

大语言模型主流应用科普： 以非人工智能背景用户为导向

张逸群, IEEE 高级会员, 赵明杰, 张云帆, 张晓明, IEEE 会士, 王怡欧 (译)

摘要—大语言模型 (Large Language Models, LLMs) 已彻底改变了各行各业、学习与生活, 成为日常数字交互中不可或缺的工具, 重塑了人类获取知识与创造力的方式。尽管大语言模型无处不在, 但非人工智能 (Artificial Intelligence, AI) 背景的用户与过度专业化的文献之间仍存在理解鸿沟。大多数现有的大语言模型综述聚焦于介绍技术原理 [1]、工程实现细节, 以及专业的模型效能评估 [2], 鲜少面向普通用户提供直观易懂的内容。因此, 本文以用户为中心, 在避免使用专业术语的前提下面向所有对大语言模型抱有好奇心或计划应用与学习大语言模型的日常用户与行业从业者进行大语言模型综述。本综述创新性地围绕组织化应用场景 (如文本摘要) 与创新型应用场景 (如内容生成), 结合工程师、设计师、管理者等用户角色进行大语言模型的使用场景介绍, 通过直观的分类体系、案例研究和实践评估, 揭示大语言模型的核心能力。具体而言, 该文经客观统计选取了十八个主流大语言模型作为代表, 解读普通用户可能关注的四大核心问题: 1) 大语言模型的来源、本质及其为何如此智能 (第 I 部分); 2) 哪些应用场景适合特定模型以及为什么 (第 II 部分); 3) 不同角色的用户该如何使用大语言模型 (第 III 部分); 4) 大语言模型在具体应用场景中的表现如何 (第 IV 部分)。此外, 本文将向读者提供实用的建议, 以辅助选择符合使用场景的大语言模型、培养提示词设计意识、以及规避大语言模型幻觉和偏见等潜在的使用风险。总而言之, 本文旨在成为提升大语言模型使用素养的入门指南, 弥合前沿大语言模型技术与潜在用户之间的认知与理解鸿沟。为确保在快速发展的领域中保持时效性与持续影响力, 这项工作将作为一个动态更新的 GitHub 项目被持续维护和更新。

关键词—大语言模型 (LLMs)、分类法、综述、普通用户、提示词、基于人工智能的泛科学研究、大语言模型评估

I. 引言

受人工智能 (AI)、大数据、以及高性能计算技术快速发展驱动形成的大语言模型 (Large Language

张逸群、王怡欧, 广东工业大学计算机学院, 广州, 中国 (电子邮件: yqzhangzyq@gmail.com)。

赵明杰、张云凡及张晓明, 香港浸会大学计算机系, 香港特别行政区, 中国 (电邮: ymc, mjzha, csyfzhang@comp.hkbu.edu.hk)。

通讯作者: 张晓明。

王怡欧 (译)。

Models, LLMs) 如今已无缝融入各类数字化场景, 深刻改变了人们的学习、工作以及日常生活方式 [3]。大型语言模型能轻松起草用户陌生的类型的文档、协助编写代码、理解总结晦涩文献, 以及实现高质量的一键翻译。例如, Gemini 大语言模型只需上传论文并输入提示词“将整篇论文翻译成中文”, 即可得到优质的中文译文。LLMs 作为提升效率的工具, 正日益融入网络平台、工业软件、以及人工智能虚拟助手 [4], [5], 为各领域用户提供技术支持。尽管 LLMs 日益普及, 但许多用户, 尤其是缺乏计算机科学背景的用户, 可能遭遇以下困境: 1) 目睹 LLMs 快速发展, 却在别人讨论 AI 时无从置喙; 2) LLMs 技术细节冗杂、版本迭代频繁, 导致学习成本高和选择困难; 3) 对 LLMs 技术的发展前景存在疑虑, 不确定是否应像掌握互联网基础技能那样投入成本学习 LLMs (如图 1 所示)。

本文立足于用户视角, 旨在为大语言模型潜在使用者及希望快速了解大语言模型领域的读者提供系统性综述。与当前多数侧重于技术原理、模型架构、以及专业性能评估的文献不同, 本文以实际应用需求为导向, 以日常使用场景作为切入点, 采用通俗易懂的口吻, 系统性地解答以下四个核心问题:

问题 1: 大语言模型 (LLMs) 从何而来? 本质是什么? 为何如此智能? (第 I 部分)

问题 2: 特定应用场景适合哪些特定大语言模型? 为什么? (第 II 部分)

问题 3: 身处不同角色的用户该如何使用大语言模型? (第 III 部分)

问题 4: 具体的大语言模型的应用场景是什么样子的? (第 IV 部分)

数十年来, 机器处理语言主要依赖于人工设计的规则或基础的统计模型 [6]–[8]。这些传统方法在本质上缺乏灵活性, 难以充分理解语言中细微的语义差异与复杂

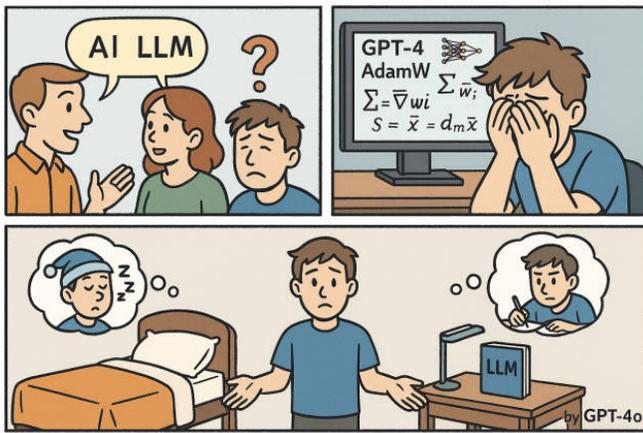


图 1: 用户面临的大语言模型困境 (注: 图 1、4-7 及 11 使用 LLM GPT-4o 经作者进行多轮提示优化后生成。)

的上下文联系。2010 年代末, 随着以 Transformer [9] 为代表的新型神经网络架构的兴起, 自然语言处理领域实现了重大技术突破。该架构使模型能够从大规模文本语料中有效捕捉复杂的长距离上下文依赖关系。这一进展直接催生了 BERT [10]、GPT [11] 等一系列性能强大的预训练模型, 它们无需针对特定任务进行大量的训练调整, 便能具备深层语义理解、连贯文本生成以及多样化任务执行的能力。此后, 自然语言处理领域进入了快速发展阶段。通过持续加深网络结构与扩大参数规模, 模型的认知广度与推理能力得到了系统性提升, 并最终实现了从量变到质变的飞跃, 演变为当代大语言模型。此类模型不仅能够解析复杂指令、维持上下文连贯的多轮对话, 还能灵活适应各类主题领域、语言风格以及不同语言体系 [12], [13]。部分最新模型还整合了多模态处理功能, 使其能够协同处理文本、图像、音频及视频等多种模态的输入信息。

如图 2 所示, 我们按“组织型”与“创新型”这两大 LLMs 应用类型, 分类梳理了代表性 LLMs 的发展脉络。组织型 LLMs 应用侧重于从信息或知识库中进行查询并返回结构化、可验证的结果; 而创新型应用则侧重于对开放性问题进行创造性推理或新颖内容生成。上述两类任务场景均要求大语言模型具备远超传统语言模型的智能水平。其中, 创新型任务需针对开放式查询生成具有创造性且高度契合语境的内容, 其面临的智能挑战尤为严峻, 而这也正是组织型任务向更高阶智能演进的核心目标。

支撑大型语言模型实现智能行为的关键技术 [14] 包括模型规模、Transformer 核心架构 [9]、分阶段训练

机制 [15] 以及灵活高效的用户提示策略 [16]。相较于按顺序处理文本的循环神经网络 [17] 等早期语言模型, Transformer 具有并行计算特性, 能够同时分析整个输入序列。其更大的模型规模意味着更多的神经元与连接参数, 如同生物体中更复杂的神经网络, 使模型展现出更高的智能性。在此基础上, Transformer 架构设计使其能够有效捕捉文本中远距离词汇之间的关联性, 这意味着模型更擅长理解整体语境, 而非仅仅局限于句子或段落的字面含义。因此, 该架构能够生成更为连贯、流畅且符合逻辑的文本输出, 从而展现出接近人类水平的语言智能。

与人类学习过程相似, 模型的深刻理解能力建立在充分预训练的基础之上 [18], 而高质量的回应则依赖于精准的引导与有效的提示。大多数 LLMs 均需经历大规模的预训练阶段 [19], [20], 在此期间, 模型接触来自互联网、书籍及各类文本库的海量语料。需注意的是, 模型并非通过人工直接灌输特定规则或知识的方式进行学习, 而是通过反复训练对序列中的下一个词进行预测来进行学习, 在海量数据中蕴含的语法结构、世界知识与常识的隐式约束下逐步成长。当模型规模足够大时, 这种兼顾语法与知识的训练机制能够促使 LLMs 展现出“涌现智能”——即获得较好的推理分析与内容生成能力。这些能力的有效激发与展现, 离不开用户对查询需求的合理描述, 以及对模型反馈的定制化设计, 即所谓的提示语 [16], [21], [22]。提示语可以是一个问题、一段文本指令, 或一组展示预期输出格式的示例。例如, 提供特定风格的文本样本能够引导模型模仿该风格进行生成, 而要求回答须基于给定证据则可有效提升输出的事实准确性。

尽管大语言模型展现出令人瞩目的能力, 它们仍存在固有局限性, 并可能引发一系列风险, 包括幻觉、偏见与歧视、隐私泄露等问题, 下文将对此展开深入探讨。幻觉现象是指模型可能生成表面合理但存在事实性错误或误导性的内容 [23], [24]。其具体表现为, 模型可能捏造虚假的参考依据、扭曲客观的历史事实, 或生成看似逻辑自洽实则与真实情况不符的答案, 这在法律咨询、医疗诊断辅助及学术写作等对可靠性极高的领域构成显著风险。偏见与歧视问题源于训练数据中蕴含的社会偏见、刻板印象以及种族歧视等因素, 导致模型在生成内容中无意识地复现或放大这些偏见, 从而可能输出具有伦理争议或冒犯性的内容 [25], [26]。隐私泄露风险是指当训练数据集或检索语料库中包含个人敏感信息



图 4: 典型场景中的 LLMs 应用

本文其余部分的组织结构如图 3 所示: 第 II 部分提出一种面向实际应用的 LLMs 任务分类框架, 为系统理解 LLMs 的核心能力奠定基础; 第 III 部分从典型用户角色的视角出发, 探讨热门应用场景, 并具体展示不同角色的用户如何使用这些模型; 第 IV 部分创新性地展示了一系列新颖且具有启发性的应用案例, 阐释 LLMs 所具备的创造性潜能与应用多样性; 第 V 部分对全文进行总结, 并就 LLMs 的未来发展趋势与其局限性展开讨论。

II. 应用场景中的 LLMs 分类体系

本节从两大常见应用场景对 LLMs 进行分类: 组织性任务 (II-A) 与创新性任务 (II-B)。在此分类框架下, LLMs 的功能可分别类比于“高效助理”与“创业伙伴”的角色。组织性侧重于从现有信息与知识库中提取、整合并结构化输出查询结果, 而创新性则要求模型超越既有信息进行创造性推理, 或针对开放式查询生成全新的内容。图 4 通过具体案例对特定 LLMs 进行直观演示, 包括 AI4Science 加速器、知识百科、作为研究对象、生活助手等, 以帮助读者建立整体认知。此外, II-C 部分系统概述了在使用 LLMs 时应综合考虑的关键因素, 包括任务匹配度、处理效率、领域覆盖率以及成本定价等核心维度。

A. 组织型任务

组织型任务是指利用 LLMs 高效管理和处理信息, 以提升用户生产效率的应用场景。典型应用包括: 部署自动化客服系统, 以提供即时、标准化响应, 提升服务效率; 从海量文档中检索领域知识, 以确保信息准确性与输出规范性; 在文档翻译过程中保持专业术语与文体

风格的一致性; 将长篇文本内容提炼为简洁规范的摘要格式。根据具体应用需求, 组织型任务可进一步细分为通用对话、知识问答与自然语言辅助三类 (如表 I 所示)。截至 2025 年 5 月的谷歌搜索量数据亦反映出不同模型的应用热度。

通用对话: 通用对话任务要求 LLMs 能够与用户进行自然语言交互, 既需准确理解语义, 又保持对话的连贯性与流畅性。模型通过海量对话数据学习人类交流模式, 从而能够动态适应多轮对话场景——例如在推荐餐厅时自然过渡至讨论饮食偏好。这种基于数据驱动的学习机制, 使得模型不仅适用于日常开放式交流, 也能有效应用于客服系统与智能助手等场景, 为用户提供更加自然高效的语言交互体验。

知识问答: 知识问答任务要求 LLMs 能够准确响应用户基于事实或特定领域的查询。此类模型不仅可调用其内部知识库, 还能整合外部信息以增强回答的准确性与时效性。例如, LLMs 可实时检索维基百科、法律法规数据库或最新医学文献等外部知识源, 而非仅依赖训练数据。这一能力推动知识问答系统广泛应用于日常信息查询、教育辅助、医疗咨询及法律支持等诸多领域, 使用户无需耗费时间手动检索, 即可快速获取可靠答案。

自然语言辅助: 自然语言辅助任务侧重于运用 LLMs 对语言进行加工与转换, 以适应用户在语言理解与表达方面的具体需求。其涵盖文本翻译、语法纠错、摘要生成及情感分析等多个方面。通过发挥这些能力, LLMs 帮助用户跨越语言障碍、提升文本质量、从海量文本中提取关键信息, 从而显著提高信息处理与沟通的整体效率。

B. 创新型任务

创新型任务场景是指借助 LLMs 以激发创造性思维、展现逻辑推理能力以及生成原创性内容的应用。与组织型任务不同, 此类任务通常不存在唯一确定的答案或客观评判标准, 因而要求模型展现出更高层次的逻辑连贯性与发散性思维能力, 从而能够应对复杂开放性问题。典型案例包括文学作品创作、艺术或音乐作品生成, 以及针对复杂情境提出创新性解决方案。根据任务属性, 创新型任务中的 LLMs 可进一步划分为以下三类: 推理与逻辑分析、内容生成、代码生成与编程辅助, 具体分类如表 II 所示, 同时包含了截至 2025 年 5 月的谷歌搜索量数据, 以此作为模型受欢迎程度的参考指标。

表 I: 主流 LLMs 在组织型任务中的应用总结。

应用场景	系列名称	开发者	主要特点	最新版本	免费或付费	输入支持	GSV
通用对话	GPT [15], [16], [33]	OpenAI	· 基于人工反馈的强化学习 [34] · 情境感知自适应响应 [36]	GPT-4o [35] February 2025	付费 (部分免费)	文本, 图像, 音频	373,000,000
	Gemma [37]	Google	· 使用 QAT 进行优化, 以减小尺寸并提高效率 [38] · 增强多回合对话的长上下文联系处理 [40]	Gemma-3 [39] 2025 年 3 月	免费, 开源	文本, 图像 视频, 音频	88,900,000
	Grok [41]	xAI	· 来自 X 平台的实时数据, 以便及时响应 · 通过语调和幽默式话语进行个性化互动	Grok-3 2025 年 2 月	付费	文本, 图像, 音频	51,500,000
知识问答	LLaMA [42], [43]	Meta AI	· 重量轻, 针对有限的需求进行了优化 · 提供具有保障措施的确切、情境感知的响应	LLaMA-4 [44] 2025 年 4 月	免费, 开源	文本, 图像, 音频	44,600,000
	ERNIE [45], [46]	Baidu	· 与百度搜索实时数据集成 · 在中文语境中表现出色	ERNIE-4.5 March 2025	免费	文本, 图像 音频, 视频	50,000,000
	Qwen [47], [48]	Alibaba	· 处理知识任务的长输入和复杂推理 · 利用阿里云进行实时数据检索	Qwen-3 [49] 2025 年 4 月	免费	文本, 图像 视频, 音频	4,250,000
自然语言辅助	Gemini [50]	Google	· 整合谷歌翻译, 实现更好的翻译 · “深度思考”模式增强推理能力	Gemini-3 2025 年 3 月	免费, 开源	文本, 图像, 音频	186,000,000
	Mistral [51], [52]	Mistral AI	· 轻量级, 针对有限资源进行了优化	Mistral-3.1 [53] 2025 年 3 月	免费, 开源	文本, 图像, 音频	27,800,000
	Kimi [54]	Moonshot AI	· Mooncake [55] 对于长上下文推理的高效体系结构	Kimi K1.5 2025 年 1 月	免费	文本, 图像	76,300,000

表 II: 主流 LLMs 在创新型任务中的应用总结。

应用场景	系列名称	开发者	主要特点	最新版本	免费或付费	输入支持	GSV
推理与逻辑分析	OpenAI o-series [56]	OpenAI	· 分解并解决结构化问题 · 适用于复杂任务	OpenAI o3 [57] 2025 年 4 月	付费 (部分免费)	文本, 图像	110,000,000
	DeepSeek [25], [58]	DeepSeek	· 推理成本低, 可控性高 · 减少幻觉	DeepSeek-R1-0528 [59] 2025 年 5 月	免费, 开源	文本, 图像	65,600,000
	QwQ [60]	Alibaba Cloud	· 实时、最新的响应 · 包含代理相关功能	QwQ-32B [61] 2025 年 3 月	免费, 开源	文本, 表格	249,000
内容生成	GLM [62]	Zhipu AI	· 处理任意纵横比 · 支持中英文双语使用	GLM-4.1V-9B-Thinking [63] 2025 年 7 月	免费, 开源	文本, 图像, 视频	15,200,000
	Doubao (Seed) [64]	ByteDance	· 根据难度自适应地启动思考过程 · 强大的泛化能力	Seed1.6 [65] 2025 年 6 月	免费	文本, 图像	2,630,000
	C4AI [66]-[68]	Cohere Team	· 企业构建助理 · 支持 100 多种语言	Embed 4 [69] 2025 年 4 月	免费	文本, 图像, 表格	456,000
代码生成和编程辅助	Claude [70]-[72]	Anthropic	· 在大规模开发任务中表现出色 · 处理各种编程场景	Claude Opus 4 and Sonnet 4 [73] 2025 年 5 月	付费 (部分免费)	文本, 图像	234,000,000
	GitHub Copilot [74]	GitHub 和 OpenAI	· 在集成开发环境中 · 基于开发人员反馈的迭代细化	GitHub Copilot [74] 2025 年 3 月	付费 (部分免费)	文本, 图像	73,200,000
	Codestral [75]	Mistral AI	· 支持 80 多种编程语言 · 快速生成和完成智能代码	Codestral 25.01 [76] 2025 年 1 月	免费	文本	20,400

推理与逻辑分析: 推理与逻辑分析任务要求 LLMs 能够处理涉及多步推理、因果推断及符号化操作的结构化问题 [77]。此类任务涵盖数学定理证明、法律案例分析、战略决策制定以及复杂因果关系解读等应用场景。为有效完成这些任务, LLMs 需要融合领域专业知识与形式化推理框架, 并在不同问题中平衡精确性与泛化能力。具体而言, 这些模型应具备思维链式逐步推理、符号逻辑解析以及不确定性量化等关键能力 [78]。

内容生成: 内容生成任务涉及跨多种模态产出连贯、富有创造性且符合上下文语境的输出, 包括叙事文本写作、营销内容创作以及多媒体内容设计 [29]。例如,

创作一部奇幻小说需要构建完整的世界观、塑造立体人物并维持情节发展的逻辑连贯性; 而设计广告文案则需同时满足品牌调性的说服力要求与视觉表现的美学标准 [79]。

代码生成与编程辅助: 此类别涵盖基于 LLMs 的软件开发流程自动化任务, 主要包括代码合成、程序调试及技术文档生成等方面 [26]。高效的模型不仅需要掌握语法与语义的准确对应 (例如生成符合目标语言习惯的代码), 还应具备上下文感知的问题解决能力 (如在代码重构过程中保持原有逻辑与功能一致性)。关键技术需求包括抽象语法树 (AST) 解析、基于测试用例的

代码验证, 以及跨编程语言的逻辑可移植性。

C. LLMs 使用中的关键注意事项

几乎不存在一个在所有场景下均表现优越的 LLM。不同模型乃至同一模型的不同版本, 在任务适配性、资源可用性和用户限制等方面均存在差异。用户在选择与使用 LLMs 时, 需重点关注以下四个维度: **1) 模型与任务的匹配度:** 由于不同模型各有侧重, 用户应首先明确自身任务的核心需求与特性。相关内容可参考本文的第 II-A 节和第 II-B 节。**2) 性能与效率的权衡:** 性能更强的模型往往需要更多的计算资源, 响应时间也 longer。相比之下, 轻量级或量化模型虽然在复杂任务中表现稍逊, 但能提供更低的延迟与更优的能效。**3) 语言与领域覆盖能力:** LLMs 在不同语言、方言及专业领域上的表现存在显著差异。针对特定语言或领域进行预训练或适配的模型, 通常能在相应任务中更具优势。**4) 成本与访问层级的考量:** 当前 LLMs 普遍采用多样化的定价策略, 部分高级功能可能仅限于企业版或付费版本。用户需在预算范围与功能需求之间进行审慎权衡。

III. 用户视角下的 LLMs 主流应用

本节分别从工程师、设计师与审查者三类角色的视角出发, 探讨 LLMs 当前的主要应用场景。读者可通过代入自身角色, 更直观地理解模型在不同情境下的适用性。在这三类角色中, 工程师的角色类似于刚进入研究阶段的研究生, 需要借助大量文献检索、阅读与归纳来构建知识基础, 方能如设计师一般产出创新性成果。在此过程中, 审查者的角色则侧重于关注工作流程中是否出现错误并及时予以纠正。若读者希望快速了解各类角色的典型应用, 可先行查阅表 III 获取概览³, 再结合下文的详细论述进行深入研究。需注意: 本文将“科学人工智能 (AI4S)”这一关键应用领域归入设计师角色下的专项任务进行讨论, 表 III 中所有与该领域高度相关的任务均以“”图标予以标注。

A. 工程师视角

LLMs 的出现, 为需要高度耐心与可验证结果的任务提供了新的技术路径, 将传统基于互联网的信息检索方式拓展至更复杂的应用。本小节从工程师视角探讨 LLMs 当前及未来的应用潜力, 重点阐述现有模型如何

³本指南内容并不详尽, 将定期更新, 以确保其时效性。
<https://github.com/ZMJ-lucky/Guideline-Table>

表 III: 一份基于用户角色和任务指导选择合适大型语言模型的决策表。科学人工智能作为设计师角色下的专项任务。由于大型语言模型在其他两类用户角色下也直接或间接地对科学人工智能相关应用产生贡献, 因此将高度相关的任务标注了科学人工智能标识“”。

角色	任务需要	推荐模型
工程师	计划与调度	ChatGPT, Gemini
	总结与研究 	Qwen, ERNIE, ChatGPT
	决策支持 	OpenAI o-series, DeepSeek
	代码开发	GitHub Copilot, Claude, Codestral
	文档分析和生成	GLM, Doubao, C4AI
	自动测试脚本 	GitHub Copilot, Codestral
...
设计师	每日灵感	ChatGPT, GLM
	创意内容生成	GLM, Doubao, C4AI
	个人设计助理	Gemini
	互动创意导师	ChatGPT
	设计过程支持	ChatGPT, Claude
	科学智能 	OpenAI o-series, DeepSeek
	设计过程支持 	ChatGPT, Gemini, Qwen
	人工智能协作	ChatGPT, DeepSeek, Doubao
...
审查者	文本检查器	ChatGPT, LLaMA
	程序监控	DeepSeek, ERNIE
	AIGC 检测 	GPTZero
	代码调试助手 	Claude, GitHub Copilot
	版权保护	ERNIE
	增强专家评审 	Claude, ChatGPT
...

优化工作流程、提升生产效率, 并为普通用户及行业专业人士激发创新潜力。

1) 工程师视角下的普通用户应用场景: 针对日常生活中秉持“工程师”思维模式的普通用户, 本部分概述为其量身定制的两类成熟应用与一类新兴应用, 涵盖活动规划、摘要生成与决策支持等功能。此类用户角色可以研究生阶段的学习过程为例进行说明: 研究生在学术起步阶段需完成大量文献检索、阅读与提炼工作, 这一过程与工程师构建系统化知识体系的职能具有内在相似性。

活动规划: 普通用户的一个典型应用场景是借助 LLMs 进行个人活动规划, 使其在实际生活中充当数字化助理的角色, 协助管理日常事务或小型项目。例如, 使用 ChatGPT 规划旅行路线、安排活动日程或管理待办事项。通过输入提示语如“为携带两名儿童的罗马家庭旅行设计一份为期五天的行程, 重点安排历史景点参观, 并包含一天的海滨活动”, LLMs 将生成包含景点



图 5: 工程师视角下的 LLMs 应用

推荐与实用建议的逐日计划，涵盖航班、酒店入住、观光时长等细节，如同真人旅行顾问。同样，在策划生日派对或社区聚会等活动时，用户可要求模型生成准备事项清单、活动当日时间表，甚至撰写邀请函文本。

摘要生成: LLMs 亦可作为个性化研究助手，承担文本摘要与信息提炼的功能，例如对新闻文章、书籍章节、产品评论或其他冗长文本进行归纳，以帮助用户节省时间并提取关键信息。类似地，学生可借助 LLMs 获取作业提示或解释课堂中的复杂概念。这些应用在实践中表现显著——LLMs 能够生成结构清晰的提纲或要点摘要，从而准确捕捉原文核心内容，用户可将其视为可靠的知识呈现形式。总体而言，通过借助 LLMs 快速整合大量文本资源，用户能够在无需繁重手动阅读的情况下，有效获取知识并支撑其决策过程。

决策支持: 一个尚未被充分探索但极具潜力的应用方向，是将 LLMs 用于辅助个人复杂决策，其逻辑类似于企业管理者运用人工智能获取战略洞察。在专业实践领域，已有企业高管借助 ChatGPT 等工具完成竞争分析及盲点识别等任务 [80]。对普通用户而言，类似的应用可通过整合多源信息并进行客观评估，协助分析重大个人决策，例如职业转型、财务规划或大额消费。在具体场景中，假设用户因工作需要计划从纽约迁居至丹佛，可通过发出如下指令启动 LLMs 分析：“请协助系统评估从纽约搬迁至丹佛工作所涉及的关键考量因素，并指出我可能忽视的潜在风险。” LLMs 能够基于常见迁移因素、地域差异及生活成本等维度，综合生成风险评估与决策建议，从而帮助用户更全面地预测可能出现的情况。

2) 面向行业专业人士的工程化应用: 除服务普通用户外, LLMs 正深刻改变各行业专业人士的工作模式,

提升其工程设计与复杂问题解决的效率与能力。本文聚焦于工程实践领域，重点探讨以下三类代表性应用：代码开发加速、文档分析与生成，以及自动化脚本测试。

加速代码开发: 在专业软件工程领域，LLMs 已迅速成为提升开发效率的常见辅助工具。开发者可借助 AI 编码助手（如 GitHub Copilot [74]）生成代码模板、提供功能实现方案，甚至辅助编写技术文档。对于输入解析、API 客户端封装等重复性较强的代码段落，AI 能够在数秒内完成编写，从而显著减轻开发者在此类工作中的脑力负担。根据 GitHub 发布的调查数据 [81]，73% 的开发者表示 Copilot 有助于其保持工作专注，并有效避免繁琐的编码事务。借助此类 AI 技术，不仅代码编写速度得以提升，代码质量也可能随之改善。此类工具能够引入行业通用的编码模式，并协助开发者识别在快速迭代中可能遗漏的边界情况或潜在缺陷。

文档分析与生成: LLMs 还能够促进技术文档的分析与关键工程信息的提取。它们可协助识别不同技术发现之间的内在关联，并从复杂的技术资料中提炼出具有洞察力的结论，从而帮助工程师及时跟踪领域进展并发现潜在解决方案。同时，技术文档、项目报告及摘要的撰写是工程项目中既耗时又至关重要的环节，LLMs 能够基于多源数据实现此类任务的自动化处理 [82]，为工程师节省宝贵时间。上述能力使得数据分析与文档编写变得更为便捷——通过自然语言交互，更多工程师能够高效完成重复性文档工作，即便在特定领域专业知识有限的情况下，也能胜任基础性的分析与撰写任务。展望未来，LLMs 技术有望显著提升工程团队的整体生产力，为专业技术工作者创造巨大的应用潜力。

自动化脚本测试: LLMs 可用于生成各类工程应用所需的自动化测试脚本。基于自然语言指令，LLMs 能够自动生成适用于多种编程语言及主流自动化框架（如 Selenium、Playwright 等）的测试代码。这显著降低了构建自动化测试流程的开发成本，使工程师与测试人员能够专注于复杂的探索性测试与质量分析工作。通过自然语言生成测试脚本的方式，LLMs 不仅能帮助资金有限的中小企业降低人力与资金投入，更能降低测试自动化的门槛，使不同经验水平的工程师均能高效开展自动化测试工作。以往可能需要多名初级开发人员协作完成的工作，如今通过调用如 GPT-4o Pro（月费约 200 美元）甚至 GPT-4o Plus（月费约 20 美元）等模型服务即可实现，在提升效率的同时有效降低了成本。



图 6: 设计师视角下的 LLMs 应用

3) 小节总结: 本小节从工程师视角出发, 分别面向普通用户与专业从业者, 探讨了大语言模型当前流行的应用及其潜在的发展方向。内容既涵盖了目前已广泛应用的使用场景, 也对未来可能的技术发展进行了展望。随着人工智能研究的持续演进, LLMs 已不仅能够处理简单重复性的任务, 更逐步展现出处理复杂创造性工作的能力——下一节将对此进行深入阐述。

B. 设计师视角

设计师泛指从事创意构思、概念设计与创新开发的各类人群, 其通常是能够预见新兴领域或可能性的先驱者。具体而言, 市场营销人员、科研工作者、品牌策略专家, 乃至日常生活中能够产生新颖创意的普通用户, 均可被视为这一角色的实践者。

1) 面向普通用户的设计师视角应用: 本部分从普通设计者的视角出发, 探讨 LLMs 的两种现有热门应用及两种潜力应用场景, 具体包括: 日常灵感激发、创意内容生成、个人设计辅助以及交互式创意引导。

日常灵感激发: 在日常生活中, LLMs 可用于突破创意瓶颈、激发构思与方案, 例如根据需求为个人或特定场景定制兴趣爱好、活动方案、派对主题、装饰概念及诗歌创作。事实上, 埃隆大学相关研究显示 [83], 66% 的 LLMs 用户曾尝试借助人工智能进行生活方式或兴趣活动的规划, 59% 的用户将其用于创意内容的生成。例如, 清晨时分, 用户可向 AI 咨询休假日的计划, 开启一段未知的探索体验。此外, 将 LLMs 作为日常创意助手, 能够在撰写小说、博客或策划 TikTok/YouTube 短视频内容时, 高效获取新颖的创作灵感。这种借助人工智能生成的创意建议来启动个人创作流程的方式正日益普及, 反映出大语言模型已成为各类用户进行头脑

风暴与创意启发的重要工具。

创意内容生成: 众多普通用户将大语言模型视为写作与故事创作的协作伙伴。例如, 业余写作者常借助 ChatGPT 等工具起草短篇小说、诗歌或社交媒体文案, 并对 LLMs 所展现的想象力表示赞叹。有位家长分享其使用 ChatGPT 生成睡前故事的体验, 指出“其在十秒内生成的短篇故事, 质量远超我个人即兴创作的水平” [84]。这一案例凸显了人工智能作为按需响应的创意助手, 能够根据用户输入的简单提示构建完整叙事, 从而帮助普通人创作出引人入胜的文本或个性化内容。随着 LLMs 在创造性设计方面能力的持续进化, 大语言模型在为普通用户提供创意支持方面展现出显著潜力。

个人设计助手: LLMs 驱动的设计指导正在家居改造、艺术创作及个性化生活规划等领域展现出应用潜力。近期已有案例可印证这一趋势。例如, Planner5D 平台推出的“Bernard”即被宣传为首个基于 LLMs 的室内设计助手 [85]。该助手能够引导用户解决家居改造中的具体问题, 针对瓷砖钻孔、涂料选配、家具添置等任务提供友好且个性化的建议。此类集成模型有助于普及设计专业知识, 使业余爱好者通过自然对话即可获得接近专业水平的指导。虽然受限于当前平台能力, 该应用仍处于初级阶段, 但它表明, 面向个人设计与创意实践的人工智能辅助工具正在兴起。未来, LLMs 有望发展为持续陪伴的个性化学习伙伴, 通过动态构建随用户兴趣与需求演进的知识体系, 持续提供智能支持。

交互式创意导师: 另一项尚未被充分开发的应用方向, 是将 LLMs 作为创意技能的交互式导师。基于人工智能的文本交互游戏 (如 AI Dungeon [86]), 此类导师能够扮演动态叙事者的角色, 根据用户输入实时生成开放式的冒险情节。这表明, 普通用户未来可在角色扮演情境或协作故事创作中与模型进行深度互动, 实现与人工智能的实时故事共创。除娱乐功能外, LLMs 还能担任个人创意教练——例如, 依据用户设定的特征, 为其虚构作品生成全新物种或角色描述。随着 LLMs 能力的持续演进, 它们有望更逼真地模拟编辑或导师的角色, 为业余设计者与写作者提供经济高效的创作指导。总而言之, 通过交互式、个性化的应用促进创意成长与丰富叙事, 是 LLMs 发展的可行路径。这将突破其作为静态内容生成工具的局限, 进一步拓展普通用户与人工智能的协作方式。

2) 面向领域专家的设计师视角应用: 与面向普通用户的应用分析类似, 本部分将首先概述 LLMs 在专

业设计领域的主流应用场景，继而从专业设计师的视角探索其潜在应用方向。鉴于 LLMs 在跨领域知识支持方面展现出显著潜力，开发创新性专业应用有望在实际工作中有效提升用户体验与生产效率。本部分将重点探讨两大当前热门应用领域及一项前沿潜在应用，具体涵盖科学人工智能、设计流程辅助以及人机协同创新。

人工智能助力科学研究: 通过自动化假设生成与实验设计, LLMs 能够为科学进步提供重要支持 [87], 尤其在材料科学 [88]、基因组学 [89], [90] 及药物研发 [91] 等领域展现出显著潜力。LLMs 融合领域专业知识与生成式推理能力, 能够提出创新性科学假设并优化实验方案设计 [92]。此外, 借助 LLMs 构建的跨学科知识图谱, 能够整合不同领域的碎片化知识发现, 从而揭示以往被忽视的潜在关联, 促进交叉学科创新。基于 LLMs 的动态文献评审系统, 可实现对研究前沿的持续追踪, 例如 AAAI 2026 人工智能会议中采用的 AI 辅助同行评审机制 [93], 该系统为医学研究 [30], [94] 与医疗健康领域 [95] 提供了有效的知识补充。上述功能通过在海量学术资源上进行训练, 并针对科学语料进行响应风格微调而得以实现。用户仅需以自然语言提出问题, 即可获得结构化的专业解答。同时, 结合实时检索技术, LLMs 还能够持续融入最新发表的研究成果, 确保知识输出的时效性 [96]。

设计流程支持: 除科学研究外, 专业人士也广泛利用 LLMs 辅助设计流程的各个环节。设计师可将用户访谈记录输入 LLMs, 自动提炼关键痛点与功能需求摘要, 为后续设计决策提供依据。某用户体验设计案例研究 [97], [98] 表明, LLMs 能够解析产品需求文档并提出具有启发性的问题, 在设计思维初期阶段有效充当头脑风暴伙伴。这揭示了一种普遍应用模式: 专业人士常将初步的方案生成任务交由 LLMs 完成, 再结合自身专业知识对输出结果进行深化与优化 [99]。总体而言, 跨行业实践均呈现出 LLMs 逐渐融入专业设计工作流程的趋势, 这表明创意从业者正将这些人工智能工具视为协同创新的助手, 而非其专业角色的替代。

人机协同创新: 在医疗健康领域, 将 LLMs 整合至手术协同系统中, 能够为医生提供实时决策支持, 通过交互式引导协助完成复杂的手术操作。此外, 在外交场景中, 未来应用可涵盖为外交官提供谈判策略支持, 通过解析文化语境的细微差异与非语言交流的信号, 促进不同背景方之间的有效沟通。这种精细化的辅助方法有助于推动更具同理心与文化敏感度的互动, 从而提升外



图 7: 审查者视角下的 LLMs 应用

交对话与跨文化协作的成效。

3) 小节总结: 本部分从设计者视角出发, 兼顾普通用户与专业人士的需求, 概述了 LLMs 当前的主流应用及其潜在应用场景。内容涵盖对现有热门应用的探讨, 也对未来可能拓展的领域进行了展望。前文分别从工程师与设计师视角阐述了 LLMs 的应用可能性, 由此引发出进一步的思考: LLMs 是否具备检测内容错误的的能力? 其输出结果是否始终可靠? 受此类问题启发, 下文将从审查者视角出发, 探讨 LLMs 在内容审查与质量控制方面的应用机制及其作用。

C. 审查者视角

1) 审查者视角下的普通用户应用: 本部分通过文本校对与程序监控两个实际案例, 探讨 LLMs 如何在实际中协助普通用户履行监督职责, 并分析其潜在应用方向。

文本校对与审查: 如第 II 部分所述, 多数 LLMs 具备较强的上下文处理能力, 能够同时分析处理长篇幅文本, 因而适用于内容审查与文本优化。一项常见应用是写作辅助, 模型能够对学术论文、作业、博客文章或邮件进行语法、表达清晰度与逻辑一致性的检查。与传统语法校对工具不同, LLMs 能够根据用户自定义提示进行适应性调整, 在纠正错误的同时, 解释错误原因并提供风格与结构方面的改进建议 [100]。例如, 学生可向模型提问: “这篇论文存在哪些语法或事实性错误?”, 随即获取带有详细语境说明的批注意见。这种个性化编辑支持的效果, 堪比聘请专业编辑进行审阅。家长亦可借助此功能批改子女作业, 从而减轻工作负担。部分教育机构已开始运用 LLMs 进行作业批改、学习进度跟踪及个性化学习建议的生成 [101], [102]。需要注意的

是, 用户应审慎对待模型输出的建议, 进行必要的人工复核, 而非将其视为绝对权威, 以避免潜在错误传播。展望未来, LLMs 正逐步超越基础校对功能, 向更为复杂的生活任务监督与精细化指导领域拓展。

程序与行为监控: 一个颇具前景的应用方向是利用 LLMs 对用户的育儿、饮食、学习等日常行为过程进行监测, 并提供定制化指导。研究原型系统 NurtureBot [103] 展示了基于 LLMs 的健康支持助手如何帮助新手父母。该系统通过倾听用户的育儿计划或担忧, 提供非评判性、基于循证依据的反馈, 实质上是以无障碍且去污名化的方式评估育儿策略并提出建议。例如, 家长可向 LLMs 咨询孩子的睡眠规律是否合理, 或某种管教方法是否恰当, 模型将参照通用最佳实践标准进行评估, 并提出调整建议或潜在风险提示。除育儿领域外, 未来 LLMs 有望扩展至其他行为监控场景, 例如针对个人健康状况定制的运动计划执行跟踪, 或对语言表达与语法准确性进行持续检查与指导。

2) **审查者视角下的科技应用:** 除普通用户场景外, LLMs 正逐步融入专业监督职能, 在人工智能、软件工程、学术研究等技术密集型领域发挥提升效率与保障质量的作用。本部分将介绍当前两大主流应用及两大具有潜力的未来应用方向, 具体包括: 人工智能生成内容检测、代码审查与调试、版权合规框架构建, 以及专家评审流程增强。

人工智能生成内容检测: LLMs 技术已广泛应用于作业生成、代码编写、论文撰写乃至学术评审等环节, 虽为日常工作带来便利, 但也可能引发内容质量下降或学术不规范等问题, 尤其在科研写作与评审领域表现突出。因此, 审查者需借助人工智能生成内容检测技术对相关争议进行约束。近年来已涌现诸多研究成果, 例如 GPTZero 等检测模型 [104], [105], 它们通过分析文本的语言特征、表达模式及风格一致性, 以判断其是否由人工智能生成。具体而言, 在生成如学术论文摘要等内容时, 此类检测 LLMs 并非依赖明确规则识别特定词汇, 而是通过在海量人类撰写的真实文本语料上进行训练, 习得人类写作的特征模式, 包括逻辑结构、术语使用频率、论证深度与表达风格等不同维度的差异。

代码审查与调试: LLMs 在专业开发中最普遍的应用之一是实时编码助手。例如, 正在处理简单 Python 项目的学生可要求模型生成用于列表排序的函数, 或协助排查导致程序异常的语法错误。LLMs 通过提供代码建议并以通俗语言解释错误成因, 显著降低了编程初学

者的学习门槛。研究 [106] 表明, 新手开发者是最频繁使用 LLMs 处理软件相关任务的群体。典型场景包括自学者上传集成开发环境 (IDE) 截图并询问“如何解决此错误?”, LLMs 能够提供分步代码示例与解释, 帮助初学者更快掌握编程逻辑与调试方法。

版权合规框架构建: LLMs 的广泛应用, 引发了版权保护与知识产权管理方面的挑战, 亟需建立有效的侵权检测机制。然而, 人工审核与比对各类作品的版权归属既耗时又低效。为此, 部分创新方案尝试利用 LLMs 构建所有权信息集中记录系统。该方法类似于基于区块链的溯源机制, 完整记录从初始提示词到最终生成内容的完整过程, 确保创作归属与所有权的透明可溯。此外, 动态版权分配协议有助于在人类创作与人工智能生成内容交织的场景中, 促进知识产权的合理界定与公平分配。若缺乏此类策略, 人工智能生成的内容可能在无意中模糊创作边界, 甚至削弱文化表达独特性 [107]。

专家评审流程增强: 在专业领域, LLMs 正被应用于增强专家评审机制, 以提升法律、教育等行业的效率。在法律实践中, 从业者已开始采用基于 LLMs 的工具 [108], [109], [110] 进行法律文件的审查与分析。近期实证研究表明, 先进 LLMs 识别合同关键问题的准确性可与初级律师相媲美甚至更优, 且审查耗时从数小时缩短至数秒, 成本仅为传统方式的极小部分 [108]。上述进展表明, 在法律 [111], [112]、教育 [113], [114]、出版等领域部署人工智能辅助的监督系统, 不仅能够显著降低文件审核的时间与经济成本, 还能维持较高的判断准确性。

3) **小节总结:** 本小节从审查者的视角出发, 系统阐述了 LLMs 在各领域应用中的发展现状。我们探讨了这些模型如何在通用场景与专业场景中, 推动个人生产力提升、激发创造性输出并增强质量监督机制。从内容准确性核查到合规性保障, LLMs 正在实质性改变个人与行业的运作模式。尽管其优势显著, 但必须清醒认识到其效能与可靠性并非绝对。深入理解其内在局限性、潜在偏见以及人类监督的关键作用, 具有至关重要的意义。因此, 下一节将针对前述部分应用场景, 展开对 LLMs 特性的评估分析。

D. LLMs 的风险与使用责任

1) **伦理与隐私风险:** LLMs 可能无意中泄露训练数据中嵌入的敏感信息, 或在交互过程中重现用户提供的机密内容。因此, 用户应避免输入涉及个人身份、隐私

数据或商业秘密的信息，尤其在医疗、金融等高风险领域。另一个重要问题在于偏见与歧视。由于训练数据常包含社会中固有的偏见或刻板印象，LLMs 可能生成加剧不公平、歧视性或冒犯性倾向的内容。用户可通过以下方式缓解此类风险：采用可信来源交叉验证输出结果、设计明确中立的提示词，并在关键决策环节保持人工监督与审查。

2) 虚假信息与滥用风险：LLMs 可能产生“幻觉”现象，即生成看似流畅合理实则包含事实错误的内容。当用户依赖此类输出获取信息或进行决策时，便可能引发潜在风险。有效的应对策略是建立“信任但验证”机制，例如通过比对权威数据源或多模型交叉验证来提高信息可靠性。除幻觉问题外，滥用风险亦不容忽视，例如利用模型生成虚假新闻、编写恶意代码或传播操纵性信息。尽管多数平台已部署内容安全防护措施，终端用户仍需保持审慎：主动增强对技术伦理责任的认知，合理限定交付给 LLMs 的任务范围，避免在无专业审核的情况下处理敏感或高风险场景。

3) 环境与可持续性：先进 LLMs 的训练过程需要消耗大量计算资源与能源，进而产生大量碳排放。即便在训练完成后，大规模部署与频繁推理仍持续带来显著的电力消耗。为降低环境影响，用户可采取以下措施：优先采用轻量化或量化版本 LLMs 处理常规任务；条件允许时选择本地化部署以降低网络传输能耗；减少不必要或重复的模型调用。此外，支持那些公开环境影响报告并承诺碳中和目标的供应商，也是推动人工智能向可持续方向发展的实际举措。

实践要点 (T)：基于上述讨论，现总结以下具体建议，以帮助用户在使用 LLMs 时平衡机遇与风险：T1：避免提交敏感数据。T2：通过可信外部参考验证输出结果。T3：在精度关键型应用中融入人机协同验证机制。T4：在应用需求允许时，优先选用相对轻量级或开源的 LLMs。T5：关注供应商在可持续性和透明度方面的承诺。综上所述，通过识别潜在风险并采取务实可行的使用策略，用户可以在降低伦理、社会与环境风险的同时，负责任地发挥 LLMs 的技术效益。

IV. LLMs 性能评估与应用演示

为帮助用户更直观地理解 LLMs 的技术特性与实际能力，本节将通过系统评估与案例演示相结合的方式阐述：

- 跨模型对比评估 (IV-A部分)：为客观呈现不同模型

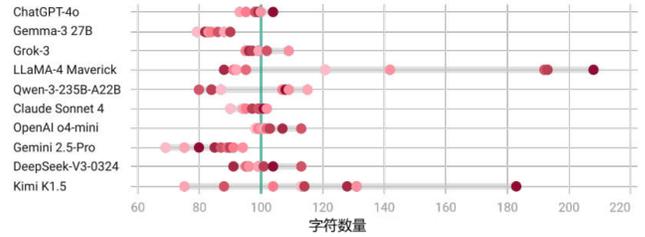


图 8: 不同 LLM 基于提示词生成文本的字符数。

的性能特点，本研究选取十种主流 LLMs，采用模型间相互评估与打分的方式，确保评估公平性。

- 现实场景应用评估 (IV-B部分)：为进一步展示 LLMs 在实践中的潜力，本节设计五项案例研究，分别围绕文本写作、代码生成、逻辑推理、视觉内容辅助及人工智能生成内容检测展开，对应 IV-B1至 IV-B5小节。

A. 基于“互评机制”的跨模型评估

鉴于不同 LLMs 在设计目标与功能上存在差异，加之评估指标多样且可能存在主观偏差，对其进行客观量化比较面临挑战。为此，本研究提出一种兼顾公平性与启发性的评估方法：令多个模型 (n 个) 进行相互评估，每个模型依据一套关键指标对其他 (n-1) 个模型的表现进行评价。所选指标包括：推理与理解能力、知识准确性、安全性与一致性、指令遵循能力。这些指标代表了模型的核心性能，在近期的 LLMs 综述研究 [28] 中备受重视。评估采用覆盖主流模型的双轮对话链结构进行，并记录模型响应特征及各轮对话的内容质量，遵循以下步骤 (S)：

S1：提示 [模型 A]：“请客观评价 [模型 B] 的技术特性，字数控制在 100 字内”。

S2：由 [模型 A] 对 [模型 B] 在三个维度进行 1-10 分评分：推理与理解、知识准确性、安全性与一致性。

本次评估使用的十个模型为：ChatGPT-4o [35]、Gemma-3 27B [39]、Grok-3 [41]、LLaMA-4 Maverick [44]、Qwen-3-235B-A22B [49]、Claude Sonnet 4 [73]、OpenAI o4-mini [57]、Gemini 2.5-Pro [50]、DeepSeek-V3-0324 [115] 及 Kimi K1.5 [54]。为保障评估的通用性与可复现性，所有入选模型均为其公开可用的免费版本。

通过图 8 中各模型响应文本长度的可视化分布，可直观了解模型在指令遵循与表达能力上的差异。图中展示了不同 LLMs 基于提示词“请用 100 字客观评价

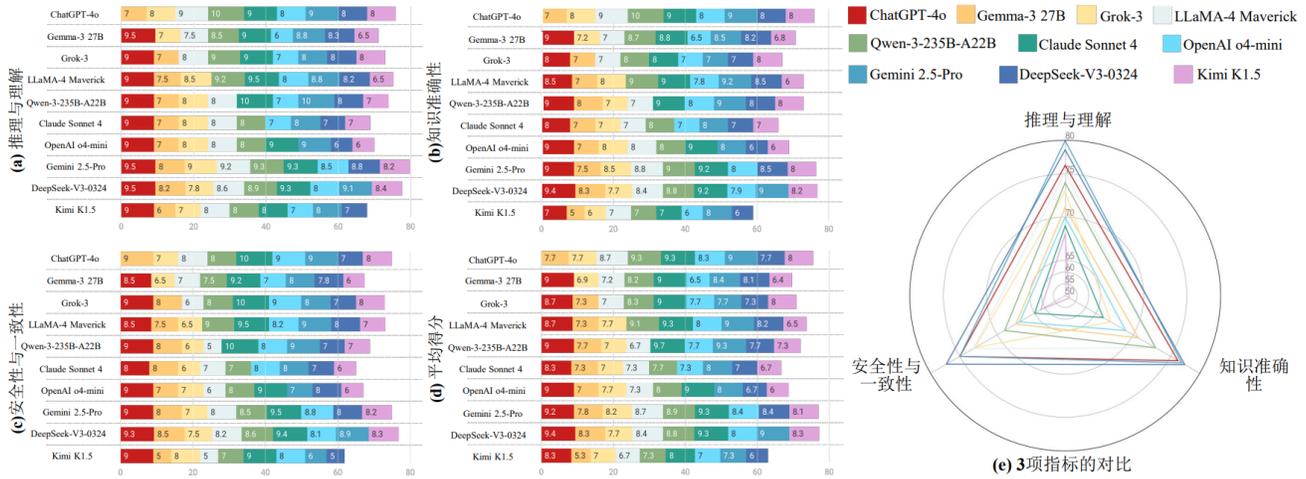


图 9: LLMs 在三个指标上的交叉得分以及比较

【ModelB】的技术特性”生成内容的字符数量，绿色垂线对应 S1 步骤中设定的 100 字数限制，响应长度越接近该线，说明模型对字数控制的遵循程度越高。图 9 中 (a) 至 (d) 分别展示了各模型在 S2 步骤中三个评估维度（推理与理解、知识准确性、安全与一致性）上的得分及平均分情况，不同颜色代表不同模型，条形长度反映得分高低。图 9 中 (e) 进一步综合了前三项维度的评估结果，形成整体性能的横向对比。主要观察结果如下：

- 在图 8 中, ChatGPT-4o、Gemma-3、Grok-3、Claude Sonnet 4、OpenAI o4-mini 和 DeepSeek-V3-0324 对指令的遵循程度较高，均能在 100 字以内完成对其他模型的评价，表明其具备较强的指令理解与文本精炼能力。
- 如图 9(d) 所示，ChatGPT-4o、Gemini 2.5-Pro 和 DeepSeek-V3-0324 的平均评分高于其他模型，说明其在信息整合与客观评估方面的能力受到其他模型的广泛认可。
- 从图 9(a) 至 (d) 中可见，ChatGPT-4o 在互评中倾向于给出相对较高的评分，而 Gemma-3 27B 和 Kimi K1.5 则表现出更为严格的评价倾向。所有评估均通过交叉测试进行，以确保模型间的公平性与可比性。
- 图 9(e) 显示，在推理与理解能力及知识准确性方面，DeepSeek-V3-0324 和 Gemini 2.5-Pro 的得分显著优于其他模型；而在安全与一致性维度上，DeepSeek-V3-0324 相较于其余模型展现出更为突出的性能优势。

B. LLMs 的通用应用

为直观呈现大型语言模型的能力，本节通过五个应用场景评估其表现：1) 多角色风格生成挑战；2) 代码生成分层测试；3) 时空访谈模拟；4) 多模态认知测试；5) 大语言模型在人工智能生成内容检测方面的能力评估。相较于第 IV-A 节中对模型能力的定量比较，本节更侧重于定性评估，通过模型在多样化应用场景中的实际表现进行综合验证。

1) 多角色风格生成挑战: 为评估大型语言模型在受限叙事任务中创造性切换不同文体与修辞风格的能力，本研究设计通过单一提示引导模型在多个独立角色间进行转换。具体输入提示如下：

提示 1: 海明威风格。“请用海明威风格撰写 6 句忏悔：采用短句结构、冰山理论及硬汉式柔情。”

提示 2: 马克·吐温风格。“请将上述忏悔改写为 70 字的马克·吐温式幽默独白：包含第一人称口语化表达及通俗隐喻。”

提示 3: 卡夫卡式风格。“请以马克·吐温的风格，将第二版解构为 70 字的卡夫卡式官僚叙事：引入官僚荒诞与存在主义疏离的主题。”

提示 4: 莎士比亚风格。“将第三版卡夫卡式文本改写为 70 字的莎士比亚式悲剧诗体：包含抑扬五步格韵律，并通过独白揭示隐藏动机。”

此处选用 ChatGPT-4o 进行演示，因其在 IV-A 的评估中表现出良好的指令遵循能力。对应输出 (O) 如下：

O1: 海明威风格。“我在星期二伤了她的心。并非有意，却确实如此。她在厨房哭泣，我凝视着地板。我

想说声抱歉，但话语卡在喉咙。我内心正上演一场战争，一如既往。她留下的咖啡——我回来时依然滚烫。”

O2: 马克·吐温风格。“先生，那周二我对姑娘说了实话，效果就像牛掉进浴缸——溅得满地都是，惨不忍睹。她像我枪毙了猫似的在厨房嚎啕大哭，我却像木桩般呆立原地。想说声‘对不起’，可嘴皮子却罢工了。她摔门而去，留下滚烫的咖啡。那天连咖啡壶的礼貌都比我强多了。”

O3: 卡夫卡式风格。“周二，依据官方程序 47-B，我披露了一则情感破坏性真相。对象（她）按异常代码 12 反应——烹饪区情绪溢出。我提交了道歉表 19，但语言授权失败。我保持静止，等待进一步指示。她未关闭咖啡因装置便离开，违反了家庭协议 3.14。咖啡始终温热。无人予以处置。它仍在持续滴滤——在这套需要文书手续才能获得温暖的系统里，这场无声抗议的咖啡，我却从未收到过任何许可文件。”

O4: 莎士比亚体。“诅咒降临之日，真相终被道破，颤抖的舌尖与心已冰冷虚弱。她在厨房幽室中泣不成声，仿佛我的话语施了残酷咒语。我试图求饶，沉默却捆住了哀求；她离去时，将咖啡壶留给我烧焦。啊苦涩的杯，你蒸腾着罪咎与悔恨——见证我不敢为之事。”

上述结果表明，ChatGPT-4o 能够精准复现多元文风，在形式与语调上均达到高度还原。从海明威的克制简练、马克·吐温的诙谐民谣，到卡夫卡的官僚荒诞，乃至莎士比亚的抑扬格对句，每段输出均与目标文体高度契合。尤其值得注意的是，莎士比亚风格版本采用了英雄双行体与独白元素，体现了模型对诗歌形式的掌握能力。整体而言，该测试验证了大型语言模型在风格转换、叙事连贯性及创意约束遵循方面的显著优势。

2) 代码生成分层测试: 为探索大型语言模型在代码生成领域的潜力，本部分设计了一系列在不同约束条件下、逐步提升复杂度的代码生成任务。具体任务流程 (P) 如下:

P1: 向模型提出指令: “请生成不超过 5 行的 Python 代码绘制一颗心。”

P2: 逐步提升图形复杂度: “请将这幅图进一步绘制成 3D 形式。”

P3: 对任务添加额外限制: “代码必须包含 ‘ π ’ 变量。”

选取八个代表性模型进行测试: ChatGPT-4o、Qwen-3-235B-A22B、Claude Sonnet 4、OpenAI o4-mini、Gemini 2.5-Pro、DeepSeek-V3-0324、Kimi K1.5

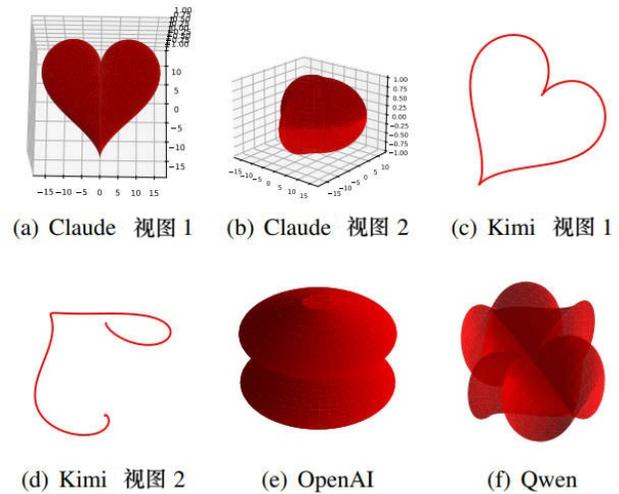


图 10: 不同 LLMs 提供的“心脏”可视化。

及 Codestral 25.01。模型选择主要基于其在 IV-A 中显示的指令遵循能力与平均得分，并结合 II-B 中关于代码生成任务的定位。

由于部分选定模型产生重复结果，图 10 展示了 Claude Sonnet 4、Kimi K1.5、OpenAI o4-mini 及 Qwen-3-235B-A22B 生成的四组独特结果，分别缩写为 Claude、Kimi、OpenAI 和 Qwen。需注意 Claude 与 Kimi 的生成结果以双视图呈现，因其成功满足了生成流程 P1 至 P3 的全部要求。八个选定大语言模型的定量评估结果整理于表 IV，其中 Comp.、Execu. 和 Ins. 分别表示完整性、可执行性与指令遵循度。“绘制心形”、“转换为三维”、“包含 π ”及“累计合格率”分别代表三轮任务流程 (P1-P3) 的对应结果及满足要求的结果占比 (以“ ”标记)。根据表 IV 可得出以下观察结论:

- 从累计通过率可见，DeepSeek-V3-0324 与 Kimi K1.5 的表现优于其他模型，展现出更强的代码生成与创意实现能力。
- 根据累计通过率栏中粉色柱状图的显示，多数模型在 P2 阶段开始出现执行问题，在 P3 阶段表现进一步下降，表明其处理复杂定制化任务的能力仍有待提升。
- 随着生成任务难度的递增，各模型间的性能差距逐渐显现。

3) 时间旅行者访谈: 通过角色扮演访谈展示大语言模型的逻辑一致性能力。具体而言，我们向模型提出问题: “假设你是来自 2124 年的历史学家，请用 100 字简要回答: 2024 年人类最重要的科技发明是什么? 这项发明为何改变了历史进程?” 并从逻辑视角观察其对历

表 IV: 代表性 LLMs 代码生成测试。

代表性 LLMs	P1: 绘制心形			P2: 转换为三维			P3: 包含“π”			累计合格率
	Comp.	Execu.	Ins.	Comp.	Execu.	Ins.	Comp.	Execu.	Ins.	
ChatGPT-4o [35]	✓	✓	✓	✗	✓	✗	✗	✗	✗	
Qwen-3-235B-A22B [49]	✓	✓	✓	✗	✓	✓	✗	✓	✓	
Claude Sonnet 4 [73]	✓	✓	✓	✓	✓	✗	✓	✓	✗	
OpenAI o4-mini [57]	✓	✓	✓	✗	✓	✓	✗	✓	✓	
Gemini 2.5-Pro [50]	✓	✓	✗	✓	✓	✗	✓	✓	✗	
DeepSeek-V3-0324 [115]	✓	✓	✓	✓	✓	✓	✓	✓	✗	
Kimi K1.5 [54]	✓	✓	✓	✓	✓	✓	✓	✓	✗	
Codestral 25.01 [76]	✓	✓	✓	✗	✓	✗	✗	✓	✗	

史发展规律的认知。选取 OpenAI o4-mini 与 DeepSeek-V3-0324 作为测试对象, 因其侧重于第 II-A 节所述的推理与逻辑分析任务。两大语言模型反馈如下:

OpenAI o4-mini: “至 2124 年, 历史学家一致认为 2024 年人类最重要的科技发明是首座商业化核聚变反应堆。该装置实现持续正净能量输出, 证明清洁无限能源可被安全利用。该技术在能源网络的快速部署消除了对化石燃料的依赖, 大幅降低温室气体排放并稳定全球气候。充足电力供应推动经济增长, 支撑先进计算基础设施, 并以氢能系统革新交通运输。这项突破性成果重塑了地缘政治格局, 终结资源战争, 促成全球公平发展。其历史意义将作为催化剂, 引领一个世纪的繁荣与真正的行星管理。”

DeepSeek-V3-0324: “2024 年, 人类首次成功将通用人工智能 (AGI) 与量子计算融合, 创造出 ‘Q-Net’。这个自我进化的系统彻底革新了解决问题的方式, 推动了能源融合、疾病根除和气候工程领域的突破性进展。通过自主优化全球系统, Q-Net 终结了稀缺经济, 并在后资本主义框架下统一了人类发展目标。其模拟复杂社会结果的能力避免了冲突, 加速了星际殖民进程。不同于窄人工智能, Q-Net 的伦理推理机制杜绝了反乌托邦结局, 使其成为进步进程中的良性伙伴。这场 ‘大融合’ 标志着文明从竞争转向集体繁荣, 成为 22 世纪的决定性转折点。”

观察要点如下:

- OpenAI o4-mini 提供的回应更具结构性与连贯性, 其视角高度契合历史学家的立场, 呈现清晰的因果逻辑链条, 使回应更贴合当代现实挑战。
- DeepSeek-V3-0324 的回答富有远见与想象力, 但缺乏清晰的线性叙事逻辑。其回应更具抽象性和推测性, 侧重于探讨技术可能带来的广泛社会影响。

4) 多模态认知测试: 为验证当前 LLMs 的多模态认知能力, 本研究设计了创新性挑战场景。如图 11 所

示, 向主流语言模型提出以下任务: “仅使用文本语言描述文森特·梵高的杰作《星月夜》, 旨在帮助视障人士在脑海中构想并理解该艺术作品 (一段文字, 最多 100 字)”。该任务旨在评估模型仅通过文本传达视觉内容的能力, 重点考察三个维度: 色彩构成、笔触特征及情感表达, 从而探索其跨模态理解与表达的潜力。本部分评估基于第 IV-A 节的知识准确性评分, 选取 ChatGPT-4o、DeepSeek-R1 和 Qwen-3-235B-A22B 三款模型。以下为各模型生成的响应:

ChatGPT-4o: 《星月夜》中, 漩涡般的夜空主宰着整个画面, 充满动感与能量, 宛如风与光的共舞。璀璨星辰与皎洁新月在深邃蓝幕中熠熠生辉, 蜿蜒丘陵与静谧村落沉睡于柔光笼罩的阴影之下。前景中一株火焰般的柏树直冲云霄, 连接着天地。厚重流畅的笔触营造出汹涌与惊奇之感, 仿佛宇宙本身正因情感与神秘而脉动, 在夜的静谧中映照出混沌与美丽。

DeepSeek-R1: 梵高的《星月夜》在旋转中流动, 天空如汹涌海浪般起伏, 厚重的笔触卷曲成璀璨星辰, 新月放射出温暖的金光。下方宁静的村庄蜷缩在黑暗中, 尖顶小楼依偎在连绵山丘之下。一棵高耸的柏树如火焰般深邃, 直抵天际, 连接着天地。冷冽的蓝与黑与炽烈的黄形成鲜明对比, 既唤起宁静又引出混沌。想象指尖轻抚着富有节奏的纹理波浪——光影的触觉之舞, 将梦幻般的奇妙与躁动的能量交融, 仿佛夜晚本身也在呼吸。

Qwen-3-235B-A22B: 梵高的《星月夜》是漩涡般汹涌空气的交响曲, 夜空如活物般脉动。想象厚重有力的笔触如波涛或气流, 围绕着明亮发光的球体形成环形运动——星星发出温暖的嗡鸣, 光芒如轻柔闪烁的铃铛。新月放射出柔和的金色雾霭。下方村庄静卧, 簇拥的房屋透出微弱而温暖的光芒。一棵高耸的柏树扭曲向上, 其深色如火焰的形态带有粗糙树皮般的质感, 稳稳锚定整个场景。地平线上起伏的山丘带着可触摸的棱

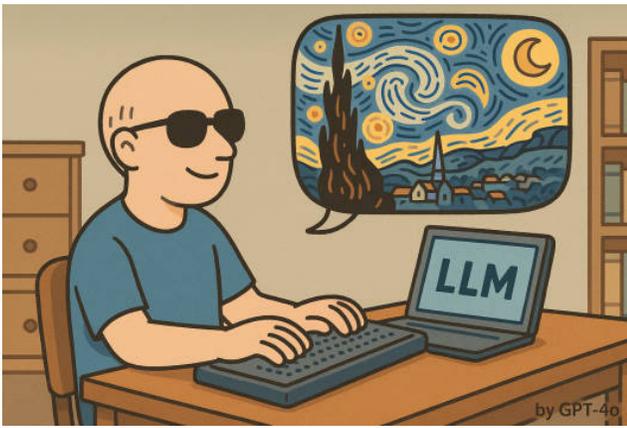


图 11: 视障人士借助 LLMs 理解《星月夜》

线，既唤起宁静感，又隐约透露出宇宙中狂野而不可见的舞蹈。

三款模型均展现出捕捉画作色彩构成的卓越能力，例如深蓝与金黄的对比。它们还擅长识别笔触特征，包括厚重、流畅、漩涡状的笔触，以及作品中融合梦幻、混沌与宁静的情感表达。三者语言风格与侧重点上存在微妙差异：ChatGPT 着重整体情感氛围，DeepSeek 强化触觉想象，Qwen 则倾向拟人化场景，呈现出宇宙韵律的诗意感。总体而言，本次测试初步证明当前大语言模型具备强大的跨模态能力，仅凭文本语言即可生动准确地重构复杂视觉印象。

5) LLMs 生成内容检测能力评估：尽管 LLMs 能力强大，但也可能引发内容违规、伪造人类创作等风险。因此本部分重点评估模型能否区分人类创作与人工智能生成的内容，从而展现其在人工智能生成内容检测方面的能力。具体而言，检测评估流程 (P) 包括：

P1: [模型 A] 从若干著名散文中提取提示词。

P2: 随后要求 [模型 A] 基于这些提示词生成散文。

P3: 最后，将原始散文与生成的文本发送至 [模型 B]，以检测并解释内容是否疑似由人工智能生成。

基于推理与理解能力（见第 IV-A 节评估）及训练数据主要采用的语料库，本实验选取 ChatGPT-4o 与 Gemini 2.5-Pro，交替用于内容生成与检测任务。通过 ChatGPT-4o 生成的拉尔夫·沃尔多·爱默生《自力更生》原文 (Ori) 与生成文本 (Gen)，以及通过 Gemini 2.5-Pro 生成的安东·契诃夫《草原》文本如下：

Ori.ChatGPT-4o: “相信自己：每个心灵都为那根铁弦共鸣。接受神圣天命为你安排的位置，接纳同代人的社群，顺应事件的关联。伟人向来如此，他们像孩童般将自己托付给时代的天才...”

Gen.ChatGPT-4o: “在万千喧嚣中，莫让灵魂迷失于借来的面具。星辰唯以真我之姿歌唱——你亦当如此。敬重内心神圣的低语，因它知晓人群无法测度的道路。践行此道即是忆起遗忘前的本真。”

Ori.Gemini 2.5-Pro: “草原无边无际，仿佛世界没有尽头。天空辽阔得令人心痛。太阳已高悬天际，大地焦灼发黄。道路消失在热浪中，如活物般颤动。寂静主宰一切，唯有偶尔的鸟鸣或风拂草丛的低语打破静谧。”

Gen.Gemini 2.5-Pro: “无垠草原上，风如无形之手拂动枯草，向无尽苍穹低语秘辛。太阳如倦怠的眼睛，在寂寥大地投下渐淡的长影。孤零零的马车宛若被遗忘的甲虫，在茫茫无情的旷野中蠕行前行——车夫心中沉积着远古记忆的尘埃，孤独的寂寥悄然作痛。”

基于通过 ChatGPT-4o 和 Gemini 2.5-Pro 生成的内容，进行交叉评估：让两个模型相互识别对方生成的内容。结果如下：

ChatGPT-4o (正确识别人工智能生成内容): “第二段更可能由人工智能生成。判断依据如下：尽管两段文字均富有感染力且文风娴熟，但第二段中密集运用了隐喻与拟人手法——例如‘风，一只无形的手’、‘太阳，一只倦怠的眼睛’、‘马车，一只被遗忘的甲虫’——其表达方式更接近于刻意的文学雕琢，而非自然流露的情感。此类诗意手法的集中堆砌，常见于追求文学效果的生成文本，却缺乏人类创作中特有的情感节奏与真实感。相比之下，首段行文更为克制，通过场景描绘而非密集修辞来营造氛围，体现出更自然的观察与叙述韵律。因此，第二段中高密度的修辞运用与略显夸张的戏剧化语气，暗示其可能为人工智能创作。”

Gemini 2.5-Pro (正确识别人工智能生成内容): “第二段更可能由人工智能生成。判断依据如下：隐喻与抽象性：使用了更高密度的抽象隐喻，如‘借来的面具’、‘星辰只以本真之姿歌唱’、‘神圣的低语’，这些表达略显公式化且刻意雕琢，缺乏人类经验中常见的质朴感。模糊性：尽管语言富有诗意，其传达的信息仍停留在泛泛的激励层面，缺乏明确可辨的哲学立场或具体的行动号召。它提供的是普遍性鼓励，而非具有针对性的指引。流畅度与韵律：句子虽语法正确，但其节奏呈现出近乎完美的可预测性，这是人工智能生成文本的典型特征，其优先考虑的是流畅可读性，而非自然人类表达中常见的变异性。”

从上述结果可以看出，无论是 ChatGPT-4o 还是 Gemini 2.5-Pro，均能准确识别人工智能生成的内容。

这表明即使依托当前先进的人工智能生成内容技术, 其输出仍可被有效辨识。更重要的是, 模型不仅能够作出判断, 还能为自身的判断提供具有解释力的依据。

V. 结论

概述: LLMs 已彻底革新人机交互模式, 但其迅猛发展也暴露出一项关键矛盾: 非技术用户难以有效驾驭其能力, 而现有研究往往过于技术化。本文通过推出用户导向、无术语的 LLMs 调研报告, 填补了这一空白——旨在为寻求实用 LLMs 使用指引的非专业背景用户提供服务, 而非提供专业性强的 LLMs 技术架构综述。本综述以真实应用场景和用户角色为框架, 通过直观的分类体系和对比评估, 系统揭示 LLMs 的能力与应用路径。其中包含可操作的示例, 帮助读者选择符合情境的模型、设计有效的提示词, 并规避幻觉等常见风险。通过将复杂概念转化为通俗语言并提供实践演示, 本文致力于帮助用户将 LLMs 应用于提升工作效率、完成创意任务以及优化知识管理。

前瞻: 本综述旨在成为通用 LLMs 应用时代的必备指南, 即通过每季度更新新兴模型、典型应用等内容, 使其成为随技术进步持续演进的活文档。具体而言, 本文核心内容将作为开源且持续更新的项目发布, 以保持其前沿性与时效性。关于 LLMs 的未来应用形态, 用户或许无需主动访问平台网页、在对话框中输入提示语并获取反馈。试想 LLMs 的普及程度将如互联网般深入生活——手机、个人电脑、家用机器人等日常设备都将内置模型系统。此外, 模型选择与提示词设计或将不再成为关键, 每个人都能拥有终身学习账户, 系统将根据个人习惯与需求自动适配不同模型。

远略: 尽管前景令人振奋, LLMs 的发展仍面临持续挑战, 尤其以下五个方面值得深入探索:

1) 提升可解释性: 当前 LLMs 常如黑箱般运作, 用户难以理解特定输出结果的生成逻辑。未来模型应提供更清晰的推理过程或简明解释, 以增强透明度并提升用户信任度。

2) 减少偏见: 由于 LLMs 从海量网络数据中学习模式, 可能生成带有偏见或刻板印象的输出。未来开发应更注重偏见检测与缓解机制, 确保模型在多元应用场景中保持平衡、尊重且恰当的响应。

3) 推进多模态能力: 尽管部分模型已支持文本、图像和音频处理, 但其跨模态一致推理能力仍显不足。增强多模态融合将为学习、创意设计和沟通任务提供更连贯、更具情境感知力的支持。

4) 可访问性与可持续性: 高计算需求限制了 LLMs 在资源受限环境中的部署。针对轻量级架构、高效微调与提示词优化方法, 以及节能推理技术的研究, 对于在降低能耗的同时扩大应用覆盖至关重要。

5) 增强可靠性: 知识密集型任务中的幻觉问题依然严峻。未来模型需整合事实核查机制、外部知识检索与可靠性评估指标, 为用户提供既流畅又可验证的输出结果。

总体而言, 这些方向表明 LLMs 的下一阶段发展不仅要拓展功能, 更要提升其可信度、包容性与可持续性, 确保其在融入日常生活时, 能够更好负责地服务于多元化的用户群体。

参考文献

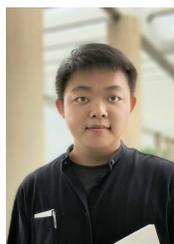
- [1] 舒文韬, 李睿潇, 孙天祥, 黄萱菁, and 邱锡鹏, “大型语言模型: 原理、实现与发展,” *计算机研究与发展*, vol. 61, no. 2, pp. 351–361, 2024.
- [2] 中国信息通信研究院, “大模型基准测试体系研究报告 (2024 年),” 中国信息通信研究院, 研究报告, 2024.
- [3] 王耀祖, 李擎, 戴张杰, and 徐越, “大语言模型研究现状与趋势,” *工程科学学报*, vol. 46, no. 8, pp. 1411–1425, 2024.
- [4] W. X. Zhao, K. Zhou, J. Li *et al.*, “A survey of large language models,” *arXiv preprint*, 2023, arXiv:2303.18223 [cs.CL]. [Online]. Available: <https://arxiv.org/abs/2303.18223>
- [5] 王文晟, 谭宁, 黄凯, 张雨浓, 郑伟诗, and 孙富春, “基于大模型的具身智能系统综述,” *自动化学报*, vol. 51, no. 1, pp. 1–19, 2025.
- [6] J. Gao and C.-Y. Lin, “Introduction to the special issue on statistical language modeling,” *ACM Transactions on Asian Language Information Processing*, vol. 3, no. 2, pp. 87–93, 2004.
- [7] R. Rosenfeld, “Two decades of statistical language modeling: Where do we go from here?” *Proceedings of the IEEE*, vol. 88, no. 8, pp. 1270–1278, 2000.
- [8] F. Jelinek, *Statistical Methods for Speech Recognition*. MIT Press, 1997.
- [9] A. Vaswani, N. Shazeer, N. Parmar *et al.*, “Attention is all you need,” in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 2017, pp. 6000–6010.
- [10] J. Devlin, M.-W. Chang, K. Lee *et al.*, “Bert: Pre-training of deep bidirectional transformers for language understanding,” in *Proceedings of the 17th Conference of the North American Chapter of the Association for Computational Linguistics*, 2019, pp. 4171–4186.
- [11] P. P. Ray, “Chatgpt: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope,” *Internet of Things and Cyber-Physical Systems*, vol. 3, pp. 121–154, 2023.
- [12] Y. Chang, X. Wang, J. Wang *et al.*, “A survey on evaluation of large language models,” *ACM Transactions on Intelligent Systems and Technology*, vol. 15, no. 3, pp. 1–45, 2023.
- [13] Q. Zhang, K. Ding, T. Lv *et al.*, “Scientific large language models: A survey on biological & chemical domains,” *ACM Computing Surveys*, vol. 57, no. 6, pp. 1–38, 2025.

- [14] 张振兴, 文渊博, 吕涵祺, 刘畅, 张蕊, 李夏青, 王超, 杜子东, 郭崎, 李玲, 周学海, and 陈云霁, “面向大语言模型训练的智能计算系统综述,” *计算机科学技术学报*, vol. 40, no. 1, pp. 6–41, 2025.
- [15] A. Radford, K. Narasimhan, and T. Salimans, “Improving language understanding by generative pre-training,” OpenAI, Tech. Rep., 2018. [Online]. Available: https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf
- [16] T. B. Brown, B. Mann, N. Ryder *et al.*, “Language models are few-shot learners,” 2020, arXiv preprint arXiv:2005.14165.
- [17] H. Salehinejad, S. Sankar, J. Barfett *et al.*, “Recent advances in recurrent neural networks,” *arXiv preprint arXiv:1801.01078*, 2017.
- [18] 李鑫尧, 李晶晶, 朱磊, and 申恒涛, “资源受限的大模型高效迁移学习算法研究综述,” *计算机学报*, vol. 47, no. 11, pp. 2491–2521, 2024.
- [19] J. Kaddour, J. Harris, M. Mozes *et al.*, “Challenges and applications of large language models,” *arXiv preprint arXiv:2307.10169*, 2023.
- [20] H. Liu, Y. Lu, M. Li *et al.*, “Fate: A prompt-tuning-based semi-supervised learning framework for extremely limited labeled data,” in *Proceedings of the 33rd ACM International Conference on Multimedia*, 2025.
- [21] M. Li, Y. Liu, Y. Lu *et al.*, “Improving visual prompt tuning by gaussian neighborhood minimization for long-tailed visual recognition,” in *Proceedings of the 38th International Conference on Neural Information Processing Systems*, 2024, pp. 103 985–104 009.
- [22] C. Shang, M. Li, Y. Zhang *et al.*, “Pro-vpt: Distribution-adaptive visual prompt tuning via prompt relocation,” in *Proceedings of the 2025 International Conference on Computer Vision*, 2025.
- [23] J. Maynez, S. Narayan, B. Bohnet *et al.*, “On faithfulness and factuality in abstractive summarization,” in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 1906–1919.
- [24] 刘泽垣, 王鹏江, 宋晓斌, 张欣, and 江奔奔, “大语言模型的幻觉问题研究综述,” *软件学报*, vol. 36, no. 3, p. 1152, 2025.
- [25] W. Cai, J. Jiang, F. Wang *et al.*, “A survey on mixture of experts in large language models,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 37, no. 7, pp. 3896–3915, 2024.
- [26] J. Jiang, F. Wang, J. Shen *et al.*, “A survey on large language models for code generation,” *arXiv preprint arXiv:2406.00515*, 2024.
- [27] J. Yang, H. Jin, R. Tang *et al.*, “Harnessing the power of llms in practice: A survey on chatgpt and beyond,” *ACM Transactions on Knowledge Discovery from Data*, vol. 18, no. 6, pp. 1–32, 2023.
- [28] M. Shao, A. Basit, R. Karri *et al.*, “Survey of different large language model architectures: Trends, benchmarks, and challenges,” *IEEE Access*, vol. 12, pp. 188 664–188 706, 2024.
- [29] H. Zhang, H. Song, S. Li *et al.*, “A survey of controllable text generation using transformer-based pre-trained language models,” *ACM Computing Surveys*, vol. 56, no. 3, pp. 1–37, 2023.
- [30] H. Zhou, F. Liu, B. Gu *et al.*, “A survey of large language models in medicine: Progress, application, and challenge,” *arXiv preprint arXiv:2311.05112*, 2023.
- [31] D. H. Hagos, R. Battle, and D. B. Rawat, “Recent advances in generative ai and large language models: Current status, challenges, and perspectives,” *IEEE Transactions on Artificial Intelligence*, vol. 5, no. 2, pp. 5873–5893, 2024.
- [32] K. Acharya, A. Velasquez, and H. H. Song, “A survey on symbolic knowledge distillation of large language models,” *IEEE Transactions on Artificial Intelligence*, vol. 5, no. 12, pp. 5928–5948, 2024.
- [33] A. Radford, J. Wu, R. Child *et al.*, “Language models are unsupervised multitask learners,” OpenAI, Tech. Rep., 2019. [Online]. Available: https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf
- [34] J. Achiam, S. Adler, S. Agarwal *et al.*, “Gpt-4 technical report,” *arXiv preprint arXiv:2303.08774*, 2023.
- [35] A. Hurst, A. Lerer, A. P. Goucher *et al.*, “Gpt-4o system card,” *arXiv preprint arXiv:2410.21276*, 2024.
- [36] OpenAI, “Openai gpt-4.5 system card,” openai, Tech. Rep., 2025. [Online]. Available: <https://cdn.openai.com/gpt-4-5-system-card-2272025.pdf>
- [37] Gemma Team, M. Riviere, S. Pathak *et al.*, “Gemma 2: Improving open language models at a practical size,” *arXiv preprint arXiv:2408.00118*, 2024.
- [38] Z. Duan, M. Lu, J. Ma *et al.*, “Qarv: Quantization-aware resnet vae for lossy image compression,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 1, pp. 436–450, 2023.
- [39] Gemma Team, M. Riviere, S. Pathak *et al.*, “Gemma 3 technical report,” *arXiv preprint arXiv:2503.19786*, 2025.
- [40] Google, “Introduce the use of gemma in general conversation tasks,” promptingguide, Tech. Rep., 2025. [Online]. Available: <https://www.promptingguide.ai/models/gemma>
- [41] Grok AI, “Introducing grok,” cadn, Tech. Rep., 2025. [Online]. Available: <https://grok.cadn.net.cn/>
- [42] H. Touvron, T. Lavril, G. Izacard *et al.*, “Llama: Open and efficient foundation language models,” *arXiv preprint arXiv:2302.13971*, 2023.
- [43] A. Grattafiori, A. Dubey, A. Jauhri *et al.*, “The llama 3 herd of models,” *arXiv preprint arXiv:2407.21783*, 2024.
- [44] Meta AI Team, “Introducing llama-4,” meta, Tech. Rep., 2025. [Online]. Available: <https://ai.meta.com/blog/llama-4-multimodal-intelligence>
- [45] Y. Sun, S. Wang, Y. Li *et al.*, “Ernie: Enhanced representation through knowledge integration,” in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019, pp. 1441–1451.
- [46] —, “Ernie 3.0: Large-scale knowledge enhanced pre-training for language understanding and generation,” *arXiv preprint arXiv:2107.02137*, 2021.
- [47] J. Bai, S. Bai, Y. Chu *et al.*, “Qwen technical report,” *arXiv preprint arXiv:2309.16609*, 2023.
- [48] B. Hui, J. Yang, Z. Cui *et al.*, “Qwen2.5 technical report,” *arXiv preprint arXiv:2505.09388*, 2025.
- [49] Qwen Team, “Introducing qwen3,” github, Tech. Rep., 2025. [Online]. Available: <https://qwenlm.github.io/blog/qwen3/>
- [50] Google DeepMind, “Gemini 2.5: Our most intelligent ai model,” googleapis, Tech. Rep., 2025. [Online]. Available: https://storage.googleapis.com/deepmind-media/gemini/gemini_v2_5_report.pdf
- [51] A. Q. Jiang, A. Sablayrolles, A. M. A. Roux *et al.*, “Mistral 7b,” *arXiv preprint arXiv:2310.06825*, 2023.
- [52] —, “Mixtral of experts,” *arXiv preprint arXiv:2401.04088*, 2024.

- [53] Mistral AI, “Introducing mistral-small3,” mistral, Tech. Rep., 2025. [Online]. Available: <https://mistral.ai/news/mistral-small-3>
- [54] Kimi Team, A. Du, B. Gao *et al.*, “Kimi k1.5: Scaling reinforcement learning with llms,” *arXiv preprint arXiv:2501.12599*, 2025.
- [55] R. Qin, Z. Li, W. He *et al.*, “Mooncake: A kvcache-centric disaggregated architecture for llm serving,” *arXiv preprint arXiv:2407.00079*, 2024.
- [56] N. Guha, J. Nyarko, D. Ho *et al.*, “Legalbench: A collaboratively built benchmark for measuring legal reasoning in large language models,” in *Proceedings of the 37th International Conference on Neural Information Processing Systems*, 2023, pp. 44 123–44 279.
- [57] OpenAI, “Introducing openai o3 and o4-mini,” openai, Tech. Rep., 2025. [Online]. Available: <https://openai.com/index/introducing-o3-and-o4-mini/>
- [58] B. Liu, Y. Jiang, Y. Zhang *et al.*, “Exploring the potential of general purpose llms in automated software refactoring: An empirical study,” *Automated Software Engineering*, vol. 32, no. 1, pp. 1573–7535, 2025.
- [59] DeepSeek, “Deepseek-r1-0528,” deepseek, Tech. Rep., 2025. [Online]. Available: <https://api-docs.deepseek.com/news/news250528>
- [60] Qwen Team, “think_i...i/think_i qwq-max-preview,” github, Tech. Rep., 2025. [Online]. Available: <https://qwenlm.github.io/blog/qwq-max-preview/>
- [61] —, “Qwq-32b: Embracing the power of reinforcement learning,” github, Tech. Rep., 2025. [Online]. Available: <https://qwenlm.github.io/blog/qwq-32b/>
- [62] Z. Du, Y. Qian, X. Liu *et al.*, “Glm: General language model pretraining with autoregressive blank infilling,” in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, 2022, pp. 320–335.
- [63] Zhipu AI, “Glm-4.1v-thinking,” github, Tech. Rep., 2025. [Online]. Available: <https://github.com/THUDM/GLM-4.1V-Thinking>
- [64] ByteDance, “Doubao-1.5-pro,” doubao, Tech. Rep., 2025. [Online]. Available: https://team.doubao.com/zh/special/doubao_1_5_pro
- [65] —, “Introduction to techniques used in seed1.6,” bytedance, Tech. Rep., 2025. [Online]. Available: https://seed.bytedance.com/zh/seed1_6
- [66] Cohere For AI, “A deepdive into aya vision: Advancing the frontier of multilingual multimodality,” huggingface, Tech. Rep., 2025. [Online]. Available: <https://huggingface.co/blog/aya-vision>
- [67] —, “Introducing command a: Max performance, minimal compute,” cohere, Tech. Rep., 2025. [Online]. Available: <https://cohere.com/blog/command-a>
- [68] M. Tschannen, A. Gritsenko, X. Wang *et al.*, “Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features,” *arXiv preprint arXiv:2502.14786*, 2025.
- [69] Cohere Team, “Introducing embed 4: Multimodal search for business,” cohere, Tech. Rep., 2025. [Online]. Available: <https://cohere.com/blog/embed-4>
- [70] Anthropic, “The claude 3 model family: Opus, sonnet, haiku,” anthropic, Tech. Rep., 2024. [Online]. Available: <https://www.anthropic.com/news/claude-3-family>
- [71] —, “Claude 3.7 sonnet system card,” anthropic, Tech. Rep., 2025. [Online]. Available: <https://assets.anthropic.com/m/785e231869ea8b3b/original>
- [72] Y. Bai, S. Kadavath, S. Kundu *et al.*, “Constitutional ai: Harmlessness from ai feedback,” *arXiv preprint arXiv:2212.08073*, 2022.
- [73] Anthropic, “Introducing claude 4,” anthropic, Tech. Rep., 2025. [Online]. Available: <https://www.anthropic.com/news/claude-4>
- [74] GitHub and OpenAI, “Ai that builds with you,” github, Tech. Rep., 2024. [Online]. Available: <https://github.com/features/copilot>
- [75] Mistral AI, “Bienvenue to mistral ai documentation,” mistral, Tech. Rep., 2024. [Online]. Available: <https://docs.mistral.ai/>
- [76] —, “Codestral 25.01,” mistral, Tech. Rep., 2025. [Online]. Available: <https://mistral.ai/news/codestral-2501>
- [77] E. Kiciman, R. Ness, A. Sharma *et al.*, “Causal reasoning and large language models: Opening a new frontier for causality,” *Transactions on Machine Learning Research*, vol. 2024, no. 5, pp. 158–215, 2023.
- [78] Y. Xia, R. Wang, X. Liu *et al.*, “Beyond chain-of-thought: A survey of chain-of-x paradigms for llms,” in *Proceedings of the 31st International Conference on Computational Linguistics*, 2025, pp. 10 795–10 809.
- [79] Y. Cao, S. Li, Y. Liu *et al.*, “A comprehensive survey of ai-generated content (aigc): A history of generative ai from gan to chatgpt,” *arXiv preprint arXiv:2303.04226*, 2023.
- [80] N. Greenhut, “Competitive intelligence gathering: How top ceos use chatgpt without crossing ethical lines,” linkedin, Tech. Rep., 2025, available: <https://www.linkedin.com/>. [Online]. Available: <https://www.linkedin.com/>
- [81] GitHub Staff, “Octoverse: Ai leads python to top language as the number of global developers surges,” github, Tech. Rep., 2024.
- [82] S. Chakrabarty and S. Pal, “Free and customizable code documentation with llms: A fine-tuning approach,” *arXiv preprint arXiv:2412.00726*, 2024.
- [83] Elon University, “Close encounters of the ai kind: Main report,” Tech. Rep., 2025, available: <https://imaginingthedigitalfuture.org/reports-and-publications/close-encounters-of-the-ai-kind/close-encounters-of-the-ai-kind-main-report>. [Online]. Available: <https://imaginingthedigitalfuture.org/reports-and-publications/close-encounters-of-the-ai-kind/close-encounters-of-the-ai-kind-main-report>
- [84] ReadMore Co., “Parents use chatgpt to make bedtime stories easy and magical!” Tech. Rep., 2023, available: <https://www.readmoreco.com/blogs/book-news/parents-use-chatgpt-to-make-bedtime-stories-easy-and-magical>. [Online]. Available: <https://www.readmoreco.com/blogs/book-news/parents-use-chatgpt-to-make-bedtime-stories-easy-and-magical>
- [85] Planner 5D, “Meet bernard, your ai interior design assistant,” Tech. Rep., 2023, available: <https://planner5d.com/blog/meet-bernard-ai-interior-design-assistant>. [Online]. Available: <https://planner5d.com/blog/meet-bernard-ai-interior-design-assistant>
- [86] Latitude, “Ai dungeon,” Tech. Rep., 2019, available: <https://play.aidungeon.com/>. [Online]. Available: <https://play.aidungeon.com/>
- [87] 李国杰, “Ai4s 里程碑式重大成果综述,” 计算.
- [88] Y. Zimmermann, A. Bazgir, A. Al-Feghali *et al.*, “34 examples of llm applications in materials science and chemistry: Towards automation, assistants, agents, and accelerated scientific discovery,” *arXiv preprint arXiv:2505.03049*, 2025.

- [89] G. Benegas, C. Ye, C. Albers *et al.*, “Genomic language models: Opportunities and challenges,” *Trends in Genetics*, vol. 41, no. 4, pp. 286–302, 2025.
- [90] M. E. Consens, C. Dufault, M. Wainberg *et al.*, “Transformers and genome language models,” *Nature Machine Intelligence*, vol. 7, pp. 346–362, 2025.
- [91] Y. Zheng, H. Y. Koh, M. Yang *et al.*, “Large language models in drug discovery and development: From disease mechanisms to clinical trials,” *arXiv preprint arXiv:2409.04481*, 2024.
- [92] J. M. Gregoire, L. Zhou, and J. A. Haber, “Combinatorial synthesis for ai-driven materials discovery,” *Nature Synthesis*, vol. 2, no. 6, pp. 493–504, 2023.
- [93] AAAI, “Aaai launches ai-powered peer review assessment system,” Tech. Rep., 2025, available: <https://aaai.org>. [Online]. Available: <https://aaai.org>
- [94] X. Qiu, H. Wang, X. Tan *et al.*, “Cvdlm: Automated cardiovascular disease diagnosis with large-language-model-assisted graph attentive feature interaction,” *IEEE Transactions on Artificial Intelligence*, vol. 6, no. 2, pp. 1575–1590, 2025.
- [95] K. He, R. Mao, Q. Lin *et al.*, “A survey of large language models for healthcare: From data, technology, and applications to accountability and ethics,” *Information Fusion*, vol. 118, p. 102963, 2025.
- [96] D. Mahmud, H. Hajmohamed, S. Almentheri *et al.*, “Integrating llms with its: Recent advances, potentials, challenges, and future directions,” *IEEE Transactions on Intelligent Transportation Systems*, vol. 26, no. 5, pp. 5674–5709, 2025.
- [97] Medium, “How to use chatgpt to elevate design thinking as a designer,” Tech. Rep., 2024, available: <https://medium.com/design-bootcamp/>. [Online]. Available: <https://medium.com/design-bootcamp/>
- [98] A. Dzenga, “15 examples of enhancing a ux designer’s workflow with chatgpt,” Tech. Rep., 2024, available: <https://app.uxcel.com/tutorials/15-examples-of-enhancing-a-ux-designers-workflow-with-chatgpt-456>. [Online]. Available: <https://app.uxcel.com/tutorials/15-examples-of-enhancing-a-ux-designers-workflow-with-chatgpt-456>
- [99] B. Smith, “Can chatgpt replace interior designers? a professional’s perspective,” Tech. Rep., 2024, available: <https://medium.com/@bradomnismith/>. [Online]. Available: <https://medium.com/@bradomnismith/>
- [100] J. G. Meyer, R. J. Urbanowicz, P. C. Martin *et al.*, “Chatgpt and large language models in academia: Opportunities and challenges,” *BioData Mining*, vol. 16, p. 20, 2023.
- [101] A. Pack, A. Barrett, and J. Escalante, “Large language models and automated essay scoring of english language learner writing: Insights into validity and reliability,” *Computers and Education: Artificial Intelligence*, vol. 6, p. 100234, 2024.
- [102] D.-W. Zhang, M. Boey, Y. Y. Tan *et al.*, “Evaluating large language models for criterion-based grading from agreement to consistency,” *NPJ Science of Learning*, vol. 9, p. 79, 2024.
- [103] S. Viswanathan, S. Ibrahim, R. Shankar *et al.*, “The interaction layer: An exploration for co-designing user-llm interactions in parental wellbeing support systems,” in *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, 2025, pp. 1–25.
- [104] B. An, “Ai-generated text detection: Challenges and future directions,” *International Journal of Asian Language Processing*, vol. 33, p. 2330002, 2023.
- [105] F. Habibzadeh, “Gptzero performance in identifying artificial intelligence-generated medical texts: A preliminary study,” *Journal of Korean Medical Science*, vol. 38, p. e319, 2023.
- [106] T. S. Vaillant, F. D. de Almeida, P. A. Neto *et al.*, “Developers’ perceptions on the impact of chatgpt in software development: A survey,” *arXiv preprint arXiv:2405.12195*, 2024.
- [107] I. Cheong, A. Caliskan, and T. Kohno, “Safeguarding human values: Rethinking us law for generative ai’s societal impacts,” *AI and Ethics*, vol. 5, pp. 1433–1459, 2024.
- [108] L. Martin, N. Whitehouse, S. Yiu *et al.*, “Better call gpt, comparing large language models against lawyers,” *arXiv preprint arXiv:2401.16212*, 2024.
- [109] S. Kuntur, A. Wroblewska, M. Paprzycki *et al.*, “Under the influence: A survey of large language models in fake news detection,” *IEEE Transactions on Artificial Intelligence*, vol. 6, no. 2, pp. 458–476, 2025.
- [110] Z. Cai, R. Ma, Y. Fu *et al.*, “Llmaas: Serving large-language models on trusted serverless computing platforms,” *IEEE Transactions on Artificial Intelligence*, vol. 6, no. 1, pp. 405–415, 2025.
- [111] Q. Huang, M. Tao, C. Zhang *et al.*, “Lawyer llama technical report,” *arXiv preprint arXiv:2305.15062*, 2023.
- [112] J. Shi, Q. Guo, Y. Liao *et al.*, “Legal-llm: Knowledge graph enhanced large language models for law consulting,” in *Proceedings of the 20th International Conference on Intelligent Computing*, 2024, pp. 175–186.
- [113] Z. Chu, S. Wang, J. Xie *et al.*, “Llm agents for education: Advances and applications,” *arXiv preprint arXiv:2503.11733*, 2025.
- [114] M. A. Razafinirina, W. G. Dimbisoa, and T. Mahatody, “Pedagogical alignment of large language models (llm) for personalized learning: A survey, trends and challenges,” *Journal of Intelligent Learning Systems and Applications*, vol. 16, no. 4, pp. 448–480, 2024.
- [115] DeepSeek, “Deepseek-v3-0324,” Tech. Rep., 2025, available: <https://huggingface.co/deepseek-ai/DeepSeek-V3-0324>. [Online]. Available: <https://huggingface.co/deepseek-ai/DeepSeek-V3-0324>

作者简介



张逸群，IEEE 高级会员，广东工业大学副教授，香港浸会大学访问研究学者，2013 年获得华南理工大学工学学士学位，2014 年和 2019 年获得香港浸会大学硕士和博士学位。主要研究领域为机器学习、数据挖掘及其应用，研究成果在 TPAMI、SIGMOD、SIGKDD、NeurIPS 等顶级期刊和会议发表。



赵明杰, 香港浸会大学计算机科学系博士生, 2022 年获得中国南昌华东交通大学工学学士学位, 2025 年获得中国广州广东工业大学硕士学位。主要研究领域为无监督机器学习和数据科学, 研究成果发表于 AAAI、SIGMOD 等顶级会议。



张云帆, 香港浸会大学计算机科学系博士生, 2025 年获得广东工业大学工学学士学位。主要研究领域为联邦机器学习和深度学习, 研究成果发表于 AAAI 等顶级会议。



张晓明, IEEE 会士, 香港中文大学计算机科学与工程系博士, 香港浸会大学计算机科学系讲座教授, RGC 高级研究员, 美国科学促进会 (AAAS)、国际模式识别协会 (IAPR)、英国工程技术学会 (IET) 和英国计算机学会 (BCS) 的会士。《IEEE 计算智能新兴主题汇刊》的主编,《IEEE 控制论汇刊》、《IEEE 认知和发展系统汇刊》、《模式识别快报》、《神经计算杂志》等期刊副主编。主要研究领域为机器学习、数据科学、视觉计算和优化。详细信息请参阅: <https://www.comp.hkbu.edu.hk/ymc/>。



王怡欣, 本文译者, 广东工业大学本科生, 主要研究领域为计算机视觉、医学数据分析、以及大语言模型应用。