings of the 2023 International Conference on Intelligent Education and Intelligent Research (IEIR)*. Piscataway: IEEE Press, 2023: 1-5.
- [163] SUTTON R T, PINCOCK D, BAUMGART D C, et al. An overview of clinical decision support systems: benefits, risks, and strategies for success[J]. *NPJ Digital Medicine*, 2020, 3: 17.
- [164] 林飞, 王飞跃, 田永林, 等. 平行药物系统: 基于大语言模型和三类

- 人的框架与方法[J]. 智能科学与技术学报, 2024, 6(1): 88-99.
- LIN F, WANG F Y, TIAN Y L, et al. Parallel drug systems: framework and methods based on large language models and three types of humans[J]. Chinese Journal of Intelligent Science and Technology, 2024, 6(1): 88-99.
- [165] LIN F, GAO T, SUN D, et al. Parallel medical devices and instruments: integrating edge and cloud intelligence for smart treatment and health systems[J]. IEEE/CAA Journal of Automatica Sinica, 2024, accepted.
- [166] NAZARY F, DELDJOO Y, DI NOIA T. ChatGPT-HealthPrompt: harnessing the power of XAI in prompt-based healthcare decision support using ChatGPT[C]//Proceedings of 2023 European Conference on Artificial Intelligence International Workshops. Cham: Springer, 2024: 382-397.
- [167] WANG S Y, ZHU Y X, LI Z H, et al. ChatGPT as your vehicle co-pilot: an initial attempt[J]. IEEE Transactions on Intelligent Vehicles, 2023, 8: 4706-4721.
- [168] GARRETT C R, LOZANO-PÉREZ T, KAEHLING L P. PDDL-Stream: integrating symbolic planners and blackbox samplers via optimistic adaptive planning[C]//Proceedings of the 30th International Conference on Automated Planning and Scheduling. California: AAAI, 2020: 440-448.
- [169] LEVINE S, CHELSEA F, TREVOR D, et al. End-to-end training of deep visuomotor policies[J]. Journal of Machine Learning Research, 2016, 17(1): 1334-1373.
- [170] SINGH I, BLUKIS V, MOUSAVIAN A, et al. ProgPrompt: generating situated robot task plans using large language models[C]//Proceedings of the 2023 IEEE International Conference on Robotics and Automation (ICRA). Piscataway: IEEE Press, 2023: 11523-11530.
- [171] CHEN B Y, XIA F, ICHTER B, et al. Open-vocabulary queryable scene representations for real world planning[C]//Proceedings of the 2023 IEEE International Conference on Robotics and Automation (ICRA). Piscataway: IEEE Press, 2023: 11509-11522.
- [172] KANNAN S S, VENKATESH V L N, MIN B C. SMART-LLM: smart multi-agent robot task planning using large language models[EB]. arXiv preprint, 2023, arXiv: 2309.10062.
- [173] LIN F, TIAN Y, WANG Y, et al. Airvista: empowering uavs with 3D spatial reasoning abilities through multimodal large language model agent[C]//Proceedings of the 2024 IEEE 27th International Conference on Intelligent Transportation Systems. Piscataway: IEEE Press, 2024.
- [174] TIAN Y, LIN F, ZHANG X, et al. Logisticsvista: 3D terminal delivery services with uavs, ugvs and usvs based on foundation models and scenarios engineering[C]//Proceedings of 2024 IEEE International Conference on Service Operations and Logistics, and Informatics. Piscataway: IEEE Press, 2024.
- [175] WANG X, YANG J, HAN J P, et al. Metaverses and DeMetaverses: from digital twins in CPS to parallel intelligence in CPSS[J]. IEEE Intelligent Systems, 2022, 37(4): 97-102.
- [176] YANG J, WANG X, TIAN Y L, et al. Parallel intelligence in CPSSs: being, becoming, and believing[J]. IEEE Intelligent Systems, 2023, 38(6): 75-80.
- [177] SOLDANI L, KINNEY R, BHAGIA A, et al. Dolma: an open corpus of three trillion tokens for language model pretraining research[EB]. arXiv preprint, 2024, arXiv: 2402.00159.
- [178] HUANG S H, DONG L, WANG W H, et al. Language is not all you need: aligning perception with language models[EB]. arXiv preprint, 2023, arXiv: 2302.14045.
- [179] GIRDHAR R, EL-NOUBY A, LIU Z, et al. ImageBind one embedding space to bind them all[C]//Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2023: 15180-15190.
- [180] HAN J M, ZHANG R R, SHAO W Q, et al. ImageBind-LLM: multi-modality instruction tuning[EB]. arXiv preprint, 2023, arXiv: 2309.03905.
- [181] YANG J, WANG Y T, WANG X X, et al. Generative AI empowering parallel manufacturing: building a “6S” collaborative production ecology for manufacturing 5.0[J]. IEEE Transactions on Systems, Man, and Cybernetics: Systems, 2024: 1-15.
- [182] WANG X, WANG Y T, YANG J, et al. The survey on multi-source data fusion in cyber-physical-social systems: foundational infrastructure for industrial metaverses and industries 5.0[J]. Information Fusion, 2024, 107: 102321.
- [183] 皮佩定, 倪清桦, 杨静, 等. 平行夏尔希里:生态资源智能管护及其可持续发展新途径[J]. 智能科学与技术学报, 2023, 5(3): 283-292.
- PI P D, NI Q H, YANG J, et al. Parallel Sharhili: a new approach to sustainable development and intelligent management of ecological resources[J]. Chinese Journal of Intelligent Science and Technology, 2023, 5(3): 283-292.
- [184] 李娟娟, 管桑田, 秦蕊, 等. 智能区块链与区块链智能:构筑 DePIN 的基础设施智能[J]. 智能科学与技术学报, 2024, 6(1): 5-16.
- LI J J, GUAN S T, QIN R, et al. Intelligent blockchains and blockchain intelligence: the infrastructure intelligence for DePIN[J]. Chinese Journal of Intelligent Science and Technology, 2024, 6(1): 5-16.
- [185] 秦蕊, 梁小龙, 李娟娟, 等. 平行科研院所:从数字化转型到智能化变革[J]. 智能科学与技术学报, 2023, 5(2): 212-221.
- QIN R, LIANG X L, LI J J, et al. Parallel scientific research institutes: from digital transformation to intelligent revolution[J]. Chinese Journal of Intelligent Science and Technology, 2023, 5(2): 212-221.
- [186] 李娟娟, 王戈, 王晓, 等. 加密管理:一种基于区块链的新型组织管理模式[J]. 智能科学与技术学报, 2022, 4(2): 145-156.
- LI J J, WANG G, WANG X, et al. Crypto management: a novel organizational management model based on blockchain[J]. Chinese Journal of Intelligent Science and Technology, 2022, 4(2): 145-156.
- [187] 杨静, 王晓, 王雨桐, 等. 平行智能与 CPSS:三十年发展的回顾与展望[J]. 自动化学报, 2023, 49(3): 614-634.
- YANG J, WANG X, WANG Y T, et al. Parallel intelligence and CPSS in 30 years: an ACP approach[J]. Acta Automatica Sinica, 2023, 49(3): 614-634.
- [188] 田永林, 陈苑文, 杨静, 等. 元宇宙与平行系统:发展现状、对比及展望[J]. 智能科学与技术学报, 2023, 5(1): 121-132.
- TIAN Y L, CHEN Y W, YANG J, et al. Metaverses and parallel systems: the state of the art, comparisons and prospects[J]. Chinese Journal of Intelligent Science and Technology, 2023, 5(1): 121-132.
- [189] 缪青海, 王兴霞, 杨静, 等. 从基础智能到通用智能:基于大模型的 GenAI 和 AGI 之现状与展望[J]. 自动化学报, 2024, 50(4): 674-687.
- MIAO Q H, WANG X X, YANG J, et al. From foundation intelligence to general intelligence: the state-of-art and perspectives of GenAI and AGI based on foundation models[J]. Acta Automatica Sinica, 2024, 50(4): 674-687.
- [190] GE J W, CHANG C, ZHANG J W, et al. LLM-based operating systems for automated vehicles: a new perspective[J]. IEEE Transactions on Intelligent Vehicles, 2024, 9(4): 4563-4567.
- [191] TENG S Y, YAN R, ZHANG X T, et al. Sora for hierarchical parallel motion planner: a safe end-to-end method against OOD events[J]. IEEE

Transactions on Intelligent Vehicles, 2024, 9(4): 4573-4576.

- [192] 王飞跃, 王艳芬, 陈慧竹, 等. 联邦生态: 从联邦数据到联邦智能. 智能科学与技术学报[J], 2020, 2(4): 305-313.

WANG F Y, WANG Y F, CHEN Y Z, et al. Federated ecology: from federated data to federated intelligence[J]. Chinese Journal of Intelligent Science and Technology, 2020, 2(4): 305-313.

- [193] 邓攀, 刘俊廷, 王晓, 等. STCTN: 一种基于时域偏倚校正与空域因果传递的时空因果表示学习方法[J]. 计算机学报, 2023, 46(12): 2535-2550.

DENG P, LIU J T, WANG X, et al. STCTN: a spatio-temporal causal representation learning method based on temporal bias adjustment and spatial causal transition[J]. Chinese Journal of Computers, 2023, 46(12): 2535-2550.

- [194] FAN J, TU J H, LI G L, et al. Unicorn: a unified multi-tasking matching model[J]. ACM SIGMOD Record, 2024, 53(1): 44-53.

[作者简介]



黄峻 (1998-), 男, 澳门科技大学博士生, 主要研究方向为平行智能、自动驾驶轨迹预测规划、提示工程、大语言模型等。



林飞 (1994-), 男, 澳门科技大学博士生, 主要研究方向为平行系统、人工智能药物研发、大语言模型、多模态感知、生成式人工智能。



杨静 (1998-), 女, 中国科学院自动化研究所多模态人工智能系统全国重点实验室博士生, 主要研究方向为众包、平行制造、社会制造、人工智能和社会物理信息系统。



王兴霞 (1996-), 女, 中国科学院自动化研究所多模态人工智能系统全国重点实验室博士生, 主要研究方向为平行智能、平行油田、故障诊断等。



倪清桦 (1999-), 女, 澳门科技大学博士生, 主要研究方向为平行智能、区块链、数字孪生、智能系统等。



王雨桐 (1994-), 女, 中国科学院自动化研究所复杂系统管理与控制国家重点实验室助理研究员, 主要研究方向为计算机视觉、图像异常检测。



田永林 (1994-), 男, 中国科学院自动化研究所博士后, 主要研究方向为平行系统、自动驾驶、场景工程。



李娟娟 (1986-), 女, 博士, 中国科学院自动化研究所多模态人工智能系统全国重点实验室副研究员, 主要研究方向为区块链智能、Web3.0与平行管理。



王飞跃 (1961-), 男, 博士, 中国科学院自动化研究所复杂系统管理与控制国家重点实验室主任, 澳门科技大学特聘教授, 主要研究方向为平行系统的方法与应用、社会计算、平行智能、知识自动化。