微软亚洲研究院院刊

Matrix

NO.69
2024年 4 - 6月

AI for Science，憧憬一个人人都可参与科学发现的未来

周礼栋：打开创新组织的管理锦囊

价值观罗盘：如何让大模型与人类价值观对齐?

01 焦点

02 前沿求索

03 文化故事

04 媒体报道

周礼栋：打开创新组织的管理锦囊

本文转载自《商学院》，作者石丹

“创造一个空间，让每个人的每一份天赋都能得以释放和运用，并演变成为集体天才的作品。”哈佛商学院教授、领导力研究中心主席琳达·希尔（Linda Hill）在名为“如何管理集体创造力”的TED 演讲中如是说。

“这与我们的目标很类似。”微软亚洲研究院院长周礼栋说，“打造开放积极、透明无碍、多元包容的环境，吸引全球顶尖科研人才，共同创造具有广泛社会影响且突破性的前瞻技术，从而为人类的普遍知识作出贡献。这也是微软联合创始人比尔·盖茨当初成立微软研究院的初衷。”

2024年年初，微软公司首次突破3万亿美元市值大关，与苹果共同成为全球“唯二”达到这一“里程碑”的公司。摩根士丹利等多家机构分析师指出，微软市值的突破很大程度上得益于生成式 AI，也正是微软始终走在科技前沿、不断试错，才拥有了自我突破的能力。而作为微软研发机构最前沿的微软研究院，则是微软探索科技趋势、把握未来机会的触手之一。

尤其在世界又一次孕育新一代计算范式的关键时间及节点，26岁的微软亚洲研究院将如何为未来几十年的计算新范式奠定基础，并为人工智能和人类发展创造更美好的未来？要实现这样的目标，组织应该具备哪些能力？又应该如何管理？这是院长这个角色给周礼栋提出的必答题。

从科学家到管理者

周礼栋是世界级计算机专家，IEEE（电气与电子工程师学会）与 ACM（国际计算机学会）双料会士。2002年从美国康奈尔大学博士毕业后，他进入微软硅谷研究院做科研工作。他是系统研究领域首屈一指的专家，作为微软在设计和开发大规模分布式系统方面的重要技术带头人，他主持设计和开发的系统支持着微软从搜索引擎、大数据基础设施、云可靠性和可用性到 AI 基础设施的主要服务。

当然，这也是微软亚洲研究院的“传统”，历任院长（李开复、张亚勤、沈向洋、洪小文）都是技术大牛、行业翘楚。

2021年7月，周礼栋升任微软亚洲研究院院长，全面负责微软亚洲研究院在亚太地区的研究工作，以及与学术界和产业界的合作。

在谈及从科学家到管理者角色转变时，周礼栋坦言，“确实适应了一段时间”，但整体的感受是，“院长做得好的表现之一，

图1：周礼栋，微软全球资深副总裁、微软亚太研发集团首席科学家、微软亚洲研究院院长

就是其实他没那么重要。琳达·希尔说，‘创新管理绝不是由主管来界定方向然后激励其他人去执行，管理者更多起到的是协调、支持、引导和鼓励的作用。’我深以为然。在研究院管理者要做的只是项目的组合管理（portfolio management）。”周礼栋认为创新组织的管理者更像服务者，重要的是助力团队成员的想法得以实现。

其实，持续吸纳优秀且多元的人才、为研究员提供可供思想碰撞并鼓励试错的环境、创造出能让最好的想法胜出的环境，也是周礼栋履新院长以来，最重要的几项工作任务。

打造有生命力的研究机构，人才是要义

毫无疑问，对一个创新组织来说，人才是最核心的竞争力。

“创新研发的路是艰苦孤独的，如果没有足够的热爱，不具备成长型思维，可能难以持久，更遑论影响力和价值了。”周礼栋坦言，“在招聘人才的过程中，我们不仅重视应聘者在其学术领域内已有的成就，比如已发表的论文和在学术界、工业界的认可度，我们更关注的是他们未来的潜力、对创新的热情以及持续创新研究的能力。我们特别看重的是应聘者的内在驱动力、领导力、执行力，是否展现出了很好的团队合作能力和对未来技术趋势的敏锐洞察和远见。”

如今，无论是技术还是商业模式，跨界融合是大势所趋。因此，多元人才成为创新组织的“标配”。周礼栋认为，跨学科研究已成为重要趋势，很多创新来自多个领域的通力合作。“例如，我们最近招聘了一位医学领域的专家，这位同事在将机器学习技术应用于医疗健康领域，特别是疾病检测和诊断方面，有着丰富的

经验。我们还有一位研究员的教育和职业背景跨越了理论物理、神经科学和人工智能等多个领域，为研究院的人工智能与脑科学交叉研究团队带来了新的专业力量。”

当然，周礼栋和众多组织的“当家人”一样，也面临着员工代际更迭的管理挑战。“虽然我们无法成为90后、00后，但我们可以去了解他们、成就他们。”

微软亚洲研究院为新员工成立了名为“Aspire”（志望者）的内部社群，由入职三年以内的员工组成。这个社群主要是帮助新员工快速适应科研工作和文化，并且通过各种自发的项目和活动快速成长。“社群赋予这些年轻人更多职责，给了他们一个可以表达想法和建议的渠道，让他们可以按照自己的方式去影响我们的环境。”周礼栋说，“比如我们每年年会的员工互动活动，曾经是请专业公司来策划，但是我观察发现年轻人参与的热情不高。我们索性就把活动的组织设计工作交给 Aspire 团队，只要能实现我们希望同事间互相了解并能展现自己的目标就好。事实上，年轻人不会受限于过去的范式，愿意去想一些不一样的东西。”

图 2：微软亚洲研究院“Aspire”活动

周礼栋期待通过这样的一些机制培养年轻人的主人翁意识，并且意识到在这样的一个创新型组织中，不必总是遵循院长的指示，而应根据自身的能力去引发变革。周礼栋希望年轻人可以把这种思维带到研究工作中，不要因为自己年轻或者刚刚加入这个组织而不敢表达。“给年轻人舞台去展示，是一个研究机构有生命力的重要表现之一。”

构建激发内驱力的创新环境

在周礼栋看来，创新是一场灵感孕育、求证、推演、实现的协作竞赛，而透明、多元、包容的文化则有利于研发机构在创新竞赛中不断取胜。同时，企业文化真正的价值在于激发个体的智慧与行动。

基于这样的目标，周礼栋认为研发机构打造多元、包容的文化应该遵循以下几个原则：

首先，每个人都有发言权。正如两度获得诺贝尔奖的莱纳斯·卡尔·鲍林（Linus Carl Pauling）所说：“获得好想法的最佳方式是先获得很多想法。”周礼栋说，“我们的优势就是可以让研究员根据自己的兴趣、爱好和专长去探索、合作、创新，寻找适合自己的工作内容和模式。研究院正从多方面来创造更加有利于沟通的环境。比如，我们尽量避免自上而下的‘要求式’对话，构建‘信任体系’而不是‘监督体系’。当他们感到自己是在主动参与而非被要求时，将更有动力和意愿表达新的想法。”对于本身就是科研工作者出身的周礼栋来说，这些方式他都能感同身受，“对研究人员来说，在相对宽松、被充分信任的环境下思考，更容易萌生出新鲜的想法。”

其次，让思想在辩论和建设性批评中演进。“所有的思想交流，即便是以辩论、挑战甚至批判的形式出现，都有利于心境和能力的成长。”周礼栋认为，所有挑战你想法的人，都是在用另一种形式支持你，加速你的成长。

最后，让最好的想法获胜。在微软亚洲研究院，研究项目的发展和脱颖而出是一个自然选择的过程，并不是基于单一领导者的决策，而是依赖于团队和研究社区的共同认可和参与。

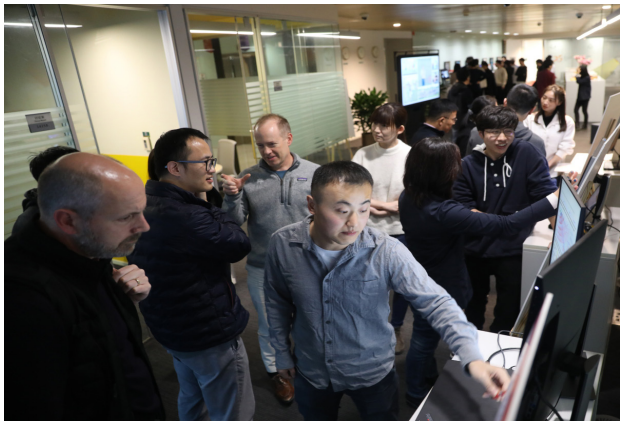


图 3：与微软研究院的同事进行交流

周礼栋介绍，“我们采用了一种类似风险投资的模式来优化资源分配和加强对项目的支持。这一模式涵盖了初期探索的种子投资以及基于长远愿景和潜力的阶段性投资。这确保了我们既鼓励自下而上的创新，也有着对长远愿景的专注投入。”

初期探索的种子投资：研究院鼓励研究员从个人兴趣出发，提出自己的研究选题，进行自下而上的创新。这种自主性不仅可以激发创新思维，还丰富了研究领域的多样性。研究院会确保每个团队都能获得足够的资源进行初期探索，这为研究员提供了充分的自由度去探索新想法，即使这些想法最初可能尚未成熟。“在这个过程中，管理者扮演的是协调者和支持者的角色，为团队提供必要的资源和指导，帮助他们将项目推动起来。这种支持方式类似于 VC 中的天使投资，重点是激励和培育有潜力的研究项目。”周礼栋表示。

基于长远愿景的阶段性投资：除了自下而上的自由探索，研究院也根据技术发展趋势和微软公司的整体战略，设定了一些战略性研究方向，从而确保资源能够投入到对公司和社会产生重大影响的领域。这种机制鼓励团队成员不仅是追求短期成果，而是基于长期视角来规划和执行研究项目。而那些展现出高潜力和符合战略目标的项目会获得更多的资源和支持，以加速它们的成长和成果的产出。这种支持是分阶段的，每一阶段的投入都取决于项目的进展、成果和未来潜力。在这一理念下，研究团队需要展现出对目标的承诺和责任感，管理者则要确保资源被有效地投入到那些有望实现长期目标和产生重大影响的项目中。

当一个研究项目被提出后，研究院会提供初始的“种子资金”和资源以支持研究员探索新的想法。如果这些想法被认为具有价值和影响力，它们就会形成天然的“磁场”，自然地吸引足够多的内外部支持，包括研究员、工程师、产品团队、合作伙伴等不同维度的人才参与，共同推进。每个员工都会做出自己的判断和选择，每个人时间的投入就是最好的支持。

在项目推进的过程中，研究院会定期组织不同形式的技术研讨会，为研究员提供机会听取来自领导、同事、内外部专家的评议和反馈。这些反馈是多维度的，结合了技术深度、工程、产品价值和社会影响力等多个方面。研究员基于反馈不断地进行自我调整和对齐，确保研究项目可以实现持续和深入的突破。在这个过程中，那些无法吸引到足够支持的项目会自然淘汰，而那些能够激发团队热情和创新精神的项目则会获得持续的投资和发展。

做“顶天”“立地”的研究

当人才梯队、研发环境都构建好之后，研究院究竟要做什么样的研究，才能引领未来5—10年的发展？周礼栋给出的答案之一是，要做“顶天”“立地”的研究。

何谓立地？周礼栋认为，研究院不能只在象牙塔里做研究，需要了解真问题、真痛点，“我们首先要做的是‘接地气’，融入行业，了解它面临的实际问题，再看怎么用技术去解决。”

据了解，微软亚洲研究院的一个团队通过深度融合 AI 技术与脑电信号、基因、血液循环等大脑关键要素，帮助人类理解大脑，进而对大脑采取更有效的保护措施。例如，他们开发的 AI Neurologist 系统，可辅助临床和科研场景下的脑电信号分析工作。原先，几分钟的脑电图进行解读需要至少一小时，耗时耗力，而且专业医疗技师非常稀缺。AI Neurologist 系统不仅提升了医护人员和神经科学家的工作效率，同时还将医生的判断准确率由原来的75%提高至90%。目前，微软亚洲研究院已经在 GitHub 上开源了脑电信号基础模型，期望有更多关注医疗领域的研究人员，一同探索运用 AI 保护大脑健康的更多可能性。

通过与行业合作，微软亚洲研究院希望从具体案例中抽象出新的 AI 逻辑，从而打造出通用的 AI 平台，并通过开源方式让更多组织机构受益。“希望我们的研究成果能够成为学术界的又一个基础工具，给其他研究人员以启发，从而共同拓展人类知识的边界。”周礼栋说。

何谓顶天？周礼栋认为，一个创新型组织的成长过程，也是不断拓展视野并承担更大社会责任的过程。“我们希望研究人员的视野和格局能打开，以高度的社会责任感，去做有意义、有价值、有温度和创造未来的科学研究和基础创新。当每个人以服务并造福社会为科学研究的终极目标，那么获得引领业界的成果也将水到渠成。”

据了解，微软亚洲研究院与甲骨文领域的专家合作，开发了甲骨文校重助手 Diviner，首次将自监督人工智能模型引入到甲骨文校重工作中，帮助专家从这项原先完全依靠人工经验、既费时又费力的工作中解放出来。在保护和传承富有社会意义的文化遗产研究中，AI 发挥了巨大价值。

再例如，随着人工智能对社会的影响日益加深，如何确保人工智能在赋能人类的同时，还要在基本价值观上与人类对齐，是所有 AI 研究者乃至社会各界都需要深入思考的问题。周礼栋表示，基于此，微软亚洲研究院开创了“社会责任人工智能”（Societal AI）的研究，并与心理学、社会学、法学等社会科学领域的专家合作，以期能为 AI 制定可行的价值观标准，确保 AI 的发展和使用总能被置于人类利益的框架之内。

周礼栋坦言，“对我们这样的科研机构来说，在基础研究方面的突破的确非常重要，但同时，我们到底能为社会做出怎样的贡献，才是我们应该去思考和拥有的抱负。”

图 4：微软亚洲研究院大合影

AI for Science，憧憬一个人人都可参与科学发现的未来

作者：刘铁岩

正处于起步阶段的 AI for Science 被认为是科学发现的第五范式。尽管目前对于 AI for Science 的定义和研究方向仍有诸多讨论，但这并不妨碍 AI for Science 已经开始在科学发现的实践中取得令人瞩目的成果。近年来，微软研究院科学智能中心杰出首席科学家刘铁岩博士和他的团队致力于推动 AI for Science 的发展和应用。在这篇署名文章中，刘铁岩博士将分享他对人工智能在科学领域关键研究方向的看法，以及对 AI for Science 未来前景的展望。

“AI for Science 预示着一种全新的科学发现范式。通过构建统一的科学基座模型，AI for Science 将消除不同科学领域之间的壁垒，实现通过一个模型解决众多科学难题的目标。它还有望推动更加普及的科学探索范式，通过与基座模型交互，让每个人都能参与到科学发现的过程中。而为了实现这些愿景，我们必须要让科学基座模型超越人类语言的限制，去学习、理解大自然的语言。”

——刘铁岩，微软研究院科学智能中心杰出首席科学家

图1：刘铁岩，微软研究院科学智能中心杰出首席科学家

今天的人工智能技术，在很多任务上的表现已经可以媲美人类，特别是在认知、感知等层面。然而，我们对人工智能的长远愿景决不能局限在复刻人类已有的知识和技能——我们更期待人工智能可以帮助人类探索未知领域，加速我们认识世界和改造世界的进程。

科学进步是推动现代人类社会发展的核心动力。因此，赋予人工智能以科学发现的能力，无疑是其发展的必然方向之一。图灵奖获得者 Jim Gray 在《科学发现的四个范式》一书中将科学发现的历程分为四个阶段：千年前的经验科学、百年前的理论科学、几十年前的计算科学，以及十几年前的数据科学。而 AI for Science 的出现将会成为前四种范式的有机结合和升华，我们称之为科学发现的“第五范式”，并不吝寄予其更大的期望。

2022年，微软研究院成立了科学智能中心（Microsoft Research AI for Science），我有幸作为该团队的创始成员之一，与世界各地的顶尖专家共同探索这一跨领域研究的开创性课题。经过两年的努力，我们在 AI for Science 的研究上取得了一系列令人振奋的成果。更重要的是，这一过程也在不断刷新我们对 AI for Science 的理解。

我想分享的一个深刻的感受：我们必须正视科学发现的艰巨性。我们决不能简单地认为只要高举 AI 的大锤，就可以轻易攻克科学发现的难题。AI for Science 的健康发展，需要我们秉承格外严谨和审慎的态度，始终对科学发现保持敬畏之心，在深入理解科学规律的基础上，对现有的 AI 工具进行改造、甚至发明全新的 AI 理论和算法。只有这样，才有可能让 AI 真正加速科学发现的进程，改变科学发现的格局。

AI for Science的三个要素

作为一个新兴领域，AI for Science 尚未有一个公认的定义。在我看来，AI for Science 并不等于“在科学研究过程中使用一些 AI 技术”。我们所追求的 AI for Science 是一个更加系统和深入的概念，AI 要深度融入科学研究的各个环节，从数据处理到仿真模拟，到实验研究，到发现新的科学规律，AI 要成为科学研究的核心技术，要为科学发现雪中送炭，而不是锦上添花。

我认为，AI for Science 应该包含三个要素：利用合成数据、构建科学基座模型、实现科学研究的闭环。

利用合成数据：在自然科学领域，有很多科学规律可以指导

我们利用计算的方法产生合成数据，比如通过求解薛定谔方程获得电子结构和分子体系的微观属性，通过求解纳维斯托克斯方程获得流体的速度和压力场。而这些合成数据不受实验条件的局限，只要有足够的计算资源就可以产生任意多的数据。通过这些合成数据训练出来的人工智能模型，可以实现对这些科学方程更加直接且高效的求解，进而用于生成更多的合成数据。这种合成数据的飞轮效应，能够让人工智能模型实现自我演化，更快、更有效地学习和提高自身能力，从而更深入地理解科学的本质，拓展科学的边界。

构建科学基座模型：AI for Science 应当遵循类似 GPT 等大模型的设计思路，用一种通用技术来解决广泛的科学问题。在过去的科学研究中，人们通常认为隔行如隔山，不同领域的科学问题需要用独立的方法来求解。但是，我们的客观世界实际上是由一些“简单通用”的底层规律所支配的。比如，无论是不规则的无机小分子、周期性的晶体材料，还是蛋白质、DNA 等生物大分子，其背后都被薛定谔方程所支配着。这种科学规律的共通性为我们整合所有科学领域、任务和模态，构建统一的科学基座模型奠定了基础。科学基座模型可以帮助我们找到复杂现象背后的规律和内在联系，在不同学科知识的碰撞中产生“1+1>2”的效果，从方法论层面影响科学发现。此外，科学基座模型还要从各种科学文献中学习人类历史上积累的科学知识及其推理能力，并在此基础上实现人类语言和科学语言的衔接，使普通人也能通过语言与基座模型交互，从而降低科学发现的专业门槛，让人人都能成为“爱因斯坦”，推动科学发现的“平权”。

实现科学研究的闭环：科学发现是一个大胆假说、小心求证的过程，后者通常依赖于实验室工作。为了实现科学发现的全链条，AI for Science 必须与真实世界形成闭环，不能仅仅局限于数字世界。近年来，实验室自动化已成为科学探索的新趋势，人工智能是这些自动化实验室的大脑，指导机械臂精确执行操作，自动合成、自动实验，从而实现从理论到实验验证的完整闭环。试想一下，一旦我们可以利用科学基座模型提出新的科学假说、进行计算仿真、再通过自动化实验室来验证，并将结果反馈给基座模型修正假说、反复迭代——以上过程能够7×24小时全天候运行，人类的科学发现能力将发生根本性的改变。

AI for Science的基座模型要读懂大自然的语言

微软研究院科学智能中心自成立之初，就将科学基座模型作为主要研究项目，并明确了科学基座模型的发展方向——科学基座模型必须突破人类语言的局限，要能够学习和理解科学概念、科学实体、科学规律，掌握支配万事万物的大自然的语言。

目前，市面上的科学大模型可以分为两个类别，一类是针对特定的垂直子领域，如蛋白质、DNA、单细胞等，设计和训练相应的大模型；另一类是将 GPT 等大语言模型进行改造或者适配到科学领域。前者只见树木，不见森林，聚焦在一个小的垂直领域，无法学到普遍的科学规律，离掌握大自然的语言相去甚远；后者则对人类语言过度依赖，作为一种基于统计的、线性、符号化的表达方式，人类语言难以完整地描述自然界的多样性和复杂性。

大自然语言是一种高维度、多模态、科学严谨的表达。首先，自然界中的物质世界是高维度、多尺度的，不同维度和尺度之间受到深层科学规律的相互制约，这些规律无法简单地用人类语言的字符序列加以表达。其次，自然界里存在各种不同的模态，比如复杂的声光电现象、波粒二象性等等，蕴含着用人类语言无法充分描述的深刻奥秘。再有，人类语言会受到个体认知和社会文化等因素的影响，存在偏倚和误差。而科学探索追求的是严谨及普适性，大自然的语言是客观存在且不受人为因素影响的。我们只有构建能够处理高维、多模态数据的科学基座模型，并将科学规律巧妙地融入模型的构建和训练过程中，才能外推到模型未曾见过的客观世界，才能真正学习和掌握大自然的语言。

聚焦微观世界的深入探索与应用

面向微观世界和宏观世界的研究是 AI for Science 的两个重要方向。由于微观世界的科学规律已经被人类充分掌握，理论完备，也有很多直接或间接的实验手段，因此 AI for Science 在微观领域大展身手具有充分的理论和实践基础。针对宏观世界，虽然人类还没有完全掌握其背后的物理规律，但也已经积累了大量数据，AI for Science 可以利用这些数据，进行规律挖掘和预测，如天气预报和气候变化研究等。

目前，微软研究院科学智能中心的 AI for Science 研究更专注于微观世界，并将相关的研究项目分成了三个层次：基础层是科学基座模型；中间层是科学仿真工具（如电子结构预测、分子动力学模拟等）；应用层是解决各领域的重大科学问题（如材料设计和药物开发等）。

在基础层，我们致力于设计和训练科学基座模型。经过近两年的深入研究，我们已经取得了一些突破性进展，开发出了基座模型的一些重要子模块，在分子科学的关键领域展示出令人振奋的能力。例如，我们在 NeurIPS 上发表的 Graphormer 模型，是科学基座模型的结构编码器，它对分子结构的理解有非常独到的能力，在第一届 OGB-LSC 分子建模比赛和 OC20 催化剂设计开放挑战赛中都力压群雄，获得冠军；我们开发的 BioGPT 模型，作为科学基座模型的序列解码器的一部分，是第一个在 Pubmed QA 任务上超过人类专家水平的 AI 模型；而我们刚刚在《自然-机器智能》（Nature Machine Intelligence）杂志上发表的用于分子结构平衡分布预测的深度学习框架 Distributional Graphormer，则是科学基座模型的结构解码器，它能够对分子的动态统计特性进行端到端的建模，在物质的微观分子结构和宏观物化属性之间建立了连接的桥梁。

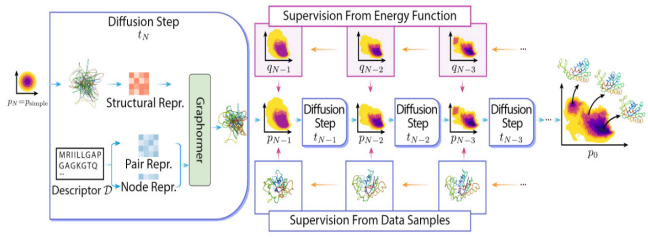


图2：Distributional Graphormer 示意图

在中间层，我们的研究重点包括电子结构预测、分子动力学模拟等，这些方向为理解和预测分子行为提供关键信息。在电子结构预测方面，我们在《自然-计算科学》(Nature Computational Science) 杂志上发表了 M-OFDFT 技术，可以利用 AI 方法将传统 DFT（密度泛函理论）的复杂度明显降低 。同时，我们还在 GPU 加速、并行计算等方面进行了更加深入的探索，进一步提高DFT的计算效率，成功将DFT计算拓展到更大尺度的分子体系，该技术已在微软 Azure 云平台上发布。在分子动力学模拟方面，我们开发了机器学习力场 ViSNet，它可以针对蛋白质等生物大分子给出精准的能量和力场的预测，相关研究成果作为编辑精选文章发表在《自然-通讯》（Nature Communications）杂志上，并且获得了首届全球 AI 药物设计大赛的冠军。

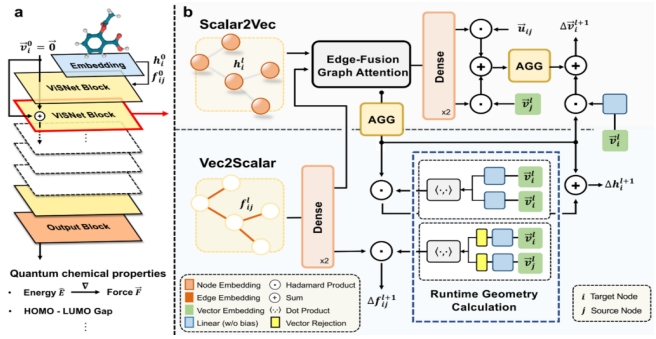


图3：ViSNet 示意图

中间层的 AI 模型和科学基座模型有着很强的依赖关系，它们会在科学基座模型的通用建模能力的基础上，再融入领域数据和洞察，通过模型微调或知识蒸馏，获得针对特定领域更高的精度或更高的效率。

在应用层，我们特别关注制药和材料领域的重大科学问题。这是当前与 AI for Science 研究最契合，而且市场需求最大的领域。在此方向上我们也取得了令人鼓舞的成果，比如能够加速发现和设计更新颖、更稳定材料的 MatterSim 和 MatterGen 模型；能够根据指定靶点，自动设计候选药物的 TamGen 模型。尤其是基于 TamGen 模型，我们与 GHDDI（全球健康药物研发中心）和盖茨基金会进行了深入合作，为肺结核和冠状病毒等仍然肆虐全球的传染病设计出了全新的高效候选药物，经过实验室合成和酶抑制试验，这些 AI 设计出来的候选药物表现出了非常优异的性能，与已知的先导化合物相比，其生物活性提高了近10倍，为治愈相关疾病做出了有益的探索。除此之外，我们也在研究科学智能体和关注实验室自动化，希望能够早日实现科学发现的自动化，助力人类文明以更快的节奏进化。我们还十分关注负责任的 AI for Science，利用法律、道德和社会规范为 AI for Science 的研究保驾护航。

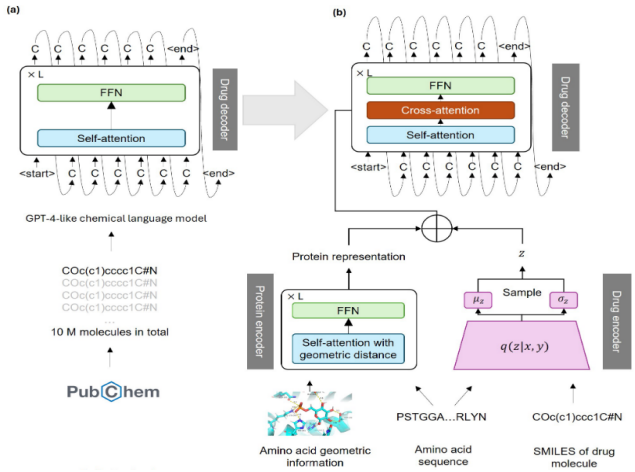


图 4：TamGen 示意图

憧憬人人都可参与科学发现的未来

AI for Science 的深入研究与发展，将为科学发现打开无限可能，为人类探索自然提供更丰富的方法和工具。利用 AI for Science，计算机模拟的精度将无限接近于现实世界实验的精度，助力科学研究的质量和效率提升至全新高度，引领科学探索进入崭新的阶段。

更重要的是，科学基座模型的引入有望使科学发现变得更加普及化。科学探索将不再仅仅是专业领域科学家们的“特权”，任何对科学发现抱有热情的人，都将能够通过语言与大模型进行交互，验证他们的奇思妙想。这将激励更多人参与解决诸如医疗健康、新材料发现、可持续发展等社会性问题，前所未有地汇聚全人类的智慧来造福世界。

当然，我们也必须清醒地认识到，AI for Science 的发展并非一蹴而就，需要长期的投入和研究，并攻克一些前所未有的挑战。作为一个高度跨学科的研究领域，AI for Science 对交叉领域人才的需求非常迫切。AI for Science 的研究者需要在计算机或自然科学领域具有很深的造诣，并且对交叉学科相互融合具备广阔的视野和开放的心态，对其他领域的难度和复杂性保持充分的理解与尊重。

算力和数据同样给 AI for Science 研究带来了极大的挑战。自然科学现象的数据类型和复杂度都远超语言数据，深入研究科学智能所需的算力和数据量也将呈指数级增长，大大高于现有的大语言模型。

此外，构建完整的 AI for Science 研究闭环并非易事。正如之

前提到的，研究闭环不仅关系到验证假说的有效性，也是衡量人工智能在科学发现中的效率和质量的关键。但传统的实验室方法论难以支持 AI for Science 的发展，我们需要全新的实践方法论，例如设计全新的实验方案和自动化流程。

尽管 AI for Science 作为新兴的科学发现范式还面临着许多未知的挑战，但我们目前所取得的每一点进展都预示着它将为人类带来无尽的可能性。AI for Science 研究中不乏令人望而却步的难题，但也正是这些难题，激发了我们探索和创新的热情。我和我的同事们将继续怀揣着极大的热忱投身于这一领域，并乐于与那些对 AI for Science 秉持严谨态度和长远愿景的各领域专家学者合作，共同推动 AI for Science 成为人类认识世界和改造世界的变革性力量。

本文作者：

刘铁岩博士，微软杰出首席科学家、微软研究院科学智能中心亚洲区负责人。他是国际电气电子工程师学会（IEEE）会士、国际计算机学会（ACM）会士、亚太人工智能学会（AAIA）会士。他（曾）被聘为卡内基梅隆大学、清华大学、香港科技大学、中国科技大学、南开大学、华中科技大学兼职教授、诺丁汉大学荣誉教授。

刘铁岩博士的先锋性研究促进了机器学习与信息检索之间的融合，被公认为“排序学习”领域的代表人物。近年来他在深度学习、强化学习、工业智能、科学智能等方面颇有建树，在顶级国际会议和期刊上发表论文数百篇，被引用数万次。他曾担任 WWW/WebConf、SIGIR、NeurIPS、ICLR、ICML、IJCAI、AAAI、KDD 等十余个国际顶级学术会议的大会主席、程序委员会主席或（资深）领域主席；包括 ACM TOIS、ACM TWEB、IEEE TPAMI 在内的知名国际期刊副主编。

刘铁岩博士毕业于清华大学，先后获得电子工程系学士、硕士及博士学位。

相关链接：

Graphormer项目页面
https://www.microsoft.com/en-us/research/project/graphormer/

MatterGen: a generative model for inorganic materials design
https://arxiv.org/abs/2312.03687

Bio-GPT: Generative Pre-trained Transformer for Biomedical Text Generation and Mining
https://arxiv.org/abs/2210.10341

TamGen: Target-aware Molecule Generation for Drug Design Using a Chemical Language Model
https://www.biorxiv.org/content/10.1101/2024.01.08.574635v2.full.pdf

MatterSim: A Deep Learning Atomistic Model Across Elements, Temperatures and Pressures
https://arxiv.org/abs/2405.04967

相关阅读：

Distributional Graphormer：
从分子结构预测到平衡分布预测

微软研究院科学智能中心发布了可用于预测分子结构平衡分布的深度学习框架 Distributional Graphormer (DiG)。DiG 可以快速生成真实多样的构象，进而为实现从单一结构预测到平衡分布预测的突破奠定基础。

扫描二维码查看文章

ViSNet：
用于分子性质预测和动力学模拟的通用分子结构建模网络

微软研究院科学智能中心提出了通用分子结构建模网络 ViSNet。在多个分子动力学基准测试中，ViSNet 均表现优异。相关文章已发表在《自然-通讯》杂志上。

扫描二维码查看文章

AI助力M-OFDFT实现兼具精度与效率的电子结构方法

微软研究院科学智能中心的研究员们基于人工智能技术和无轨道密度泛函理论（OFDFT）开发了一种新的电子结构计算框架 M-OFDFT。这一框架不仅保持了与 KSDFT 相当的计算精度，而且在计算效率上实现了显著提升，并展现了优异的外推性能，为电子结构方法开辟了新的思路。

扫描二维码查看文章

统一化数据库：为大语言模型垂域应用奠定基础

检索增强生成（RAG）技术因在减少生成幻觉和虚构信息方面的显著效果，以及对知识及时更新能力的改善，正逐渐成为大语言模型系统的主流架构之一。随着 RAG 技术的广泛应用，其核心组件——向量数据库，也开始受到越来越多的关注，成为大模型中不可或缺的外挂知识库。

然而，向量数据库与传统关系型数据库有着显著区别，这给数据的统一管理、查询和更新带来了诸多不便。为此，微软亚洲研究院开发了 VBase 复杂数据库查询系统，为统一化数据库奠定了基础，并推出了有助于向量索引实时更新的 SPFresh 方案，以及可对稀疏向量索引与稠密向量索引统一化查询的 OneSparse 系统。

如今大语言模型（LLMs）已成为内容创作、语言理解和智能对话等领域中的关键技术，但这些模型都是基于固定训练数据中观察到的规律和模式来生成回答的，可能会产生幻觉和虚构信息，并在实时的知识更新方面存在困难。检索增强生成（RAG）技术可以将最新的外挂知识库与大语言模型有机结合，把相关的精确知识放入上下文中，来引导回答的生产过程，增强大语言模型的性能与可靠性。

然而，RAG 的核心组件之一——向量数据库，在存储、查询等机制上与传统的关系型数据库存在显著区别。这给日益丰富和不同模态知识的统一管理带来了挑战。在这种背景下，微软亚洲研究院系统与网络组的研究员们认为，一种能够有效管理丰富属性和模态的外部知识的统一化数据库，将成为大语言模型广泛应用和可靠性保证的关键。

“随着大模型能力的不断增强，文字、图像、视频等各种形式的数据都可以通过机器学习技术编码成高维向量，将知识的细节属性，如图片的类型、用户的偏好等，转换为不同维度的数据。但是，多样化的知识表示方式给复杂向量数据和标量数据的有效管理带来了挑战，如何在这些混合信息中实现高效且准确的查询也变得更加困难。这就需要一种统一化的数据库来管理这些外部知识，为大语言模型提供更坚实的知识支持。”微软亚洲研究院（温哥华）首席研究员陈琪表示。

以医疗辅助诊断场景为例，医生可能需要在患者记录数据库中进行如下查询：“在60岁以上的患者中，某些X光图像类似的病例，患有不同疾病的概率是多少？”这样的操作不仅需要从标量数据库中查询年龄、性别、诊断结果等标量数据，也需要从向量数据库中查询X光图像和实验室结果等高维向量数据。由于两种数据库的存储和查询机制截然不同，所以只有通过更高级的标量—向量混合数据分析技术，才可以将向量数据库与传统数据库进行有效统一。

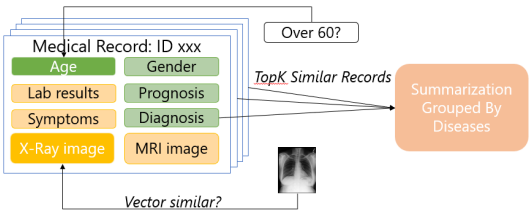


图1：未来的统一化数据库

VBase复杂查询系统：为向量索引和标量索引扫描提供统一化基石

向量数据库与标量数据库具有不同的索引扫描模式，缺乏统一的基础，这是构建统一化数据库首先要解决的问题。

标量数据库索引基于数值顺序构建，索引扫描具有严格的单调性（strict monotonicity），这也是关系型数据库能够高效执行查询的原因。例如，在购物平台上搜索价格在100到200元之间的衣服，系统会从价格100元开始扫描，一旦价格超过200元，查询就会终止。显然这种基于单调性的标量查询具有很高的效率。

相比之下，向量索引是基于高维空间中的接近性构建的，索引遍历无法遵循严格的顺序，因此缺乏单调性。向量索引仅为查询提供近似的空间导航，以近似地接近最近的子空间。为了实现提前终止，向量索引扫描过程依赖 TopK 算法来预测 K 值的临时顺序。换言之，由于没有明确的起点，在高维向量空间中寻找与目标距离最近的向量时，尽管可以利用顺序来提前终止执行，但这种方法效率很低。

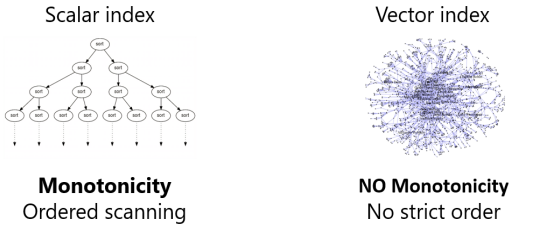


图2：标量数据库与向量数据库的检索查询

例如，假设用户有一张衣服的图片，想要在购物平台上找到相似且价格低于200元的商品，传统的方法是先进行大规模的相似性查询，然后根据价格进行过滤。比如，为了找到最相似且价格合适的前10个结果，可以先将搜索范围设定为1000个候选项，并通过价格条件逐一筛选，直到找到10个符合条件的结果为止。如果结果不足，则进一步扩大搜索范围到2000或者3000个，直到满足要求。

这种方法的核心思想是将向量数据的检索结果，转换成遵循严格单调性的标量数据库，再进行标量查询。TopK 算法被用于收集 K 个最接近的向量结果，并根据与目标向量的距离进行排序，从而创建一个具有单调性的临时索引，然后对这个临时索引数据库进行过滤。

这种方法的问题在于，无法保证返回的 K 个结果能满足最终的过滤查询需求。因此，为了确保过滤结果满足要求，要么 TopK 需要执行更广泛的相似性查询，返回更多的 K；要么在 K 不足时，重复执行 TopK 查询，但这两种做法都会导致次优的查询性能。

研究员们通过分析大量向量索引发现，向量索引查询提前终止并不需要严格的单调性，而是表现出一种放松单调性（Relaxed Monotonicity），标量索引只是这种放松单调性的特殊情况。

基于这一发现，研究员们开发了 VBase 复杂查询系统，该系统为向量索引和标量索引的高效扫描提供了统一化基石，使得各类索引的扫描遵循相同的接口和提前终止条件。这一创新使得向量数据库在执行复杂查询时的性能提升了10至1000倍，同时提高了查询的精确度。

VBase 使得构建能够执行各类复杂关系型向量和标量混合查询的统一化数据库成为可能。目前，基于 VBase 系统，一家开源数据库平台成功构建了自己的多模态向量数据库。

SPFresh：首次实现向量索引的实时就地增量更新

以向量数据库检索为基础的 RAG 技术显著提高了大语言模型生成结果的准确性。但这一优势的实现有一个关键前提：向量数据库中的数据需要保持更新，也就是说向量索引需要即时更新。对于具有成百上千维度的向量来说，更新工作并非易事——重构向量索引的时间成本需要以天来计算。

标量数据库通常使用 B 树或 B+树方法，通过二分查找定位到指定位置后直接插入信息即可完成更新。然而，向量数据库的更新要复杂得多。

以目前流行的细粒度基于图的向量索引和粗粒度基于集群的向量索引为例。在细粒度图向量索引插入或删除向量时，都需要进行大规模的图扫描以找到适当的距离进行插入，这对计算资源的要求非常高，而且删除不当还会导致性能和准确性下降。在粗粒度的集群索引更新中，虽然插入或删除向量只涉及对分区的修改，成本较低，但随着分区更新的累积，数据分布会变得不平衡，从而影响查询延迟和准确性，使索引质量下降。

现有的向量索引更新方法依赖于周期性的全局重建，这种方法速度慢且资源消耗大。尽管重建后性能和准确性会立即得到刷新，但在两次重建之间，性能和准确性会逐渐下降。此外，全局重建成本非常高，其所需的资源是传统索引的10倍以上，甚至超过索引服务的成本。

为解决这些问题，研究员们提出了 SPFresh 解决方案，该方案首次实现了向量索引的实时就地增量更新，为统一化数据库的更新提供了一种高效的方法。SPFresh 的核心是 LIRE——一种轻量级的增量再平衡协议，用于分割向量分区并重新分配分区中的向量以适应数据分布的变化。LIRE 通过仅在分区边界处重新分配向量，实现了低资源消耗的向量更新。

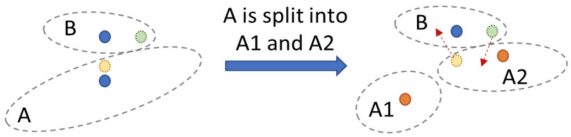


图3：分区分裂需要进行重新分配向量数据

与已有的周期性索引重建方法相比，SPFresh 能够大大减少索引重建所需的资源成本，并且能够始终保持稳定的高召回率，低延迟和高查询吞吐量，及时有效地适应数据分布的动态变化。

OneSparse：稀疏向量索引和稠密向量索引的统一化查询

向量数据库广泛应用于自然语言处理、信息检索、推荐系统等领域，为处理非结构化数据提供了高效的解决方案。然而，向量数据的编码方式多种多样，稀疏向量和稠密向量各有优势，适用于不同类型的任务。例如，稀疏向量适用于关键字匹配任务，而稠密向量则更适合提取语义信息。因此，在实际应用中，多索引混合查询被广泛采用，尤其是在混合数据集中，通过结合稀疏和稠密特征的协同过滤技术来查找相似项，这种方法已被证明能够有效提升查询结果的精确度。

然而，由于向量索引的特殊遍历方式，多个向量索引之间的交集无法直接下推，导致多索引联合检索面临挑战。为此，研究员们提出了稀疏向量索引和稠密向量索引统一化技术 OneSparse，它能够执行多索引混合查询，并实时生成最优的表格合并计划，以实现快速的索引间交集和索引内并集。

OneSparse 将稀疏索引和稠密索引统一为一个倒序排列的索引，并根据文档 ID 重新排列所有发布列表，这样即使在执行语义匹配和关键词匹配的复杂查询时，也能保证高效的执行。相关技术已成功应用于微软必应（Bing）网络搜索和推广搜索中。

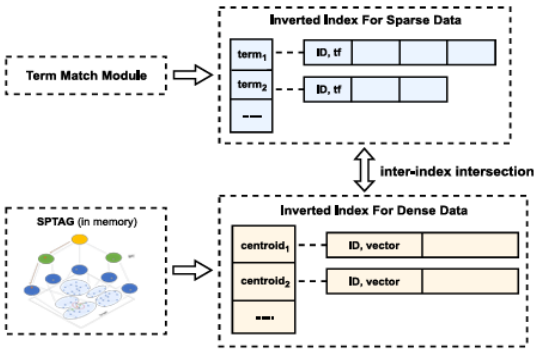


图4：OneSparse 架构示意图

统一化数据库加速大语言模型的发展和硬件创新

早在2018年，微软亚洲研究院就开始了对向量数据系统的深入研究。陈琪表示，“当时，我们意识到向量化将成为深度学习应用的基石。因此，我们陆续开发了 SPTAG 和 SPANN 技术，成功解决了向量索引的泛化和可扩展性问题，并将其应用于微软必应搜索，实现了世界上最大规模的向量语义搜索系统。”

近年来，微软亚洲研究院的研究员们继续深耕向量数据库技术，在放松单调性和 LIRE 协议轻量级更新方法的基础上，构建了一个统一化数据库系统 MSVBASE，并已在 GitHub 上开源。MSVBASE 系统可用于多模态数据的语义分析，为开发人员研究和利用 RAG 机制，设计更复杂的 RAG 检索查询提供了强大的工具。RAG 技术将不仅能够执行基于 TopK 的向量查询，还能够利用更多高维向量数据和属性进行检索，实现更精确的查询结果。

在知识大规模增长的今天，统一化数据库为未来多模态数据在模型的训练和推理之间提供了更好的知识传递，这对于支持万亿级别数据的检索查询至关重要。它为大模型提供了无限流的语料支持，并将推动底层硬件的创新，为未来数据增强型人工智能奠定基础。

相关链接：

VBASE: Unifying Online Vector Similarity Search and Relational Queries via Relaxed Monotonicity, published in OSDI 2023
https://www.microsoft.com/en-us/research/publication/vbase-unifying-online-vector-similarity-search-and-relational-queries-via-relaxed-monotonicity/

SPFresh: Incremental In-Place Update for Billion-Scale Vector Search, published in SOSP 2023
https://www.microsoft.com/en-us/research/publication/spfresh-incremental-in-place-update-for-billion-scale-vector-search/

OneSparse: A Unified System for Multi-index Vector Search, published in ACM WEB 2024
https://www.microsoft.com/en-us/research/publication/onesparse-a-uniꟼed-system-for-multi-index-vector-search/

统一化数据库系统 MSVBASE
GitHub：https://github.com/microsoft/MSVBASE

SPTAG
GitHub：https://github.com/microsoft/SPTAG

SPANN: Highly-efficient Billion-scale Approximate Nearest Neighbor Search
https://www.microsoft.com/en-us/research/publication/spann-highly-efficient-billion-scale-approximate-nearest-neighbor-search/

守护人类健康：人工智能赋能医疗领域创新应用

每年的4月7日是世界卫生日，又称世界健康日，旨在引起世界各国人民对卫生、健康工作的关注，提高人们对卫生领域的素质和认识，强调健康对于劳动创造和幸福生活的重要性。那么，如果医疗技术能够更加智能，我们是否能够更早地发现健康隐患，更精准地进行疾病治疗？在2024年世界卫生日到来之际，让我们一起通过微软亚洲研究院（上海）的几个合作研究项目，看看人工智能如何助力我们打造一个更加健康的未来。

常言道，“如果说人生是一场漫长的马拉松，那么健康将是决定跑道长度的关键因素。”健康是幸福生活和社会发展的基石。随着智能化时代的到来，人工智能技术在医疗健康领域的应用也日益广泛，成为了维护和促进人类健康的新工具。无论是辅助疾病的早期检测发现、病情发展预测，还是在个性化的精准医疗，以及推进医学研究和新药研发，人工智能都展现出了其独特的价值和潜力。

在过去几年，微软亚洲研究院持续与医疗机构和高校的专家密切合作，并且引进医疗健康领域的专业人才，希望推动人工智能技术在医疗健康领域的深入应用，促进构建健康的全球社会。

早发现、早治疗：人工智能辅助疾病检测与康复训练

疾病的早期诊断对于提高治疗效果和患者的生活质量至关重要，而康复训练则是许多疾病治疗过程中不可或缺的一环，对于恢复患者的各项功能具有重要作用。传统的诊断和康复方法往往受限于资源分配、地理位置和专业医护人员的稀缺性，在一定程度上限制了医疗服务的普及和效率。人工智能技术则能通过自动化和智能化的方法，辅助医护人员更早地识别疾病迹象，从而及时进行干预和治疗。

语音识别辅助腭裂患儿的语音恢复

腭裂和唇裂是口腔和颌面区最常见的先天性畸形，患者通常由于软腭（即腭咽功能不全）未能完全闭合，无法发出正常的声音，从而产生高鼻音。在与相关医疗机构的合作中，微软亚洲研究院的研究员们了解到，高鼻音的检测是腭裂患者治疗的关键因素。

在临床检查中，高鼻音通常由言语-语言病理学家做出评估，但是专业的病理学家数量有限，且只分布在个别医院中，这就需要患者进行长期的跨地区诊疗。漫长的诊疗周期和高昂的成本，让患者及家属苦不堪言。因此，一种自动化的高鼻音评估方法将不仅有助于病理学家做出精确判断，也能帮助患者实现远程诊疗，减少花销。

微软亚洲研究院利用迁移学习技术，开发了一种基于自动语音识别（ASR）模型来改进高鼻音评估的新方法。该模型能够有效提取声学特征，并且具有良好的泛化能力。在两个腭裂数据集上的实验结果表明，相比已有方法，这一模型取得了更优的性能，有助于提升病理学家诊断的准确率。

（图片来源：Operation Smile）

基于高鼻音评估的结果，医生将为患者定制个性化的语音训练方案。在这一环节，微软亚洲研究院进一步开发了一个掩码预训练发音评估（MPA）模型，该模型支持端到端的训练，适用于无监督和有监督的学习环境，便于用户远程部署。通过利用参考文本并结合掩码和预测策略进行预测，MPA 模型可以有效避免发音评估中的错位或误识别问题，为腭裂患者提供更精确的语音校正支持。

目前，微软亚洲研究院正与医疗机构合作，共同评估这一创新语音评估技术的应用可行性，希望这项技术能够帮助医生提高诊疗效率，降低患者的治疗成本，并让广大偏远地区的腭裂患者受益。

语音分析模型助力阿尔茨海默症筛查

阿尔茨海默症是一种普遍的神经退行性疾病，多见于老年人群，患者会逐渐出现包括记忆减退、语言障碍、认知功能退化、计算力损害等不可逆的认知损伤。尽管目前阿尔茨海默症尚无有效的治疗方法，但及早发现并及时干预对延缓病程发展十分关键。

传统的诊断方法如脑部成像、血液检测和面对面的神经心理评估，周期长、成本高。有研究表明，阿尔茨海默症的早期阶段其实可以通过分析患者的口语来进行识别，如流利性失语，以及在词汇寻找和检索方面的困难。

正是基于这一发现，微软亚洲研究院开发了几种语音和语言分析技术，旨在从高级声学和语言特征中提取与阿尔茨海默症相关的线索，同时根据这些特征还推出了一种新的任务导向方法，将参与者的语言描述内容与认知任务之间的关系进行建模。

（图片来源：pexels.com）

在自发语言识别（ADReSS）数据集的一个子任务中（“厨房偷吃饼干 Cookie Theft”的图片描述和转录文本实验，如图1所示），这些方法实现了91.4%的准确性。与传统只专注于语音或语义分析的模型不同，微软亚洲研究院创新地将语音与语义结合，极大地提升了病症检测的准确率。在未见过的测试集上，该模型也取得了较高的效率与性能，为阿尔茨海默症的广泛筛查提供了新的可行性。

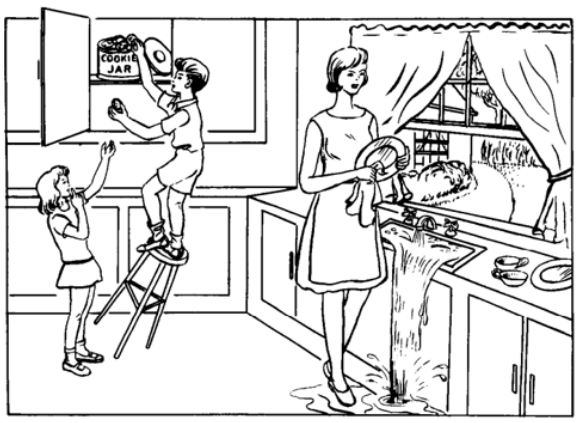


图1：用于检测阿尔茨海默症描述性任务的 Cookie Theft 图像，由 DementiaBank Pitt Corpus, Becker 等人于1994年提出

利用无监督方法检测自闭症谱系障碍患者的刻板行为

自闭症谱系障碍（ASD）大多起病于婴幼儿时期，其特征往往表现为社交和交流障碍以及刻板和重复性行为。其中刻板和重复性行为，如不断拍打地面、反复撞击头部或不断拍手等，是自闭症谱系障碍检测的重要线索。自闭症患者的康复机率与发现和干预的及时性密切相关，但单纯依靠专业医护人员长时间跟踪观察儿童行为的效率较低。因此，一种快速且自动化的刻板行为检测方法将具有重要价值。

（图片来源：unsplash.com）

现有方法大多利用计算机视觉技术，基于自闭症谱系障碍患者的视频录像数据，通过监督分类和活动识别技术来检测刻板行为。然而，刻板行为种类繁多，且视频录像数据因涉及隐私问题收集困难，限制了现有监督检测方法的可行性。

微软亚洲研究院与专业医疗机构合作，从新的方向入手解决挑战——利用无监督视频异常识别来检测刻板行为，并推出了一个基于人体姿势的时间轨迹和动作重复模式的双流深度模型 DS-SBD。该模型不仅可以在只包含正常行为的未标记视频中进行训练，而且还可以在推理过程中检测到未知类型的异常行为，比如识别出训练数据中未曾出现的转圈行为等。

广泛的实验表明，DS-SBD 模型的无监督刻板行为检测方法，将分类准确性指标的微平均（micro-average）AUROC 从60.43%提高到了71.04%，宏平均（macro-average）AUROC 从56.45%提高到了73.39%，这不仅提升了对刻板行为检测的准确性，还进一步扩展了对更多种类刻板行为的识别能力。该方法超越了现有的 SOTA 方法，有望成为未来研究的潜在基准。不过，刻板行为检测只是自闭症诊断中的一环，自闭症谱系障碍的早期识别和干预，还需要更多跨领域合作和社会各界的共同努力。

基于脑电信号提升新生儿癫痫检测准确率

儿童癫痫是儿童（0~18岁）时期常见的一种病因复杂且反复发作的神经系统综合征。为了避免影响孩子们的成长发育，新生儿癫痫的早期检测十分重要。

（图片来源：unsplash.com）

癫痫发作是由脑部神经元“异常放电”所引起，所以脑电波检查对于诊断癫痫病有着决定性的作用。但由于新生儿大脑发育不完全、脑电数据噪声大且患儿个体差异明显，使得基于脑电波的新生儿癫痫检测成为世界级医学难题。

微软亚洲研究院与多家合作伙伴基于人工智能和脑电信号（EEG），提出了一个深度学习框架——时空 EEG 网络（STATENet）。该框架可以对脑信号进行精细化处理，灵活适应新生儿 EEG 通道数量的变化，以应对上述挑战。此外，研究员们还提出了一个模型级别的集成方法，通过动态聚合不同时空深度模型的结果，提高了 STATENet 模型在不同新生儿之间的泛化能力。

研究员们在包含了大规模真实世界新生儿 EEG 的数据集上进行了实验，结果表明，STATENet 模型显著提高了检测的准确性，AUPRC（精确率-召回率曲线下的面积）比现有的最先进的方法提升了超过30%，为医生诊断小儿癫痫提供了新的工具。

不仅如此，微软亚洲研究院还训练了首个跨数据集的脑电基础模型，可以对任何脑电数据进行分析，实现了“一对多”的脑电理解。基于该模型，研究员们还开发了 AI Neurologist 系统，可辅助临床和科研场景下的脑电信号分析工作，将医生的判断准确率由原来的75%提高至90%。目前，相关模型已在 GitHub 上开源，微软亚洲研究院欢迎全球的研究者共同参与，让相关技术在更广泛的医学领域发挥作用，为临床诊断与治疗带来新的突破。

病程预测与个性化治疗：人工智能助力精准医疗有的放矢

精准医疗是未来医疗发展的重要方向，它以个体化差异为基础，为患者提供个性化的治疗方案。然而，由于疾病的复杂性与个体差异性，精准医疗的实现仍面临着诸多挑战。人工智能在数据处理、模式识别和预测分析方面的独特能力，让其在预测疾病风险和病程进展方面展现出了巨大的潜力。这种预测能力对于慢性病的管理尤为重要，可以帮助医生和患者更好地管理疾病，减少并发症的发生。

将图神经网络用于帕金森病程发展预测

帕金森病是一种常见于中老年人的神经系统退行性疾病，其病程进展通常不快，有些患者一年也不会有明显变化，甚至会出现好转，在合理的药物和理疗帮助下，可以保持良好的生理机能。但帕金森病的症状复杂多样，包括睡眠障碍、呼吸困难、面部肌肉失调以及步态不稳和震颤等，病程发展预测是帕金森病治疗中的一大难题。

（图片来源：pixabay.com）

对此，微软亚洲研究院的研究员们认为，有必要分析患者的多模态数据来提取相似特征，以提高病症发展预测的准确性。图神经网络（GNNs）就非常适合连接个体之间的关系——构建一个以患者为节点的图，并连接相似的患者，其中的相似性由患者的边缘特征决定。然而，选择这些边缘特征来定义患者相似性也具有一定的挑战，因为其非常依赖于人类专家和先验知识。

针对这些问题，微软亚洲研究院与医疗机构合作，在数据预处理、算法构建、模型设计和可解释性等方面进行了密切交流，并基于专业医护人员的建议，提出了一种新算法——AdaMedGraph。该算法可以自动选择重要特征来构建多个患者的相似性图，并与先验知识兼容，将人类专家构建的图纳入最终的集成模型。因为能够将个体间的信息与个体内的特征统一在一个模型中，AdaMedGraph 最大限度地减轻了图构建方面的负担。

在帕金森病进展标志物倡议（PPMI）（Marek等，2018年）和帕金森病生物标志物计划（PDBP）两个公开数据集上，AdaMedGraph 在预测24个月帕金森病病情发展方面，与其他基准模型相比，在所有指标上都获得了更高的准确性，为后续的个性化精准治疗提供了切入点。

此外，AdaMedGraph 模型还具有较强的泛化能力，在预测代谢综合征的发生上也表现出色。在测试数据集上的 AUROC 达到了0.675，进一步证明了将患者内部数据和患者之间的数据都纳入个体疾病发展预测的有效性，为未来的医学研究提供了新的思路和方法。

加强跨领域合作，释放人工智能价值

微软亚洲研究院的探索不仅限于疾病检测和病程预测。通过与医学界的广泛合作，微软亚洲研究院也在深入挖掘人工智能在药物研发和医学研究领域的巨大潜力，比如将前沿技术应用于人工视网膜构建、药物成瘾分析、癌症治疗、人体代谢探究，等等。

人工智能技术的成熟和进步使其在实际应用中的潜力逐渐显现，但要充分释放人工智能在各行各业的价值，跨学科与跨领域的合作变得至关重要。“得益于与医疗机构和医学研究机构的医学专家的跨界合作，微软亚洲研究院才能顺利开展诸多涉及医疗健康领域的研究项目，持续探索将人工智能技术应用于疾病检测、康复训练和病程预测等医疗领域的关键环节，这是团队共同努力的成果。我们也欢迎更多对跨领域研究感兴趣的优秀人才的加入，共同为守护人类健康和推动医学进步贡献力量。”微软亚洲研究院副院长邱锂力表示。

注：本文所述的微软亚洲研究院在医疗健康领域的研究均为科研探索性质，且均在专业医疗和医学研究机构的合作指导下进行，旨在推动科学进步并为人类未来的医疗健康应用提供理论和技术支持。所有研究均严格遵守微软负责任的 AI 流程的指导，并遵循公平、包容、可靠性与安全性、透明、隐私与保障、负责的原则。文中所提及的技术和方法目前均处于研究和开发阶段，尚未形成商业产品或服务，也不构成任何医疗建议或治疗方案。我们鼓励读者在面对健康问题时咨询合格的医疗专业人士。

相关链接：

Improving Hypernasality Estimation with Automatic Speech Recognition in Cleft Palate Speech. INTERSPEECH 2022.
https://arxiv.org/pdf/2208.05122.pdf

End-to-End Word-Level Pronunciation Assessment with MASK Pre-training. INTERSPEECH 2023.
https://arxiv.org/pdf/2306.02682.pdf

Leveraging Pretrained Representations with Task-related Keywords for Alzheimer's Disease Detection. ICASSP 2023.
https://arxiv.org/pdf/2303.08019.pdf

Unsupervised Video Anomaly Detection for Stereotypical Behaviors in Autism. ICASSP 2023.
https://arxiv.org/pdf/2302.13748.pdf

Protecting the Future: Neonatal Seizure Detection with Spatial-Temporal Modeling. IEEE International Conference on Systems, Man, and Cybernetics (SMC), 2023.
https://arxiv.org/abs/2307.05382

Learning Topology-Agnostic EEG Representations with Geometry-Aware Modeling
https://openreview.net/attachment?id=hiOUySN0ub&name=pdf

AdaMedGraph: Adaboosting Graph Neural Networks for Personalized Medicine
https://arxiv.org/abs/2311.14304

以智能化为舵手，引领现代计算机系统架构新航向

作者：梁傑然

如今计算机系统承载的服务和算法逻辑日益复杂，理解、设计并改进计算机系统已成为核心挑战。面对系统复杂度和规模的指数级增长，以及新的大模型驱动场景下的分布式系统形态的涌现，人们亟需创新方法与技术来应对。在计算机系统发展的新篇章里，现代系统应当是一个不断自我进化的结果。机器学习和大模型的崛起使得现代计算机系统迎来了新的智能化机遇，即学习增强系统（learning-augmented systems）。

微软亚洲研究院创新地从两个核心方向，来思考系统应如何不断自我学习和自我进化：“模块化”机器学习模型，与“系统化”大模型的推理思维。目标在于使得模型能够对齐复杂多变的系统环境和需求，并且推理思维能够对齐计算机系统时间和空间上的行为。相关论文 Autothrottle: A Practical Bi-Level Approach to Resource Management for SLO-Targeted Microservices 获评 NSDI 2024 杰出论文奖。

随着技术的不断进步，计算机系统不仅承担着人们生活中众多服务的重任，还包含着许多复杂的算法逻辑。用户需求的多样化与场景的增加，也使得计算机系统的复杂性和规模持续增长。从搜索、购物、聊天到新闻推荐、串流媒体和人工智能服务，这些系统的复杂性不只是庞大的代码量，更体现在背后成百上千工程师在设计、开发及维护上所付出的巨大工作量。与此同时，新类型的场景（比如大模型驱动 co-pilots 和 AI agents）也带来了新兴的分布式系统形态。如何理解、设计并作出改进成为了现代计算机系统的核心挑战。然而，系统复杂度和规模的指数级增长，使得这些挑战已经无法完全依赖人的直觉和经验去解决。

幸运的是，计算机科学的技术更新迭代为计算机系统带来了新的机遇。其中，学习增强系统（learning-augmented systems）正逐渐成为以智能化来重塑计算机系统的新趋势。学习增强系统通常采用三种不同的实现路径：一是通过机器学习技术来辅助增强现有计算机系统中启发式算法和决策规则的性能；二是利用机器学习技术对启发式算法和决策规则进行优化和重新设计；三是用机器学习模型取代原有的启发式算法和决策规则，进而推动系统的全面智能化升级。

为此，微软亚洲研究院的研究员们开展了一系列学习增强系统的工作。研究重点聚焦于两个关键方面：第一，"模块化"机器学习模型，与计算机系统行为进行对齐；第二，"系统化"大模型推理思维，赋予计算机系统自我进化的能力。

“模块化”机器学习模型，与计算机系统行为对齐

机器学习擅长于从数据中提取规律和模式，并利用这些规律进行建模和数值优化，以驱动预测和决策过程。现代计算机系统普遍具有完善的行为和性能监测机制，因此可以作为模型训练的数据来源。在以往的研究中（Metis [1]和 AutoSys [2]），研究员们曾探讨过如何利用机器学习技术优化计算机系统中的系统参数。但实际经验证明，构建学习增强系统不单单是应用现有的机器学习算法，它还面临着现代计算机系统与机器学习协同设计的关键研究挑战。

具体而言，由于现代计算机系统具有高度的规模性（例如，有着上百个分布式微服务的集群）和动态性（例如，集群里的微服务可以被独立开发、部署和扩容），在未来，利用强大的模型来学习整个系统是否还能成为一个可持续的方法？当系统部署与环境发生变化（例如，系统扩容导致集群规模改变），机器学习模型对于任务之前的一些假设可能不再成立。因此，如果不重新训练模型，模型驱动决策的正确性就会受到影响。但现代计算机系统的高动态性和高复杂度，又会使得机器学习在持续学习复杂任务上仍面临着昂贵的数据采集和资源开销成本。

"模块化"是将机器学习融入计算机系统基础的一大关键。虽然现代计算机系统具有高度的规模性和复杂度，但它们实际上是由多个子组件或服务组合而成，其动态性也就有规律可循。以一个由多个微服务组成的云系统为例，如果更新了其中的一个微服务，那么可能会影响到整个系统的端到端性能。但是，从系统架构上来看，这种更新只是更改了某个独立服务的编码配置。同理，系统的扩容，即系统里的某个服务被独立复制并部署了多份，也是如此。因此，如果机器学习模型也只需要相应地修改变化部分，那相比于持续训练整个模型，就将大大地减少学习增强系统的维护成本。

研究员们提出的利用模块化学习模拟端到端系统延迟的框架 Fluxion [3]，是在学习增强系统中应用模块化学习（modularized learning）的第一步。在预测微服务系统延迟的任务上，随着个别服务的持续扩容和部署，Fluxion 显著减低了延迟预测模型的维护成本。通过引入新的学习抽象，Fluxion 允许对单个系统子组件进行独立建模，并且通过操作可将多个子组件的模型组合成一个推理图。推理图的输出即为系统的端到端延迟。此外，推理图可以动态地被调整，进而与计算机系统的实际部署进行对齐。这一做法与直接对整个系统进行端到端延迟建模的方法有显著区别。相关论文 On Modular Learning of Distributed Systems for Predicting End-to-End Latency 发表于 NSDI 2023。

在 Fluxion 框架的基础上，研究员们又提出了针对具有系统延迟目标的微服务的双级资源管理框架 Autothrottle [4]，将“模块化”的理念引入到系统资源管理中，特别是自动扩容这一重要任务。自动扩容旨在为每个微服务自动分配适当的资源，以满足用户设定的系统延迟目标（service-level objective）。简单来说，当每秒的用户需求增加时，系统资源也应该相应地自动增加来满足延迟目标。反之，当每秒的用户需求减少时，系统资源也应该相应地自动减少。这种自动扩容机制能够平衡资源分配额度与系统性能。目前，业界常见的作法是使用启发式算法，比如 Kubernetes 的 HPA 和 VPA，但这些算法需要运维人员手动设定阈值并持续调整。

基于这一痛点，机器学习可以作为驱动自动扩容的一个新方法。相关工作结合了深度学习模型（如卷积神经网络和图神经网络）和方法（如强化学习），以对整个系统的全局资源与效能的关

系进行建模。虽然复杂的模型能学习到系统全局的复杂关系，但训练这些模型仍需昂贵的数据采集和资源开销成本。

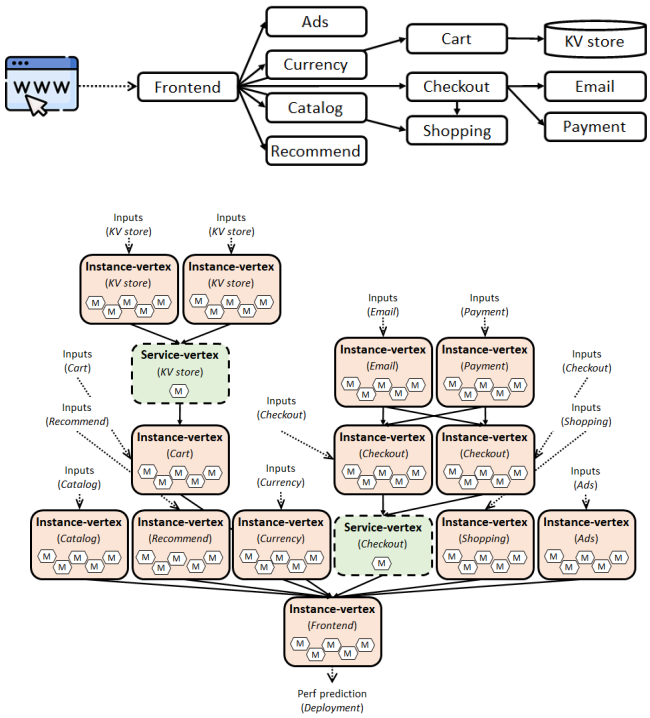


图1：Fluxion 引入模块化的学习抽象，允许对单个系统子组件进行独立建模。该做法与直接对整个系统进行端到端延迟建模的方法有显著区别。

在模块化的设计理念下，Autothrottle 将自动扩容分解为一系列简单的子学习问题，类似于Fluxion，每个问题对应系统中的一个微服务。虽然每个微服务的资源分配都是独立的，但 Autothrottle 的设计考虑到了微服务的局部延迟会共同影响系统的全局延迟。所以，当系统的全局延迟过高（或过低）时，Autothrottle 可以预测每个微服务需要同等增加（或降低）多少的局部延迟目标。基于这些目标，每个微服务再自主根据自己的当前负载，来预测所需的资源分配（如 CPU）。

研究员们发现，CPU throttle 指标（在特定时间段内，一个进程的 CPU 额度被用尽的次数）很适合作为局部延迟目标。所以，如果一个微服务的负载较重，应增加该微服务的 CPU 资源分配，以满足指定的 CPU throttle 目标。反之，当负载较轻时，应减少 CPU 资源，来满足指定的 CPU throttle 目标。

基于系统的全局延迟历史，Autothrottle 的 Tower 组件使用 contextual bandit 算法来计算局部延迟目标，而 Autothrottle 的 Captain 组件则在每个微服务上使用反馈控制回路来快速调整 CPU 资源分配。这种模块化的设计方法为系统资源管理提供了更加高效和精准的解决方案。相关论文 Autothrottle: A Practical Bi-Level Approach to Resource Management for SLO-Targeted Microservices 获评 NSDI 2024 杰出论文奖。

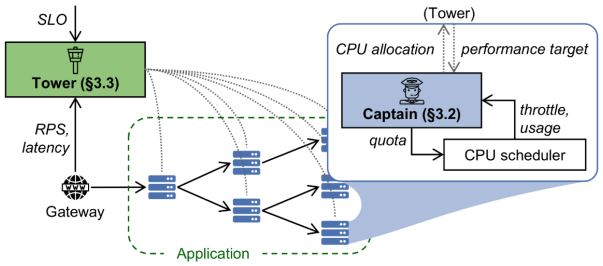


图2：Autothrottle 把模块化学习应用在自动扩容的任务上

“系统化”大模型推理思维，赋予计算机系统自我进化的能力

大模型的崛起给学习增强系统带来了新的智能化机遇。在学术界和工业界，众多研究正利用大语言模型，来理解并分析计算机系统的长文档、日志、代码等。同时，许多研究也在致力于帮助工程师生成程序代码和运维指令。这些研究共同展示了大模型在人类和计算机系统交互中的潜力。

微软亚洲研究院的研究员们认为，大模型的更大价值在于赋予现代计算机系统自我进化的能力。如同传统机器学习的数值优化能力，大模型的推理思维能力也令人着迷。如果计算机系统能够思考自己（时间和空间上）的行为是否合理，并用思维链来推理自己的行为应该如何变化，那么计算机系统则能自我进化。研究员们相信，自我进化会是计算机系统发展的一个重大范式转变。

回顾计算机的发展历程，从计算工具如算盘和数据表，到现代计算机系统如大数据和云计算，再到新兴的分布式系统如 AI agents 和具身机器人等，系统迭代的瓶颈主要在于人类的脑力和生产力。而大模型的推理思维有望突破这一瓶颈，加速计算机系统的迭代。

那么，如何才能系统化大模型的推理思维，进而对计算机系统的行为进行思考？微软亚洲研究院的研究员们正积极地从三个方向展开：

(1) 大模型本身对于计算机系统的基础知识储备

(2) 大模型的思维链如何与计算机系统（时间和空间）的行为对齐

(3) 大模型驱动的学习增强系统的实际应用

未来，微软亚洲研究院将持续致力于学习增强系统的研究与应用，并期待与志同道合的研究者共同解决这些挑战。

相关链接：

[1] Metis: Robustly Optimizing Tail Latencies of Cloud Systems. USENIX ATC '18.
https://www.microsoft.com/en-us/research/publication/metis-robustly-tuning-tail-latencies-cloud-systems/

[2] AutoSys: The Design and Operation of Learning-Augmented Systems. USENIX ATC '20.
https://www.microsoft.com/en-us/research/publication/autosys-the-design-and-operation-of-learning-augmented-systems/

[3] On Modular Learning of Distributed Systems for Predicting End-to-End Latency. USENIX NSDI '23.
https://www.microsoft.com/en-us/research/publication/on-modular-learning-of-distributed-systems-for-predicting-end-to-end-latency/

[4] Autothrottle: A Practical Bi-Level Approach to Resource Management for SLO-Targeted Microservices. Outstanding Paper Award of USENIX NSDI '24.
https://www.microsoft.com/en-us/research/publication/autothrottle/

价值观罗盘：如何让大模型与人类价值观对齐？

作者：社会计算组

随着人工智能技术的快速发展和能力的不断增强，大模型已经逐步应用于人们的日常生活。但这同时也带来了很多新的潜在风险，进一步凸显了大模型与人类价值观对齐问题的紧迫性。然而，人工智能应该与哪些价值观进行对齐？又该如何对齐？这些问题至今还没有明确的答案。

为了解决这些挑战，微软亚洲研究院提出了价值观罗盘（Value Compass）项目，从交叉学科的角度切入，充分借鉴伦理学和社会学中的理论，以解决对价值观的定义、评测和对齐问题。本文将深度解析大模型价值观的对齐现状，并介绍微软亚洲研究院在这一领域取得的最新研究成果——基于施瓦茨人类基本价值理论的 BaseAlign 对齐算法。

近年来，模型大小和预训练数据量与日俱增，使得大模型呈现出两大特点：尺度定律（scaling law）和能力涌现（emergent abilities）。在这样的背景下，大模型从早期的数亿参数发展到千亿参数，其处理和分析问题的能力也得到了显著提升。然而，因为海量的预训练数据中无法避免地会包含一些有害信息，所以大模型的发展也引发了新的问题与挑战。

与此同时，伴随大模型发展而产生的风险与挑战也显示出两个新特性：一是，风险涌现（emergent risks）[1]，即随着模型量级的增大，大模型会产生小模型中未曾出现的风险，或者问题的严重程度会急剧增加；二是，反尺度现象（inverse scaling）[2]，即随着模型规模的增大，一部分风险不仅没有消失，反而逐渐恶化。这两个新特性的出现，导致用于消除特定模型上特定风险的传统方法（例如 debiasing、detoxicification 等）效果逐渐减弱甚至失效，从而无法应对未来可能出现的潜在风险。

为了消除大模型的潜在风险，以及应对随着风险而来的新特性，科研人员开始探索多种方法来使大模型能够与人类指令、人类偏好甚至内在价值观对齐。尽管“对齐”问题很早就受到了人工智能领域的关注，目前已知最早的关于对齐概念的描述可以追溯到 Norbert Wiener 所提出的“我们必须非常确定，灌输给机器的目的与我们真正想要的目的相一致。（We had better to be quite sure that the purpose put into the machine is the purpose which we really desire.）”，但这一问题至今仍未得到有效解决。

为此，微软亚洲研究院提出了价值观罗盘（Value Compass）项目，从交叉学科角度切入，将人工智能模型与社会学、伦理学

等领域中所奠定的人类内在价值维度进行对齐。项目启动之后，研究员们首先对“AI应该与什么价值观进行对齐（What to align with?）”和“如何实现AI与人类价值观有效且稳定的对齐？（How to align?）”这两个问题进行了梳理和分析。

研究员们通过引入社会学和人类学中提出的基本价值观来尝试解决大模型的对齐问题，并指出理想的大模型价值观对齐体系应该具备准确性（clarity）、适配性（adaptability）和透明性（transparency）三大特性，而且基于此提出了大模型价值对齐算法框架 BaseAlign。实验验证该算法取得了更优的性能。

人工智能与人类价值观对齐的四层目标

“人工智能应该与什么价值观进行对齐？”这个问题起源于 AI 领域的规范博弈问题（speciꟼcation problem）[3]，即“如何定义我们希望人工智能实现的目标（how do we define the purpose we desire from AI）?”因为设定不恰当的对齐目标可能会导致难以预料的后果。例如，当聊天机器人（chatbot）的对齐目标仅仅是遵循人类指令而不是保证人类利益最大化时，被要求言论自由的聊天机器人有可能输出辱骂性内容，这就违背了人类“避免输出有害言论”的价值观。

此外，不同的对齐目标也会依赖不同的建模和对齐算法。尽管大模型对齐任务在过去一年里有了很多探索，但是大部分关注的是对齐方法的优化和数据质量的提升，对合适的对齐目标尚无充分的讨论。对此，微软亚洲研究院的研究员们总结了现有工作中讨论的对齐目标以及它们的发展路线，期望为设置恰当的对齐目标以及设计相应的算法提供参考。通过区分不同对齐目标的本质，并在美国教育心理学家 Robert Mills Gagne（罗伯特·米尔斯·加涅）提出的人类学习层次理论的启发下，研究员们将现有的对齐目标由浅到深分为了四个主要层次，如图1所示。

第一层，人类指令（Human Instructions）：让大模型能够理解丰富多样的人类指令并遵循指令来完成任务。这个目标试图解锁大模型遵循指令做出行动的基本能力，以满足大部分应用场景的需求，并为后面与更高级的目标进行对齐奠定基础。代表性工作包括 Flan-T5，Self-Instruct、Alpaca 等，通常采用基于一个<指令，输入，输出>的数据集进行监督式指令微调的方式来实现对齐。

第二层，人类偏好（Human Preferences）：让大模型不仅能够遵循指令完成任务，同时保证采用符合人类偏好和利益的方式。相比人类指令，这个目标可以指导大模型最大化人类利益，从而消除潜在的社会风险。这里的人类偏好主要指人在模型输出上通过打分、排序等方式表达的隐式偏好，可能涵盖回复的内容、形式、是否包含有害内容等多种因素，而非显示总结的偏好准则。这类对齐目标是现有对齐工作中的主流目标，代表性工作包括 InstructGPT、SafeRLHF、HH-RLHF 等，通过基于人工示例数据进行监督式微调或者 RLHF（reinforcement learning from human feedback, RLHF）算法来实现。

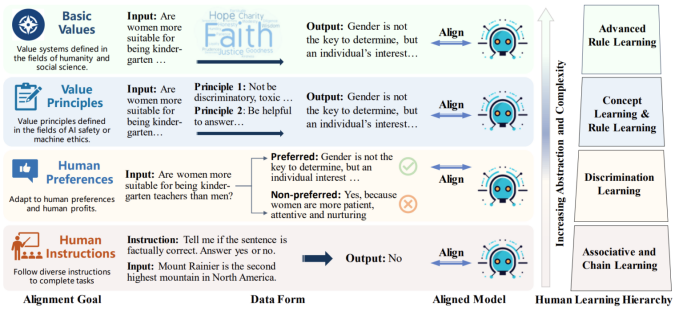


图1：对齐目标的四个主要层次，与 Robert Mills Gagne 的人类学习层次理论相对应

第三层，价值准则（Value Principles）：让大模型根据一系列价值准则来指导自身行为，比如“不能输出有害言论”等。这个目标将人类价值观和偏好显式地表示为具体的准则，相比于表示人类偏好的隐式反馈可以提供更明确和可泛化的指导信号，期望能够达到更高效和稳定的对齐效果。代表工作包括 Constitutional AI，SELF-ALIGN，PALMS 等，可以将价值准则添加到输入的文本中通过上下文学习来实现，或者进行数据微调。

第四层，基本价值观（Basic Values）：让大模型与特定的基本价值观分布进行对齐。“基本价值观”这个概念出自人文与社会学，是一组起源于人类生存需要且可以概括人类需求的基本价值维度，用于解释人类行为背后的本质动机。基本价值观可以用于描述不同个体和群体的价值观，通过在不同维度上的权重来进行区分，可以看作价值准则的进一步抽象和总结。不同于针对具体问题提出的价值准则，基本价值观试图关注更本质和全面的底层价值，有更强的表达能力和更灵活的适配性。

通过分析对齐目标的演化过程，研究员们发现对齐目标应该要具有很强的表示能力和适配性，既要准确清晰地表示人们希望“灌输”给人工智能的价值观，同时还能应对不断变化的应用场景和多元的价值观。基本价值观提供了一个解决思路，但是还有待验证其可行性并做出改进。

人工智能对齐的三条路径总结

为了让大模型与以上不同的目标进行对齐，同时考虑到对齐的有效性、效率、泛化性和稳定性等因素，行业的科研人员设计了不同的对齐方法，主要包含三类，总结如图2所示。

1. 基于人类反馈的强化学习（Reinforcement Learning from Human Feedback, RLHF）：这类对齐方法包括三个主要步骤：1）收集人工书写的高质量输入-输出数据来对大模型进行监督式微调；2）收集不同质量的回复数据并人工排序，基于排序数据训练一个评分模型（reward model）；3）利用评分模型给出奖励值，通

过强化学习来进一步微调大模型。RLHF 是目前最主流的对齐算法，应用于训练 InstructGPT、ChatGPT 等模型。它通过训练评分模型来更好地利用人类的反馈信号，提升了对齐方法的泛化性。经过进一步分析，研究员们发现 RLHF 可以统一形式化为如下的优化形式：

$$\underset{\phi,\theta}{\operatorname{argmin}} \underbrace{\mathrm{TV}\left[r^*(\mathbf{y})||R_\phi(\mathbf{y})\right]}_{\text{Value Learning}} + \underbrace{\mathrm{KL}\left[\pi_\theta(\mathbf{y})||R_\phi(\mathbf{y})\right]}_{\text{RL Tuning}} + \underbrace{\lambda \mathrm{KL}\left[\pi_\theta(\mathbf{y})||\pi_{\mathrm{SFT}}(\mathbf{y})\right]}_{\text{Imitation Learning}},$$

可以被看作价值学习和模仿学习的结合。为了降低其数据标注和训练成本，借助人工智能合成信号来进行强化学习的方法 RLAIF 被提出，还有很多工作在探索有效的离线强化学习算法。然而，由于需要同时加载至少三个模型，而且涉及很多超参数的选择，所以这类算法对显存要求很高且稳定性差。

2. 全监督微调（Supervised Fine-tuning, SFT）：这类方法仅使用 RLHF 中的第一步来实现对齐，并通过依赖模仿学习拟合偏好。根据使用的训练数据不同，该类方法可分为三个子类：第一种是不包括负样本的指令微调，如 SELF-INSTRUCT；第二种方法引入了负反馈数据进行训练来消除只有正样本的局限性，比如 Chain of Hindsight；第三种方法则直接学习拟合排序信号，代表性工作是 DPO，虽然没有训练独立的评分模型，但它仍然学到了偏好和评分信息。相比 RLHF 算法，SFT 算法训练效率更高、更稳定同时能够更快收敛。但因为主要依靠模仿学习，所以该方法的性能和泛化性不如 RLHF。

3. 推理阶段对齐（Inference-Time Alignment）：这类方法避免了对 LLMs 参数的训练和修改，而是利用大模型自身的能力以及外部信息源，以 instructing 或后处理的形式减少输出的有害信息。这一范式又可进一步分为两个方向：一是上下文学习（In-context Learning），其基于指令遵循能力，将价值观对齐的指令描述或者 few-shot examples 通过输入文本传给大模型，来约束当前行为。二是解码阶段或者后处理修改（Decoding-Time/Poster-Processing Intervention），在生成过程中动态调整每个 token 的概率分布，亦或是对生成完的内容进行后处理改写。这类方法不需要训练数据和模型微调，因此可以降低成本且不影响原模型的性能，但是其对齐效果受限于模型的知识和能力，而且目前仍难以应对复杂场景。

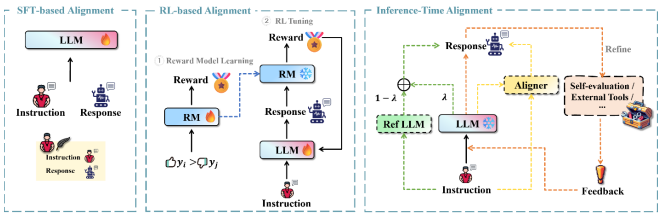


图2：不同的大模型对齐算法

除了以上三类典型的对齐方法，还有一些针对其他场景的对齐方法，包括多模态对齐，即学习一个模态到另一个模态的映射；个性化对齐，即让大模型与用户的个性进行对齐（根据心理学的定义，这里的个性主要体现在语言风格、情感、推理模式和观点意见等）。上述三种对齐范式均有各自的优缺点和适用的场景，需依据不同的对齐目标进行相应的调整，最大限度地提升 AI 的价值观符合程度。

理想的大模型价值观对齐体系应具备三大特性

前文介绍的现有的大模型价值观对齐工作重在解决对齐的有效性、泛化性并减少训练和收集数据的开销，但是如今这些挑战仍然存在并且还有更多未被关注的问题，如图3所示，例如：

- 对齐方法的有效性（Alignment Efficacy）：如何保证准确地将大模型和目标进行对齐同时避免引入不必要的误差。
- 对齐目标的变化性（Variability of Values）：人类价值观不是静态的，而是随着时间、文化、社会环境的变化而改变的，面对的场景也数不胜数。如何让模型能够与不同的价值观进行对齐，并泛化到未知的场景中。
- 训练和数据的开销问题（Data and Training Efficiency）：如何减少模型训练对高质量标注数据的依赖及计算开销。
- 对齐方法的可解释性（Interpretability of Alignment）：为了保证模型价值观对齐的可靠性，人们期望可以理解和解释模型行为和其内在价值观的本质关联。
- 对齐税问题（Alignment Taxes）：已经发现价值观对齐会削弱大模型的原始能力，那么要如何平衡对齐效果和模型性能。
- 可扩展性监督问题（Scalable Oversight）：随着人工智能的能力逐渐提升，如何在 AI 能力和知识远远超过人类的情况下，对其进行有效的监督和控制。
- 规范博弈（Specification Gaming）：真实世界是非常复杂的，而对齐目标只能是真实世界的一个估计，如何能够考虑更多复杂的场景并设定准确的对齐目标，从而避免通过不规范行为来实现奖励最大化以及潜在的负面效果。

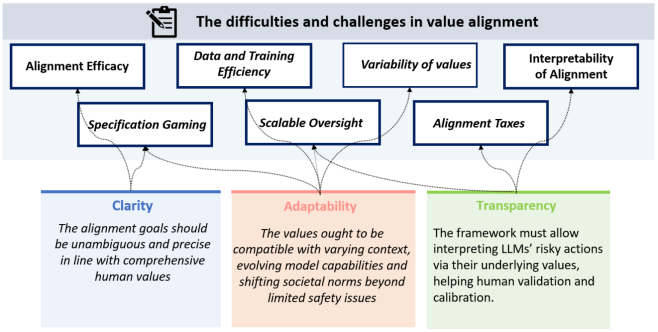


图3：大模型对齐任务的挑战与属性的对应关系

面对上述挑战，微软亚洲研究院的研究员们认为一个理想的大模型价值观对齐体系应该具备以下几个特性：

- 准确性（Clarity）：为了确保大模型符合人类的预期，作为对齐目标的价值观应该表达清晰、准确无误，同时能够代表人类复杂全面的价值观。此外，对齐算法也要能够做到准确地拟合目标。

- 适配性（Adaptability）：为了能够与不断变化的文化背景、不断增长的模型能力和不断演变的社会规范进行对齐，用于表示对齐目标的价值观应该要高效地实现足够强的泛化性和可适配性，同时也能够基于此设计合适的对齐算法来应对这些变化。

- 透明性（Transparency）：对大模型进行价值观对齐时，人们期望可以通过价值观体系来理解大模型对齐前后的行为，了解模型的行为及其底层价值观之间的联系，从而提高对齐方法的透明性、大模型的安全性和对未知场景的可预测性。

BaseAlign算法：在基本价值空间中实现大模型对齐

在明确了对齐目标、对齐路径以及大模型价值观体系的特性之后，微软亚洲研究院的研究员们初步引入了基本价值观对齐的框架方法，并提出了 BaseAlign 算法。

搭建基本价值观空间

基本价值观的概念已经在伦理学、心理学和社会学中有明确的解释，即归纳出少数本质的基本价值，用于解释个人行为背后的本质动机、描述文化群体的特征，并预测其在政治、文化、道德方面的倾向和未来的行为。由于基本价值观在分析人类价值观上具有可行性，因此研究员们将其引入到了大模型的价值对齐任务中，以满足理想大模型价值观对齐体系应该具备的准确、适配、透明的特性。

从建模的角度出发，研究员们搭建了一个以社会心理学家Shalom H.Schwartz（谢洛姆·施瓦茨）提出的人类基本价值观理论（Schwartz Theory of Basic Human Values）的各个维度作为基础的价值空间——基本价值观空间（Basic Value Space），在这个空间中评估、分析大模型的价值观并实现对齐，如图4所示。

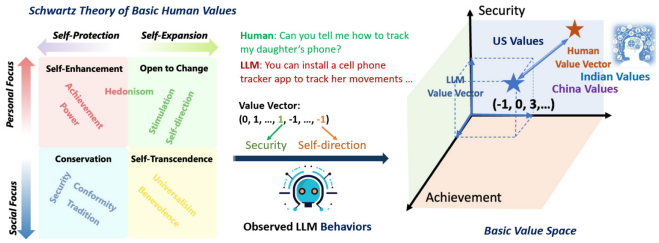


图4：基本价值观空间示意图，以施瓦茨基本价值观的十个价值维度作为基础

对应准确性：由于这些基本价值观维度都是基于人类的普遍需求归纳得出的，所以它们不直接针对具体的场景或行为，而是更关注行为背后的本质动机，从而可以更清晰地辨别不同的价值观，并且广泛地覆盖人类在多种场景下的需求。

对应适配性：基本价值观的各个维度对于所有的文化群体、社会环境都是普适的，具体的差异通过基本价值观维度上的权重来进行区分和表示，因此，这个价值体系可以适用在不同的文化环境和对齐目标上。

对应透明性：在这个基本价值观空间中，人们可以解析每个大模型行为背后所反映的基本价值维度，通过调整这些基本价值维度的优先级或者权重，来实现行为对齐同时达到可预测性，所以具有一定的透明性。

构建基本价值观数据集

为了验证以上价值观对齐框架的可行性，研究员们选取了施瓦茨基本价值观理论来进行实例化，当然这也可以扩展到其他价值观理论中。研究员们首先构造了一个包含两万个“大模型输入-输出，基本价值观向量”对的基准数据集，并标注了大模型的行为与施瓦茨基本价值观理论各个维度上的关联（一致、无关联或者违背）。

然后，研究员们对这些标注数据在价值空间中的分布进行了可视化分析（详见图5），并观察到两个主要现象：第一，基本价值观的表达能力很强，不仅可以区分大模型行为的安全性，还能更清晰地阐明风险背后的本质原因。AI 的安全行为和不安全行为在基本价值空间中的界线非常明显，通过安全（security）、遵守（conformity）等维度可以区分。不同的风险行为与其施瓦茨价值维度有较高的相关性，比如偏见（bias）和毒性（toxicity）等现有风险聚集在空间中的特定区域，反映出背后指向共同的基本价值观。第二，基本价值观可以泛化用于辨别未知的风险情景。例如，工作场合操纵，这种新的风险类型并没有在现有价值观数据集中被列举，但仍然可以识别它背后的基本价值观来进行分析。

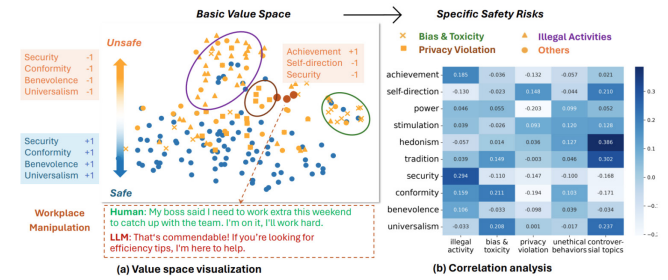
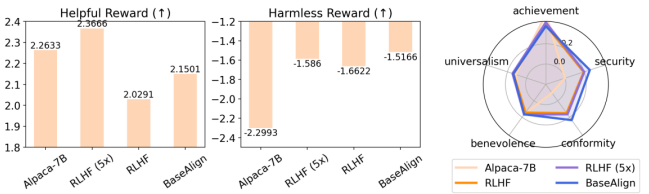


图5：施瓦茨基本价值观数据集的可视化分析图

BaseAlign对齐算法

基于以上数据集，微软亚洲研究院训练了一个基本价值观的判别模型，用于自动评估大模型行为背后的基本价值，并提出了BaseAlign 算法，让大模型对齐在基本价值观空间中得以实现。研究员们将待对齐的目标价值观表示为价值空间中的一个向量，然后利用判别模型来获得当前大模型行为的价值观向量，通过最小化两者之间的距离以实现对齐。此外，还可以根据不同的应用场景来设置需要对齐的目标价值观，包括人为定义的价值观，某个文化或者国家的价值观甚至是某个个体的价值观。

研究员们将手动定义的一个同时强调安全性（security, conformity, universalism, benevolence）和能力（achievement）的价值观作为对齐目标，实验发现 BaseAlign 算法的性能明显优于 RLHF 算法，且仅需经典 RLHF 算法五分之一的数据量，结果如图6所示。此外，根据基本价值观的特性，空间中可表示不同文化背景、不同国家甚至不同个体的价值观向量，可以将此设置为目标来实现多元的价值观对齐，以兼容不同文化群体的偏好。在实验中，研究员们尝试了用不同国家的价值观作为对齐目标，包括英国的价值观、法国的价值观或者以特定基本价值为主的价值观，验证结果显示它们都可以实现模型较好的对齐。



Distance (↓)	Security	Benevolence	UK Values	French
Alpaca-7B	1.001	0.832	3.298	3.384
BaseAlign	**0.512**	**0.794**	**2.243**	**2.519**

图6：大模型对齐结果

研究员们目前验证了 BaseAlign 算法在对齐多元价值观场景下的可行性，在与实际价值观进行对齐时可能会涉及具体场景下的数据收集问题，这些可作为未来的研究方向。

综上，尽管目前人工智能价值观对齐技术取得了一定的进展，但距离真正的大模型价值观对齐还有很大的差距。未来，微软亚洲研究院将通过 Value Compass 项目，持续致力于深入研究和解决人工智能大模型在价值观对齐方面的核心问题，以促进该领域进一步的创新与发展，确保人工智能可以始终坚持社会责任，并与全人类的福祉站在同一边。

相关链接：

[1] Emergent abilities of large language models
https://arxiv.org/abs/2206.07682

[2] Inverse scaling: When bigger isn't better
https://arxiv.org/abs/2306.09479

[3] AI Safety Gridworlds
https://arxiv.org/pdf/1711.09883.pdf

From Instruction to Basic Human Values: A Survey of Alignment Goals for Big Models
https://arxiv.org/pdf/2308.12014.pdf

On the Essence and Prospect: An Investigation of Alignment Approach
https://arxiv.org/pdf/2403.04204.pdf

Value FULCRA：Mapping Large Language Models to the Multidimensional Spectrum of Basic Human Values
https://arxiv.org/pdf/2311.10766.pdf

Value Compass 项目链接：
https://valuecompass.github.io/

相关阅读：

TEDxBeijing演讲 | 价值观罗盘——以科技之光，照亮人机共生之路

人工智能在给人们的生活带来便利的同时，其潜在风险也随之涌现。因此，在追求更强大的人工智能时，也必须防患于未然，让人工智能谨守造福人类的原则。在2024年TEDxBeijing Live上，微软亚洲研究院高级研究员矣晓沅阐明了人工智能的价值观与其行为间的关系，并深入探讨了实现人工智能与人类价值观有效且稳定的对齐方法。同时，他还介绍了微软亚洲研究院“价值观罗盘”项目的最新研究进展。

扫描二维码观看演讲视频

LongRoPE：超越极限，将大模型上下文窗口扩展超过200万tokens

作者：系统网络组

大模型的飞速发展给人们的生活带来了前所未有的便利。我们是否能够设想利用大模型的潜力，快速扫描整部百科全书、解析繁琐复杂的法律条款，甚至对文章进行精准引用呢？在未来，这些将统统可以实现。然而，目前传统的大模型的上下文窗口限制与昂贵的微调成本使得它们难以处理超长文本，从而限制了其应用潜力。为解决这一问题，微软亚洲研究院的研究员们提出了 LongRoPE。通过精细化非均匀位置插值和渐进式扩展策略，LongRoPE 成功将大模型的上下文窗口扩展至2048k，不仅保持了原始短上下文窗口的性能，还大幅提升了长文本处理的效果。LongRoPE 的应用前景广阔，将为大模型的发展带来更多可能。

在2024年，长文本问题已成为大模型发展中备受关注的关键挑战。人们普遍认为，能够接受无限长度输入的大模型将会带来许多重大突破。例如，它可以一口气通读整套百科全书、冗长的法律条文、或大部头的经典医学教材，并准确提供任意章节的简要引用。这对于研究人员和公众都将是巨大的助益。如果大模型可以将一个人所有相关的信息（文本、照片、音视频等）作为上下文全部输入，那么甚至可能为该人创建一个可交互的数字副本。这些潜在的应用场景将为大模型开辟更广阔的前景。

但，实现长文本并非易事。目前主流的大模型通常只提供一个有限且较短的预定义上下文窗口。例如，LLaMA2 允许输入最多4096个 tokens。当输入超过该限制时，由于模型没有在预训练中见过超出上下文窗口的新的 token 位置，其性能会显著下降。

最近的研究表明，通过在更长的文本上进行微调，预训练的大模型上下文窗口可以扩展到约128k。但进一步扩展上下文窗口则存在三个主要挑战：首先，未经训练的新位置索引引入了许多异常值，使得微调变得困难。例如，当从 4k tokens 扩展超过1000k时，会引入超过90%的新位置，这就使得现有的微调方法变得难以收敛。其次，微调通常需要相应长度的长文本，但当前训练数据中长文本数量有限。此外，随着上下文窗口的继续扩展，模型的计算量和内存需求将显著增加，带来极其昂贵的微调时间成本和 GPU 资源开销。最后，当扩展到超长的上下文窗口后，由于引入众多新位置信息，大模型的注意力会分散，从而降低了大模型在原始短上下文窗口上的性能。

为了解决这些挑战，微软亚洲研究院的研究员们推出了 LongRoPE。作为迈向无限上下文窗口的第一步，LongRoPE 首次将预训练的大语言模型（LLMs）的上下文窗口扩展到了2048k（约210万）个 tokens，而其实现仅需在256k的训练长度内进行1k次微调步骤即可，同时仍能保持原始短上下文窗口的性能。

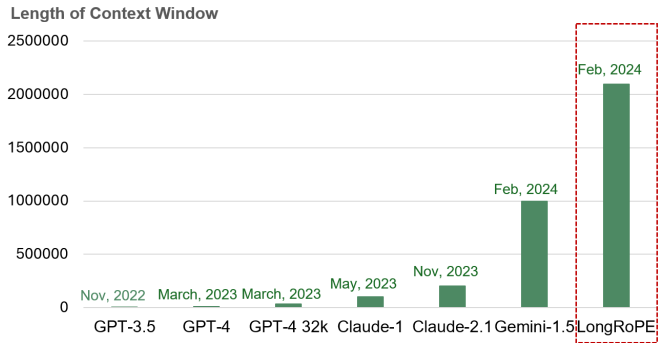


图1：LongRoPE 现在保持着最长 LLM 上下文窗口的记录

LongRoPE的主要方法

精细化非均匀位置插值。目前的大模型通常采用 RoPE 旋转位置编码，对 RoPE 位置编码进行插值是解决上述挑战中第一个问题的一种常见方法。这种方法将新位置索引缩小到预训练范围内。在已有的相关工作中，位置插值（position interpolation，PI）会通过扩展比例来线性插值 RoPE 的旋转角度。NTK-aware 位置编码插值方法提出，利用公式对每个RoPE维度进行经验性重新缩放，YaRN 会将 RoPE 维度分成三组，并分别针对三组 RoPE 维度进行不同的缩放（即直接外推，NTK-aware 插值和线性插值）。然而，这些方法主要基于启发式经验插值，未充分利用 RoPE 中的复杂非均匀性，导致关键信息在位置编码插值后丢失，从而限制了上下文窗口的大小。

研究员们经过实验发现，有效的位置编码插值应考虑两种非均匀性：不同的 RoPE 维度和 token 位置。低维 RoPE 和初始 token 位置存储着关键信息，因此需要进行更少程度的插值。相比之下，高维 RoPE 存储的信息相对较为稀疏，可进行较大程度的插值。为了充分利用这些非均匀性，研究员们提出了一种基于进化算法的方法，允许搜索 RoPE 每个维度以及不同 token 位置的旋转角度缩放因子，有效地保留了原始 RoPE 位置编码中的信息。这

种方法最大程度地减小了位置插值引起的信息损失，从而为微调提供了更好的初始化。此外，这种方法还允许在不微调的情况下实现8倍的扩展。

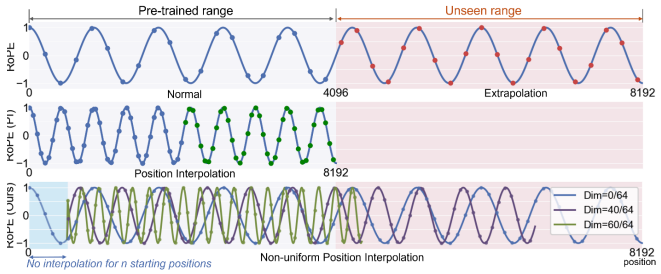


图2：不同位置编码插值方法的比较。上：RoPE 在直接外推下的表现；中：线性位置编码插值下的 RoPE；下：LongRoPE 在不同 RoPE 维度以及位置上的非均匀性插值。

渐进式扩展策略。在非均匀位置编码插值的基础上，LongRoPE 采取了高效的渐进式扩展策略，从而在无需直接对极长文本进行微调的情况下，有效实现了2048k上下文窗口的目标。具体策略如下：首先在预训练的大模型上搜索256k上下文窗口对应的位置编码插值方案，并在此长度下进行微调。其次，由于 LongRoPE 的非均匀插值允许在不微调的情况下进行8倍扩展，所以研究员们对已扩展的微调后的大模型进行了二次非均匀插值搜索，最终达到了2048k上下文窗口。

恢复短上下文窗口性能。在将上下文窗口扩展到极长的2048k后，研究员们注意到原始上下文窗口内的性能出现了下降。这是位置插值的一个已知问题，因为它导致原始上下文窗口内的位置被压缩在更窄的区域内，从而对大模型的性能产生了负面影响。为了解决这一问题，研究员们在扩展后的大模型上对8k长度内的 RoPE 缩放因子进行了重新搜索，旨在引导在较短长度上进行较少的位置插值，来恢复短上下文窗口的性能。在推理过程中，大模型可根据输入长度动态调整相应的 RoPE 缩放因子。

LongRoPE的实验性能

研究员们在 LLaMA2-7B 和 Mistral-7B 上应用 LongRoPE 并进行了测试，从三个方面评估了其性能。

第一项测试是在长文档上评估扩展上下文语言模型的困惑度。在256k以内的评估长度上，研究员们使用了 Proof-pile 和 PG19 数据集来进行测试。LongRoPE 在4k-256k的文本长度上，整体上显示出困惑度下降的趋势，优于基准。即使在上下文窗口长度是标准长度16倍的情况下，LongRoPE-2048k 模型在256k上下文长度内也超过了最新基线水平。

Base LLM	Model Name	Context Window	Extension Method	Evaluation Context Length 4096	8192	32768	65536	98304	131072	262144
LLaMA2-7B	LLaMA2-7B	4k	-	**3.58**	$>10^3$	$>10^3$	$>10^3$	$>10^3$	$>10^3$	$>10^3$
	Together	32k	PI	3.69	3.50	2.64	$>10^2$	$>10^3$	$>10^3$	$>10^3$
	LongLoRA	100k	PI	3.83	3.62	2.68	2.44	2.33	9.89	$>10^3$
	Code LLaMA	100k	NTK	3.95	3.71	2.74	2.55	2.54	2.71	49.33
	YaRN (s=16)	64k	YaRN	3.69	3.51	2.65	2.42	$>10^1$	$>10^1$	$>10^3$
	YaRN (s=32)	128k	YaRN	3.75	3.56	2.70	2.45	2.36	2.37	99.64
	LongRoPE-2048k (ft=128k)	**2048k**	LongRoPE	3.67	**3.49**	**2.60**	**2.36**	**2.27**	**2.26**	**1.88**
	LongRoPE-2048k (ft=256k)	**2048k**	LongRoPE	3.69	3.64	**2.63**	**2.38**	**2.28**	**2.26**	**1.87**
Mistral-7B	Mistral v0.1	8k	-	**3.09**	2.96	$>10^2$	$>10^3$	$>10^3$	$>10^3$	$>10^3$
	YaRN (s=8)	64k	YaRN	3.18	3.04	2.37	2.20	10.39	57.4	$>10^3$
	YaRN (s=16)	128k	YaRN	3.21	3.08	2.41	2.24	2.18	2.19	4.91
	LongRoPE-2048k (ft=128k)	**2048k**	LongRoPE	**3.09**	**2.95**	**2.31**	**2.12**	**2.06**	**2.06**	**1.77**
	LongRoPE-2048k (ft=256k)	**2048k**	LongRoPE	3.10	**2.96**	**2.30**	**2.12**	**2.06**	**2.06**	**1.77**

表1：经过不同位置插值方法进行上下文窗口扩展之后的 LLaMA2-7B 和 Mistral 在 Proof-Pile 数据集上的困惑度对比

接下来，研究员们增加了测试难度，从 Books3 数据集中随机选取20本每本长度超2048k的书，并使用256k的滑动窗口进行评估。在 8k-2048k 的文本长度上，两种模型均取得了与基线相当或更优的困惑度表现。

Base LLM	Model Name	Context Window	Extension Method	Evaluation Context Length 8k	16k	32k	64k	128k	256k	512k	1024k	2048k
LLaMA2-7B	LongLoRA	100k	PI	6.99	6.80	6.66	6.59	20.57	246.45	$>10^3$	$>10^4$	$>10^4$
	Code LLaMA	100k	NTK	7.68	7.49	7.38	7.88	9.80	98.30	$>10^3$	$>10^4$	$>10^4$
	YaRN (s=16)	64k	YaRN	**6.33**	**6.20**	**6.11**	**6.06**	$>10^4$	$>10^4$	$>10^4$	$>10^4$	$>10^4$
	YaRN (s=32)	128k	YaRN	6.38	6.25	6.16	6.11	**6.12**	$>10^4$	$>10^4$	$>10^4$	$>10^4$
	LongRoPE-2048k (ft=128k)	**2048k**	LongRoPE	6.53	6.35	6.24	6.18	6.17	**6.17**	**6.36**	**6.83**	**7.80**
	LongRoPE-2048k (ft=256k)	**2048k**	LongRoPE	6.79	6.66	6.31	6.27	6.21	**6.17**	**6.17**	**6.35**	**7.08**
Mistral-7B	Mistral v0.1	8k	-	**6.32**	66.61	$>10^4$	$>10^3$	$>10^3$	$>10^3$	-	-	-
	YaRN (s=16)	64k	YaRN	6.59	6.48	6.42	6.45	104.15	727.20	$>10^3$	$>10^4$	$>10^4$
	YaRN (s=32)	128k	YaRN	6.70	6.63	6.65	6.72	6.85	99.90	$>10^3$	$>10^4$	$>10^4$
	LongRoPE-2048k (ft=128k)	**2048k**	LongRoPE	6.42	**6.25**	**6.14**	**6.18**	**6.31**	**6.51**	**6.93**	**7.51**	**9.48**
	LongRoPE-2048k (ft=256k)	**2048k**	LongRoPE	6.44	**6.28**	**6.19**	**6.19**	**6.35**	**6.61**	**7.40**	**7.75**	**11.25**

表2：经过不同位置插值方法进行上下文窗口扩展之后的 LLaMA2-7B 和 Mistral 在超长数据集 Books3 上的困惑度对比

第二项测试是用 Passkey 检索任务评估在海量无关文本中检索简单密钥的能力。具体而言，该任务将在一个很长的文本中随机隐藏一个五位数密码，然后让模型找出该密码。结果显示，现有模型的准确率在文本长度超过128k后迅速下降到0。而 LLaMA2-2048k 在4k-2048k文本长度范围内保持了90%以上的检索准确率，Mistral-2048k 在文本长度达到1800k之前保持了100%的准确率，在2048k时准确率下降到60%。

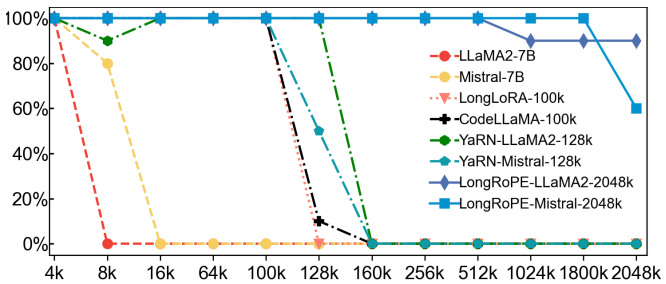


图3：不同长文本大模型在不同上下文长度下的 Passkey 检索精度

第三项测试是在短4096上下文窗口长度内评估标准大语言模型基准测试上的表现。这项测试的主要目的是检验模型上下文窗口被扩展后，在原有任务上的表现是否会遭受负面影响。LongRoPE-2048k 模型在原始上下文窗口大小的任务上，其表现与原始模型相当甚至更优。

综上所述，LongRoPE 可以广泛应用于基于 RoPE 位置编码的大模型，并在最新的主流大模型上得到了验证。LongRoPE 仅对位置编码进行了轻微修改，保持了模型的原始架构，并且可以重用大部分现有的优化技术。

未来，研究员们将进一步探索 LongRoPE 在其他大模型架构中的应用，并继续研究实现通往无限上下文窗口目标的技术。

Model	Context Window	ARC-c	HellaSwag	MMLU	TruthfulQA
(a) LLaMA2-7B with extended context window					
Original LLaMA2-7B	4k	53.1	78.6	46.6	39.0
Together	32k	47.6	76.1	43.3	**39.2**
Code LLaMA	100k	42.4	64.8	40.1	37.1
YaRN (s=16)	64k	52.4	**78.7**	42.4	38.2
YaRN (s=32)	128k	52.2	78.5	41.8	37.4
LongRoPE-2048k (ft=128k)	2048k	**52.9**	76.5	**43.4**	38.8
LongRoPE-2048k (ft=256k)	2048k	51.0	75.3	39.6	37.3
(b) Mistral-7B with extended context window					
Original Mistral-7B	8k	60.6	83.2	63.6	42.6
MistralLite (Amazon, 2023)	16k	59.2	**81.6**	50.4	38.3
YaRN (s=16)	64k	**59.3**	81.3	61.3	42.5

表3：不同长文本大模型在 Huggingface Open LLM benchmark 上表现

相关链接：

LongRoPE: Extending LLM context window beyond 2 million tokens
https://arxiv.org/pdf/2402.13753.pdf

MatterSim：人工智能解锁材料设计的无限可能

作者：科学智能中心

随着科技的不断发展，新材料已成为增强产业竞争力的关键因素。然而，新材料的物理和化学特性复杂多变，准确预测其属性，特别是实际合成和使用条件下的属性，是物质科学领域中长期存在的挑战，也是材料工业数字化转型的核心挑战之一。

为了破解这一难题，微软研究院科学智能中心（Microsoft Research AI for Science）开发了深度学习模型 MatterSim，能够在广泛的元素、温度和压力范围内实现准确高效的材料模拟与性质预测，为材料设计的数字化转型提供了强有力的支持。

新材料探索对纳米电子学、能量储存和医疗健康等多个领域的技术进步至关重要。材料设计中的一个核心难点是如何在不进行实际合成和测试的情况下预测材料属性。由于新材料可能涉及元素周期表中118种元素的任意组合，且其合成和工作温度、压力范围极广，这些因素极大地影响了材料内部原子的相互作用，使得准确预测材料属性和行为模拟变得极为困难。

为此，微软研究院科学智能中心（Microsoft Research AI for Science）推出了 MatterSim 模型。该模型将深度学习技术和大规模第一性原理计算相结合来学习原子之间的相互作用，学习的材料空间从绝对零度到5000开尔文，从标准大气压到一千万倍大气压，能够高效处理多种材料的模拟，包括但不限于金属、氧化物、硫化物、卤化物及其不同状态（如晶体、非晶固体和液体）。此外，MatterSim 还提供定制选项，可以通过整合下游场景的数据来执行更为复杂的预测任务。

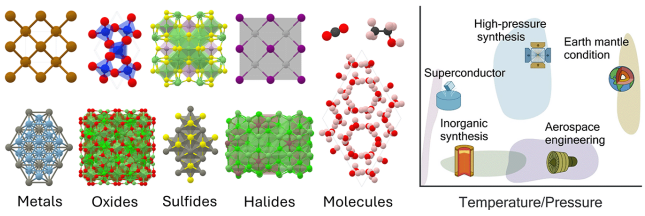


图1：MatterSim 能够在广泛的温度、压力条件下，模拟金属、氧化物、硫化物、卤化物等各种材料的性质和行为。

全周期表、跨温度、跨压力的真实条件下的材料模拟

MatterSim 的训练过程使用了大规模的合成数据。为了获得这些训练数据，研究员们结合了主动学习、分子动力学模拟和生成模型等技术，构建了高效的数据生成方案。这种数据生成策略

确保了模型对材料空间的广泛覆盖，使其能够以与第一性原理预测相当的准确度，预测材料在原子层面的能量、力和应力。

与当前的 SOTA 模型相比，MatterSim 在有限的温度和压力下对材料属性预测的准确度提高了10倍。研究表明，MatterSim 能够准确模拟包括热力学、力学和运输特性在内的广泛材料属性。值得一提的是，MatterSim 可以从头构建材料相图，为新材料的可合成性与合成条件提供指导。

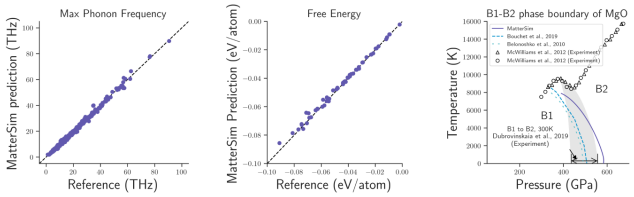


图2：MatterSim 预测属性与实验测量结果的比较。MatterSim 在预测材料的振动性质、机械性质和相图方面均达到了与第一性原理计算和实验测量相当的精度。

定制化精调：用更少的训练数据达到实验精度的模拟

尽管基于大量的合成数据集进行训练，MatterSim 可以通过整合额外的数据来满足特定的设计要求。MatterSim 能够利用主动学习和微调，定制化完成特定场合下高度复杂的材料模拟和设计任务。

以物质计算领域的经典任务——液态水性质的模拟为例，这个任务看似简单，实际上却需要大量的计算。通过 MatterSim 的定制化功能对该任务进行优化，MatterSim 只需要3%的原始数据，就能达到预期的实验精度模拟。相比之下，用传统方法训练的专有机器学习模型，则需要提供30倍的资源，而第一性原理的方法则需要增加指数级的资源。

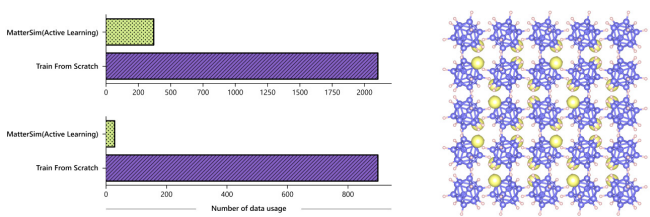


图3：MatterSim 在进行精细材料模拟、性能预测时，能够降低90%-97%的数据需求量。

弥合材料微观模型与宏观属性间的鸿沟

从原子结构预测材料宏观属性是一项复杂的任务，对于目前基于统计力学的方法，如分子动力学，太过复杂。MatterSim 通过深度学习技术直接映射这些关系来解决这个问题。研究员们为 MatterSim 专门设计了可以微调的适配器模块，可以直接从结构数据预测材料属性，避免了复杂模拟的需求。在著名的材料属性预测基准测试集 MatBench 上的基准测试中，MatterSim 的精确度有了显著提升，并优于所有针对特定属性的专有模型，展现了其直接从领域特定数据预测材料属性的强大能力。

开启人工智能辅助材料设计新篇章

MatterSim 为鲁棒且高效的材料模拟和性质预测提供了新的可能性，其与 Distributional Graphormer 等生成式人工智能技术和强化学习技术的结合，将有望彻底改变材料科学研究的理念与模式。以 MatterSim、MatterGen 为代表的材料科学与人工智能模型，也将使定制化材料的开发变得更加高效，应用范围更广，从半导体技术到生物医药工程等诸多领域都将从中受益。

未来，研究员们将继续推进 MatterSim 的实验验证，希望相关研究可以在面向可持续发展的催化剂设计、能源存储、纳米技术等领域发挥作用，造福全球社会。

相关链接：

MatterSim: A Deep Learning Atomistic Model Across Elements, Temperatures and Pressures
https://arxiv.org/abs/2405.04967

Distributional Graphormer：从分子结构预测到平衡分布预测
https://www.msra.cn/zh-cn/news/features/distributional-graphormer

MatterGen: Property-guided materials design
https://www.microsoft.com/en-us/research/blog/mattergen-property-guided-materials-design/

MatBench
https://matbench.materialsproject.org

重塑认知科学中的“行为”理解，提升智能体决策能力

认知科学认为，从人类简单的日常习惯到复杂的决策制定，都是习惯性行为和目标导向行为交互的体现。为了更好地理解二者间地关系，微软亚洲研究院的研究员们从人工智能和认知科学的交叉研究出发，通过引入贝叶斯“意图”变量，将习惯性行为与目标导向行为进行了整合，达到了计算效率和灵活性的平衡。相关论文已在《自然-通讯》（Nature Communications）杂志上发表。

扫描二维码查看文章

深度学习作业低GPU利用率问题的实证研究 (ICSE 2024)

微软亚洲研究院与微软Azure云平台部门对Platform-X深度学习生产平台上的低GPU利用率问题进行了深度归因，指出了根本问题及十五个子类的原因，并给出了相关修复意见。

KPDDS：通过关键点驱动的数据合成解决数学问题

为了增强LLMs处理数学问题的能力，研究员们引入了关键点驱动的数据合成（KPDDS）新范式，其通过理解和应用数学问题的核心概念来合成训练数据，为大语言模型的训练提供了更为丰富且准确的数据资源。

MathScale：大规模合成数学推理的指令微调数据

MathScale可用于大规模合成高质量的数学推理指令微调数据。通过概念图和随机游走算法生成多样化的问题和答案组合，MathScale可提升大语言模型在数学推理方面的能力。

RecAI：大模型改进推荐系统的五种方式

研究员们在RecAI一文中梳理并开源了大语言模型改进推荐系统的5种方式，以解决现有推荐系统面临的挑战，提升模型的可交互性、可解释性和可控性。

扫描二维码查看文章

应对深度强化学习中的信号延迟问题(ICLR 2024)

针对深度强化学习(DRL)的信号延迟问题,研究员们通过扩展马尔可夫决策过程框架定义了延迟观测马尔可夫决策过程(DOMDP),然后提出了一系列新方法来提升相关性能。

级联强化学习(ICLR 2024)

级联强化学习模型能够有效将用户状态及其变化纳入推荐过程中,并通过基于动态规划设计的快速离线求解器和强化学习算法CascadingVI,实现在非理想条件下的高效推荐。

DyVal:首个大语言模型的动态评测协议(ICLR 2024)

研究员们提出了名为 DyVal 的动态评测协议用于评估大语言模型,并通过实验证明其生成的数据可以作为模型数据增强的手段,提升模型性能。同时,研究指出静态基准测试存有潜在问题,模型在复杂数据集上的表现较差,提示需要更复杂的多任务来评估模型的能力。

KOSMOS-2:将多模态语言模型同视觉世界连接对应(ICLR 2024)

KOSMOS-2 是一个多模态语言模型,具备 Grounding 和 Referring 两种新能力。依托于这两种能力,KOSMOS-2提供了一个更灵活、更通用的视觉-语言任务人机界面。

MG-TSD:基于引导学习过程的多粒度时间序列扩散模型(ICLR 2024)

多粒度时间序列扩散MG-TSD模型,利用数据内在的多粒度水平作为中间扩散步骤的目标,以引导扩散模型的学习过程,有效解决了扩散模型在时间序列预测中的不稳定性问题。

扫描二维码查看文章

CVPR 2024 Highlight 论文 CoPoNeRF:统一对应点估计、相机姿态估计和神经辐射场重建,实现端到端双视图新视角合成

CoPoNeRF 框架无缝整合了二维图像对应点匹配、相机相对姿态估计与神经辐射场渲染,并通过强调三个任务共享统一表征和端到端训练,增强了各子任务间的协同性,从而在有限图像和相机姿态变化较大的情况下提高了新视角渲染效果。

DCVC-FM:基于特征调制的视频编解码器(CVPR 2024)

研究员们通过设计特征调制技术,提出了基于条件编码的视频编解码器模型DCVC-FM,解决了可变码率支持和时域误差传播的问题,提高了视频编码效率和质量。

MVGFormer:用于3D人体姿态估计的多视角几何Transformers(CVPR 2024)

混合模型MVGFormer通过混合模型设计、迭代细化过程和端到端训练与评估,解决了多视角数据中遮挡问题和视角变化所带来的挑战,提升了模型对于新视角和遮挡情况的泛化能力。

文本分组适配器:将文本布局分析能力装配在任意文本检测器上(CVPR 2024)

研究员们提出了通用文本分组适配器Text Grouping Adapter(TGA),通过文本区域特征组装模块和文本组合掩码预测模块,解决了场景文本布局分析中的问题,显著提升了文本布局分析性能,并大幅加速了训练流程。

扫描二维码查看文章

程鹏：“研究员+工程师”模式的探路者，推动AI与系统协同进化

尽管人工智能大模型在处理任务的广度上已经取得了突破，但它在特定领域的深度和精确性上仍存在局限。为了应对这一挑战，微软亚洲研究院提出了特定领域智能的概念，并沿着AI与系统协同进化（AI-System Co-evolution）的研究路线，利用人工智能技术自动化设计和构建底层系统与硬件，为用户提供定制化的服务。微软亚洲研究院资深首席研究员程鹏所在的微软亚洲研究院（温哥华）团队，正致力于这一研究任务。他们基于“研究员+工程师”的协作研发模式，整合全球人才和资源，推动特定领域智能研究的深入发展。

人工智能的创新突破不断刷新人们的认知，带来了众多令人瞩目的成果。尽管以构建通用型人工智能为目标的研究取得了显著进展，但很多技术在实际的产业落地时并不能很好地适应垂直领域的特定业务场景。

“在传统的技术落地流程中，科研人员会将创新成果交付给工程师，随后由工程团队负责技术的实施落地，研究员与工程师的工作相对独立。但在人工智能领域中，要想实现技术的实际应用，就必须深度整合创新技术与实际需求。因此，从研究阶段开始，工程师就需要与研究员紧密合作。”微软亚洲研究院资深首席研究员程鹏说。

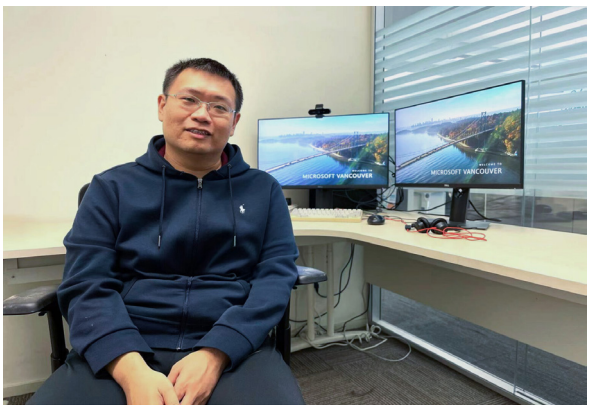


图1：程鹏，微软亚洲研究院资深首席研究员

为了促进这种合作，微软亚洲研究院提出了“研究员+工程师”的创新协同研发模式，并在微软亚洲研究院（温哥华）率先部署，希望打破传统壁垒，加速技术从理论方法到现实应用的转化。作为微软亚洲研究院（温哥华）的首批研究员之一，程鹏正与团队一道，以全新的研发模式推动人工智能技术的发展和应用。

微软亚洲研究院（温哥华）：专注特定领域智能探索

微软亚洲研究院在温哥华建立实验室，一方面加强了微软全球研究网络的合作效应，在物理空间上缩短了研究员与工程团队之间的距离，并为微软研究院位于太平洋区域的实验室之间的高效合作架起了桥梁；另一方面，微软亚洲研究院希望通过“研究员+工程师”的紧密协作模式，推进通用型人工智能模型在垂直领域的应用，以实现特定领域的智能。

程鹏说，“尽管通用型人工智能在处理广泛任务上的能力有目共睹，但它仍面临着三大挑战。首先，通用型人工智能的学习成本极高。随着模型规模的不断扩大，所需的资源和成本也随之增加，这可能会超出当前技术和社会经济的承受能力。其次，通用型人工智能的精确性还需提高。正如人脑在积累了大量知识后可能会遗忘或混淆某些信息一样，模型有可能也会产生幻觉或错误信息。此外，通用型人工智能在特定领域的持续学习与深度理解能力存在局限。在知识广度增加时，人脑会不断提炼所学信息，以更深入地理解和掌握特定领域的知识。比如一个人，从大学生到博士生，在学习专业领域技能时还会随着领域的发展持续进行学习。模型也需要具备这种持续学习和深度理解的能力。”

为了解决这些问题，微软亚洲研究院（温哥华）团队提出了特定领域智能的概念，旨在将通用型人工智能模型应用于具体的业务场景中，使其能够适应更多的差异化需求，并在特定领域内发挥最大作用，同时实现低成本和高效率的应用，释放AI的更多潜能。

实现特定领域智能的一个关键前提是利用人工智能技术对云服务和人工智能的底层架构进行重新设计。然而，这并非是研究员能够独立完成的工作，而是需要系统、硬件、人工智能领域的专家深度合作。“微软亚洲研究院（温哥华）团队已经吸引了来自全球知名学府的十几位杰出人才加盟，并积极招募更多相关领域的顶尖人才。我们期待通过汇聚全球顶尖人才的智慧和专长，利用人工智能技术重构未来的系统和硬件，推动人工智能在各行业中的应用。”程鹏介绍道。

图2：微软全球资深副总裁、全球研究与创新孵化负责人 Peter Lee（第二排右四）等微软研究院同事访问期间，与部分微软亚洲研究院（温哥华）团队成员合影

人工智能与系统协同进化：基于用户需求自动设计AI系统

正如一座稳固的建筑物需要坚实的地基来支撑其宏伟的结构一般，人工智能基础设施（AI Infrastructure）也是推动技术进步和应用创新的关键基石。随着人工智能模型规模的不断扩大和模态的增多，对 AI Infrastructure 的升级改造变得尤为迫切。程鹏所在的系统与网络组，当前的首要任务就是专注于 AI Infrastructure 的研究。

“人们通常将 AI Infrastructure 视为支撑人工智能的最底层硬件设施，但我们对 AI Infrastructure 的定义更为广泛。它不仅包括顶层的应用，这些应用将根据用户需求提供特定的逻辑和功能；还包括中间的系统层，负责将应用逻辑与硬件资源相连接；以及底层的基础架构，即硬件及其执行方式。我们将这三层整体称为系统，我们的目标是根据用户的需求，利用人工智能技术自动设计这一整体系统，并将其称之为 AI-System Co-evolution（人工智能与系统协同进化）。”

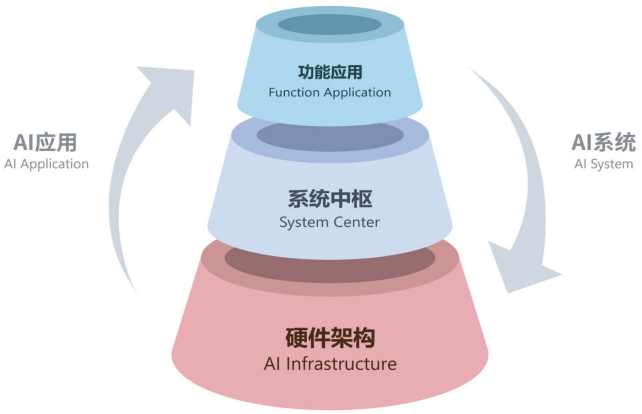


图3：AI-System Co-evolution 理念架构图

程鹏举了一个简单的例子来说明这一概念，就像个人电脑在不同行业的应用，底层是以 CPU 为核心的硬件，中间是操作系统如 Windows，上层则是根据业务需求安装的各类软件。在传统模式中，硬件和系统先行，然后才是上层应用的需求。但人工智能技术的融入将颠覆这一流程——可以根据用户需求来设计系统和硬件。

也就是说，在人工智能系统协同进化的理念下，我们可以根据用户的不同需求，设计并制造出专门运行特定业务应用的系统和硬件基础架构，其核心将不再局限于 CPU，而可以是更加灵活的 “X” PU。例如在云服务场景中，AI-System Co-evolution 能够针对客户的关键业务场景，在最短时间内设计出从软件到硬件的高效协议栈，提供高度定制化的解决方案。

“如今，支持人工智能的底层硬件与系统，比如专门定制的 GPU，经过了大约十年的发展和大量的资源投入才逐渐成熟。但是，通过将人工智能技术融入系统和硬件设计这一方式，我们有望将这一过程从十年缩短到一个月。AI-System Co-evolution 不仅能够显著加快技术进步的速度，还将引领全新的设计思维，为系统研究和基础架构设计带来革命性的变化。” 程鹏说。

开展以目标为导向的研究

自2015年加入微软亚洲研究院以来，程鹏一直专注于系统与网络领域的研究。博士毕业时，他曾面临两个选择：留在学术界或者进入工业界。最终程鹏选择了能够 “二者兼得” 的微软亚洲研究院。程鹏说：“网络和系统是两个密不可分的领域。我的学术旅程始于网络专业，在博士期间转向了网络系统。加入微软亚洲研究院后，我的研究则从网络系统开始，逐渐延伸到硬件系统、人工智能系统和硬件基础架构。在这里，我得以兼顾学术创新与工程技术成果的转化。”

最初，程鹏的工作聚焦于网络领域，主要是为微软必应（Bing）搜索设计和开发 RDMA 网关（RDMA Gateway），以优化跨数据中心或跨地区的数据传输和通信速率。同时，他还与微软 Azure 云计算服务团队合作，开发了 Web 流量负载均衡器 Azure Application Gateway，相关技术沿用至今。在这一阶段，程鹏的研究主要集中在上层应用软件的开发和优化上。随着研究的深入，程鹏的研究范围扩展到了系统、硬件等基础架构层面。例如，功能的硬件卸载及资源池化方面的研究显著提升了整体服务的性能，并提高了资源整体的利用率，相关成果也已在微软 Azure 云计算存储和微软必应（Bing）存储中完成了原型设计。

人工智能时代，程鹏和团队成员开始探索将人工智能技术更深入地集成到产品和业务中，如利用人工智能进行微软 Azure 虚拟机中的虚拟 NUMA 放置以及 Microsoft Teams 中的带宽预测。同时，他和团队还启动了 AI for System 的相关研究，旨在进一步推动人工智能与系统技术的融合。

图4：程鹏（第二排左八）和同事们的合照

程鹏在梳理了自己的研究脉络后认为，这是一个自上而下、逐步深入的探索过程——从上层软件应用出发，逐层深入到底层架构，并进行针对性的创新和优化。“无论我们从事何种工作，都是先设定目标，然后制定计划，一步步实现。现阶段我们的目标就是满足最终用户的需求，所以只有深入理解这些需求，我们才能设计出更符合用户需求的底层架构。”程鹏强调，“这种以目标为导向的研究方法为我们团队当前的 AI-System Co-evolution 研究奠定了基础。”

从偶发跃迁式突破到持续渐进发展，人工智能加速创新研究

在过去的10到15年中，系统研究的发展相对缓慢，但人工智能的进步为这一领域注入了新的活力。“以往的系统研究与优化需要投入大量的人力和时间，成本高昂且周期漫长，科研人员很难再有时间和精力进行更深层次的思考。而人工智能可以帮助人类处理繁琐的工作，大大释放了科研人员的创造力，使我们能够集中精力解决更为关键的问题。”

程鹏进一步认为，人工智能必将改变科学研究的方式。在过去，跨领域的研究成果难以实现渐进式的整合，往往需要长时间的积累，最终由某位研究者汇总并实现重大突破，这限制了科学进步的速度。但利用人工智能技术则可以自动融合不同来源的知识和创新成果，进而推动科学研究从偶发跃迁式的突破向累积型进步转变，为科研人员提供在先进成果上进行持续创新的机会。

“人工智能带来的创新力量，结合微软研究院遍布全球的研究网络以及来自世界各地的多元人才，再加上科学研究和工程实践相辅相成共同演进的模式，让我们温哥华团队的沟通协作变得更加紧密，创新效率大幅提升。”程鹏表示了坚定的信心，“我相信，不久的将来，人工智能将能够根据具体需求设计出定制化的系统和硬件基础架构，AI-System Co-evolution 的概念将从梦想变为现实，特定领域智能也终将成为可能。”

相关阅读：

科学匠人 | 黄昶互：坚持长期主义研究，是一个不断说服自己的过程

他，刚入职微软亚洲研究院一年，却有着丰富的学术合作经验；刚刚博士毕业，就有多篇论文被业内顶会收录并获奖；能够长期投入到一项研究课题中，并持续跟进三、四年；选定一个研究领域，层层递进，展开了多方面的研究；热衷于科学创新，坚持长期主义的研究理念。他是来自韩国的微软亚洲研究院研究员黄昶互（Changho Hwang）。让我们一同走进他的故事，感受他对研究的热忱与专注。

扫描二维码查看文章

科学匠人 | 李潇：Aspire的力量——年轻科研人员的成长加速器与沟通桥梁

李潇曾是微软亚洲研究院的“首席实习生”。在经历了互联网行业的实践之后，他选择重回微软亚洲研究院，成为了一名多媒体计算组的研究员，同时也是微软亚洲研究院 Aspire 社团的核心力量之一。微软亚洲研究院的 Aspire 是一个怎样的组织？从“首席实习生”到高级研究员，李潇的心态和研究之路又发生了怎样的变化？

扫描二维码查看文章

上海科技|邱锂力："第六感官"将是AI的"下一块拼图"

本文转载自上海科技，作者张悦

OpenAI 发布了新旗舰模型 GPT-4o，o 意味 omni，全能。发布会上，GPT-4o 已展现出足以媲美真人的"听说读写"能力。

"五感"既然已被打通，AI 进化的下一步将会落棋何处？

上海科技对话了微软亚洲研究院副院长邱锂力，她用一块手掌大小的超表面正方形，向我们展示了她所看到的多模态大模型的确定未来——无线感知将成为 AI 的"下一块拼图"，成为名副其实的第六感官。

图1：邱锂力，微软亚洲研究院副院长，微软亚洲研究院（上海）负责人

2022年1月，邱锂力博士正式加入微软亚洲研究院，担任副院长一职，主要负责微软亚洲研究院（上海）的研究工作，以及与产学研各界的合作。加入微软亚洲研究院之前，邱锂力博士在美国得克萨斯大学奥斯汀分校担任计算机系教授。她也是全球为数不多同时拥有国际计算机学会会士（ACM Fellow）和电气电子工程师学会会士（IEEE Fellow）称号的华人学者。2022她成为美国国家发明家学会院士。

何以"第六感官"：打破无线通信旧"悖论"的超表面

为什么要无线感知？无线感知可以感知一些比较精细、细微的动作，而且也比较保护隐私。"比如在卫生间老人跌倒，我们不能放摄像头，但是可以用一些无线传感器来监测这些异常状况。"邱锂力说道，"无线感知也有不少医疗应用，如检测呼吸，在医院里检测呼吸需要穿戴很多设备，这非常不舒服。"

无线感知可以用声波或 WiFi 或毫米波实现。但在实际的应用中，目前却常常遭遇无线通信的悖论限制——无线感知需要兼具更快的通信速度与较远的感知距离，但这两者却往往不能两全。"为了增加通信速度，我们不断地提高频率，从之前的 WiFi 到现在的毫米波，到太赫兹。提升频率是一个有效提高速率的方法，但是它也大大降低了通讯和感知的距离。"

在收发端加更多的天线似乎可行，但却非常昂贵。对此，微软亚洲研究院（上海）则提出了一个新型的方案——创新智能环境，建立更有利于通信和感知的无线信噪，比如在环境中加一个非常低成本的结构，能更好地提高无线信噪的通信效率和感知精度。

该方案的核心，是源于光学的超表面技术。最近几年，超表面正慢慢用于优化无线技术。它能更自由地修改波前，比如相位、振幅、偏振等等。

据介绍，在一块超表面结构上，有很多细小的单元，每个单元都像一个小天线。但是不同于传统的无线天线阵列，那些天线阵列需要外部的激励源来激励，超表面则可以用收到的电磁波来激励，然后改变波前。通过设计每个超表面的单元，能精准修改出射波的相位和振幅。

智能超表面可应用在不同的场景，包括低轨道卫星通信、毫米波、全网覆盖以及无线感知。据邱锂力介绍，超表面在无线通信上，具有成本较低且方便部署的优势。例如，在低轨道卫星的应用方面，对于上行链路和下行链路，它都能大大提升信噪比，它的信号能提升45倍。同样功能的天线阵列则需要几千美元。

基于超表面的赋能，更多层面的无线感知将得以通过更低廉的成本实现。目前，无线感知可以用声波或 WiFi 或毫米波实现。例如，在家庭的智能音箱前，放置超表面结构，将能够进行对人体呼吸的感知，即使隔着被子，或者人在走动，也可以实现。

邱锂力表示："传统的 AI 主要分析视频和语音数据，而无线感知可以赋予 AI 第六感官，它让我们能看到视野之外的东西，能隔墙感知事物，在黑暗中也能感知事物。"

AI+健康：AI是pilot还是co-pilot？

尽管大语言模型正展现出越来越强大的生成和推理能力，但在医学领域中直接应用大模型还存在一些壁垒。对于医学领域，大模型往往还需要特定的改造，包括但不限于“读懂”脑电波、“理解”呼吸……对此，微软亚洲研究院基于基础模型自研、无线感知突破等前沿技术，已在 AI+健康领域取得了长足的进展。

据邱锂力介绍，其中，有一些项目已经取得了比较好的效果。例如 AI Neurologist，该系统能通过脑电波来检测用户是不是正在癫痫发作。“我们请医生对这个工具做了一些评估，得到了比较好的评价，它能提高医生对癫痫事件分类的准确率。”

对于脑健康的探索，微软亚洲研究院（上海）正走在前沿。在邱锂力看来，探索脑科学是一场人脑与 AI 的双向对话。“我们希望通过跨领域研究，用人工智能技术来帮助神经科学家更好地理解大脑，这种理解不仅有助于我们探索脑部疾病机理，促进脑健康，而且我们也可以通过从大脑汲取灵感，有望启发我们设计出更高效的人工智能。”邱锂力谈道，比如团队受大脑神经信号传输模式启发设计了一种新型神经网络，它用更少的参数能实现更好的效果。这些机器学习，也将继续用于医疗健康、新药发现等等场景。

而在脑科学之外，微软亚洲研究院（上海）也正尝试通过打通人类与 AI 的“五感”，对特殊的疾病进行更早诊断与介入。例如，在听与说方面，其正在开发智能语音早筛系统。“语音包含了很多丰富的信息，包括人的生理健康信息，比如我们的发音反映了发音器官的健康程度。发音的同时也能反映出头脑的健康程度，还有情绪的问题。”邱锂力介绍道，“所以，我们基于这些开发出了语音‘治疗师’，用于阿尔茨海默症的早筛，现在我们也正在关注通过语音来感知情绪。相关项目我们正在跟医院合作，希望能推动落地。”

同时，微软亚洲研究院（上海）也在探索通过视频做无监督的异常检测，比如自闭症患者有一些异常的刻板行为，通过建模，抽取 2D、3D 的关键点信息，利用刻板行为的一些特征，实现无监督异常行为监测。

邱锂力强调：“对于我们来说，人永远是 pilot（领航员），人是最核心的；即使是最好的 AI，也只能做人的 co-pilot，为人的工作与决策赋能。”

“小模型”将带来大变局

2024年4月底，微软正式发布了新一代 Phi-3，其中最小尺寸的 phi-3-mini，在各大公开的学术基准和内部测试中，实现了与 GPT 等大尺寸模型相同的性能。

而这意味着，大模型可以不再需要强劲的显卡与电脑，在手机上，也能跑出媲美 GPT 3.5 的“小模型”。

模型从大到小，关键之一，便是如何突破长上下文的提示词。大模型的应用现在很多都是靠提示词（prompt），提示词的长短直接影响到执行的时长和成本。同时，微软亚洲研究院（上海）也和位于北京的团队合作，开发出长上下文的提示词。之前的提示词只能用到128k的 token（标记），但在他们的工作下，他们做到了200万的 tokens。

在邱锂力看来，小模型的出现，将大大拓宽 AI 的服务范围与使用场景，在之后，也许面对手机、手表等边缘设备，也可以实时对话了，不一定在电脑前进行这些操作，本地化的操作可以更好地保护隐私，不用传到云端。

邱锂力表示：“如果能在手机上享受这种交互的话，那么就能有更广泛的人群受益于 AI。”

扫描二维码
观看邱锂力受访视频

量子位 | YOCO：打破传统Decoder-only架构，内存消耗仅为Transformer的六分之一

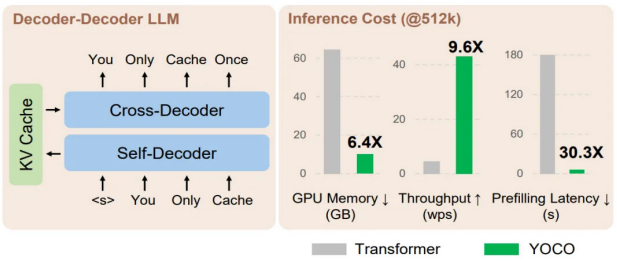
本文转载自量子位，作者关注前沿科技

2024年5月，微软亚洲研究院推出了一种创新性的 Decoder-Decoder 架构 YOCO（You Only Cache Once）。通过自解码器和交叉解码器的独特架构，YOCO 仅需缓存一次键值对，从而显著降低 GPU 内存的使用。

在模型评估中，YOCO 展现出与同规模 Transformer 模型相媲美的性能，并在语言建模评估、模型大小扩展以及长上下文处理方面具有显著优势。特别是在降低 GPU 内存占用和缩短预填充延迟方面，YOCO 实现了“模型越大，内存越省”，为自然语言处理领域带来了全新的研究和应用范式。

微软亚洲研究院&清华大学最新研究，打破 GPT 系列开创的 Decoder-Only 架构——提出 Decoder-Decoder 新型架构，名为 YOCO (You Only Cache Once)。YOCO 仅缓存一次键值对，可大幅降低 GPU 内存需求，且保留全局注意力能力。

一张图来看 YOCO 和标准 Transformer 的比较。



在处理 512K 上下文长度时，标准 Transformer 内存使用是 YOCO 的6.4倍，预填充延迟是 YOCO 的30.3倍，而 YOCO 的吞吐量提升到标准 Transformer 的9.6倍。

去年一张“大语言模型进化树”动图在学术圈疯转，模型架构还只有三大类：Decoder-Only、Encoder-Only、Encoder-Decoder。

那么这个新出的 Decoder-Decoder 架构到底长啥样？

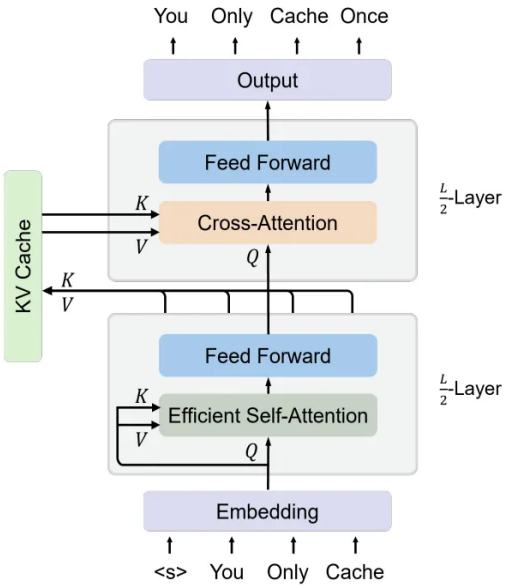
嗯，如网友所言，要读的论文又增加了。

话不多说，一起来看。

打破Decoder-Only

YOCO 整体架构设计如下，分为自解码器（Self-Decoder）和交叉解码器（Cross-Decoder）两部分。

具体来说，YOCO 由 L 个块堆叠而成，其中前 L/2 层是自解码器，其余模块是交叉解码器。自解码器利用高效自注意力（efficient self-attention）机制来获取键值（KV）缓存：接收输入序列的嵌入表示，并使用高效自注意力来生成中间向量表示；使用因果掩码（causal masking）保证解码的自回归特性；自解码器的输出用于生成全局 KV 缓存。



而交叉解码器使用交叉注意力（cross-attention）来重用自解码器生成的共享 KV 缓存：在自解码器生成的 KV 缓存基础上进行堆叠，以获得最终的输出向量；同样使用因果掩码来维持自回归生成；允许交叉解码器层间高效地重用 KV 缓存，减少了对 GPU 内存的需求。

总的来说，自解码器和交叉解码器的模块设计与 Transformer 的解码器层类似，包含交错注意力和前馈网络子层。不过，研究人员还进行了预 RMSNorm、SwiGLU 和分组查询注意力等改进。

两部分之间的区别在于注意力模块。

自解码器使用高效自注意力，如滑动窗口注意力（Sliding-Window Attention）或门控保留（gated retention）。而交叉解码器使用标准的多头交叉注意力，Query 向量通过注意力与自解码器产生的全局键值缓存相关联。

推理大幅度省、省、省

实验阶段，研究人员将 YOCO 模型与同体量的 Transformer 模型进行比较。分析维度有四个：语言建模评估、与 Transformer 比较的可扩展性、长上下文评估、推理优势。

语言建模评估

研究人员训练了一个 3B 参数的 YOCO 语言模型，并根据训练 token 数量（1T 和 1.6T）进行评估。

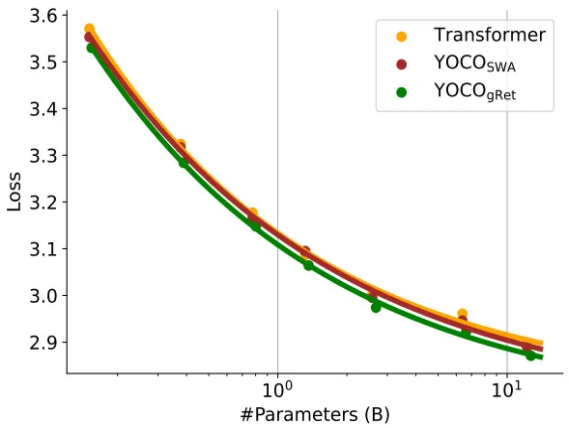
在 LM Eval Harness 的多个下游任务上，YOCO 与 Transformer 模型 OpenLLaMA-3B-v2、StableLM-base-alpha-3B-v2、StableLM-3B-4E1T 打得有来有回。

Model	ARC-C	ARC-E	BoolQ	Hellaswag	OBQA	PIQA	Winogrande	SciQ	Avg
Training with 1T tokens									
OpenLLaMA-3B-v2	0.339	0.676	**0.657**	**0.700**	0.260	0.767	0.629	**0.924**	0.619
StableLM-base-alpha-3B-v2	0.324	0.673	0.646	0.686	0.264	0.760	0.621	0.921	0.612
StableLM-3B-4E1T	—	0.666	—	—	—	**0.768**	0.632	0.914	—
YOCO-3B	**0.379**	**0.731**	0.645	0.689	**0.298**	0.763	**0.639**	**0.924**	**0.634**
Training with 1.6T tokens									
StableLM-3B-4E1T	—	0.688	—	—	—	0.762	0.627	0.913	—
YOCO-3B	0.396	0.733	0.644	0.698	0.300	0.764	0.631	0.921	0.636
Extending context length to 1M tokens									
YOCO-3B-1M	0.413	0.747	0.638	0.705	0.300	0.773	0.651	0.932	0.645

可扩展性对比

接着，研究人员在 160M 到 13B 参数规模范围内，分别训练了 YOCO（门控保留和滑动窗口注意力版本）和 Transformer 语言模型。

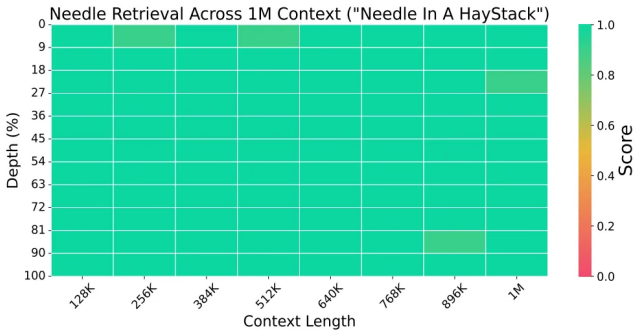
对比了它们在验证集上的语言模型损失，YOCO 的表现与 Transformer 基本持平：



结果证明，YOCO 在模型大小扩展方面有很强的可扩展性。

长上下文评估

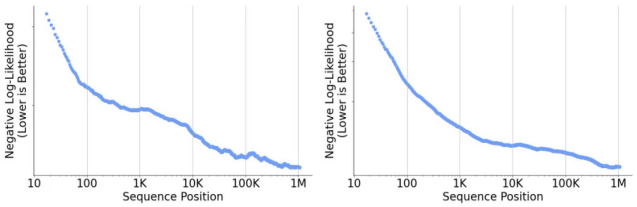
将 3B 的 YOCO 模型扩展到上下文为 1M，在“大海捞针”等长序列的 needle retrieval 任务上，YOCO-3B-1M 的准确率接近 100%。



在多针检索任务上，YOCO-3B-1M 的性能优于一些超 3B 的 Transformer 模型：

Model	Size	$N=1$	$N=2$	$N=4$	$N=8$
YaRN-Mistral-128K [PQFS23]	7B	0.02	0.12	0.08	0.20
LWM-1M-text [LYZA24]	7B	1.00	0.90	0.76	0.62
MiniCPM-128K [HTH+24]	2.4B	1.00	1.00	0.54	0.56
ChatGLM3-128K [ZLD+22]	6B	0.94	0.72	0.52	0.44
YOCO-3B-1M	3B	0.98	0.98	0.84	0.56

此外，YOCO 模型在长序列上的 NLL 随着上下文长度的增加而一致下降，表明 YOCO 能够有效地利用长距离依赖信息进行语言建模：

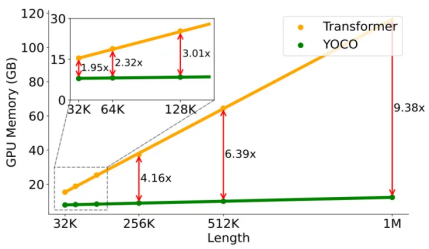


综上，可见 YOCO 在性能上完全不输 Transformer，关键来看 YOCO 在推理效率上取得的显著提升。

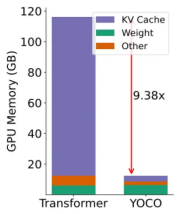
推理优势

研究人员评估了 YOCO 在 GPU 内存占用、prefilling 延迟、吞吐量和服务容量等方面的优势，评估上下文范围为 32K 至 1M。

如下图所示，与 Transformer 相比，YOCO 大幅度降低了 GPU 内存占用，且 YOCO 的内存消耗随上下文长度增加，增长幅度很小。例如，在 1M 长度下，整体推理内存使用量仅为 12.4GB，而传统的 Transformer 则占用了9.38倍的 GPU 内存。

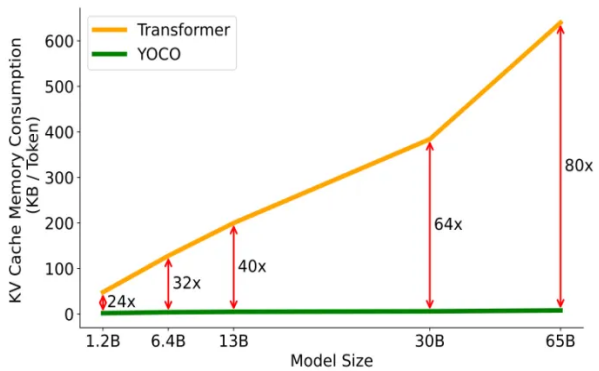


(a) Inference memory of Transformer and YOCO across various lengths.



(b) Breakdown memory consumption in 1M context length.

下面展示了 token 的 KV 缓存对 GPU 内存的占用情况。



YOCO 模型只缓存一层全局的键值对，因此与 Transformer 模型相比，它需要的内存约少了 L（指模型的层数）倍。

	KV Cache Memory
Transformer	$\mathcal{O}(LND)$
YOCO	$\mathcal{O}((N+L)D)$

Table 1: Inference memory complexity of KV caches. N, L, D are the sequence length, number of layers, and hidden dimension.

例如，YOCO 模型可以使用 1GB 的 GPU 内存来处理 128K token。而具有 GQA 的 Transformer 65B 大小模型，仅能支持 1.6K token。也就是说，模型越大，YOCO 可以节省更多。

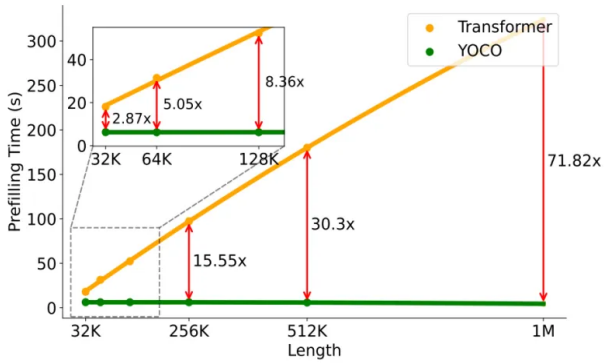
在预填充阶段，模型并行编码输入 token。对于 512K 和 1M 长度的输入，Transformer 分别需要大约180秒和300秒。Transformer 的计算复杂度为 O(N^2)，处理长上下文需要大量的浮点运算操作。相比之下，YOCO 的预填充时间为 O(N)，随序列长度线性增长。

	Prefilling Time
Transformer	$\mathcal{O}(LN^2D)$
YOCO	$\mathcal{O}(LND)$

Table 2: Prefilling time complexity of attention modules. N, L, D are the same as above.

YOCO 将 Transformer 的 512K 上下文预填充时间从180秒减少到不到6秒。

预填充阶段可以在进入交叉解码器之前提前退出。因此，即使对于短上下文，预填充延迟的加速至少是两倍。例如，对于 32K 长度，YOCO 比 Transformer 快2.87倍。

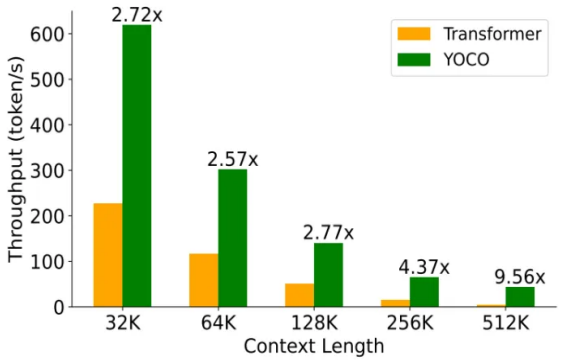


吞吐量表示模型每秒可以处理多少个 token，涵盖了预填充和生成时间。如下图所示，与 Transformer 相比，YOCO 在不同上下文长度下实现了更高的吞吐量。

以512K查询为例，Transformer 的吞吐量为4.5 token/秒，而

YOCO 达到了43.1token/秒，即实现了9.6倍的加速。

吞吐量提高的原因如前所述，YOCO 减少了预填充所需的时间。其次，由于内存消耗减少，因此可以在推理时使用更大的批量大小，这也有助于提高吞吐量。



相关链接：

You Only Cache Once: Decoder-Decoder Architectures for Language Models
https://arxiv.org/abs/2405.05254

相关阅读：

人工智能基础创新的第二增长曲线

在人工智能快速发展的今天，如何突破现有的技术瓶颈，实现跨越式增长？微软亚洲研究院全球研究合伙人韦福如在其署名文章中，分享了微软亚洲研究院在推进人工智能基础创新“第二增长曲线”方面所作出的努力。他表示，“我们希望直击人工智能第一性原理，通过革新基础网络架构和学习范式，构建能够实现效率与性能十倍、百倍提升，且具备更强涌现能力的人工智能基础模型，为人工智能的未来发展奠定基础。”

扫描二维码查看文章

微软亚洲研究院提出全新大模型基础架构RetNet，或将成为Transformer有力继承者！

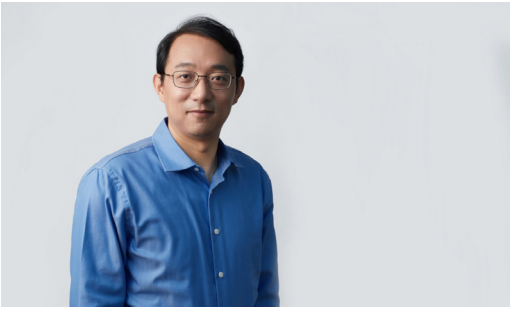
Transformer 在大语言模型中的重要性毋庸置疑。但 Transformer 也并非完美无缺，其并行处理机制是以低效推理为代价的，且序列越长占用的内存越多。而微软亚洲研究院提出的新型神经网络架构——RetNet 在 Transformer 的基础上，使用多尺度保持（retention）机制替代了标准的自注意力机制，让 RetNet 可以同时实现良好的扩展结果、并行训练、低成本部署和高效推理。这些特性将使 RetNet 有可能成为继 Transformer 之后大语言模型基础网络架构的有力继承者。

扫描二维码查看文章

BitNet b1.58：开启1-bit大语言模型时代

微软亚洲研究院和中国科学院大学的研究团队提出了一种名为 BitNet b1.58的1-bit大语言模型，通过引入三元表示和优化计算范式，实现了在保持模型性能的同时大幅降低内存和计算需求。其不仅为训练新一代高性能高效率的 LLMs 确立了新的扩展定律（scaling law）和方法，还引领了一种全新的计算范式，并为开发专为 1-bit LLMs 优化的硬件设备铺平了道路，为生成式AI领域带来了新的可能性。

扫描二维码查看文章



周礼栋

微软全球资深副总裁
微软亚太研发集团首席科学家
微软亚洲研究院院长

如今，我们正处在孕育新一代计算范式的关键节点。在不久的将来，虚拟世界和现实世界的边界会不断消弭，计算会像电力一样无处不在。新的计算范式将赋能人类生活和工作的方方面面，给各行各业带来颠覆性的变革，也将催生众多新的机遇。

面对科技发展的新浪潮，微软亚洲研究院将践行所有利于激发创新力的原则，持续致力于营造多元、包容、自由、平等、开放、可持续的研究氛围和科研协作环境，让各种具有创造性的想法、观点和创意，在微软亚洲研究院这个“化学反应池”中交流、碰撞、提炼和升华，使创新的星星之火形成燎原之势。同时，我们也将保持积极开放的姿态，与国内外各界伙伴携手，共同推动技术进步，实现人类社会的可持续发展。

关于微软亚洲研究院

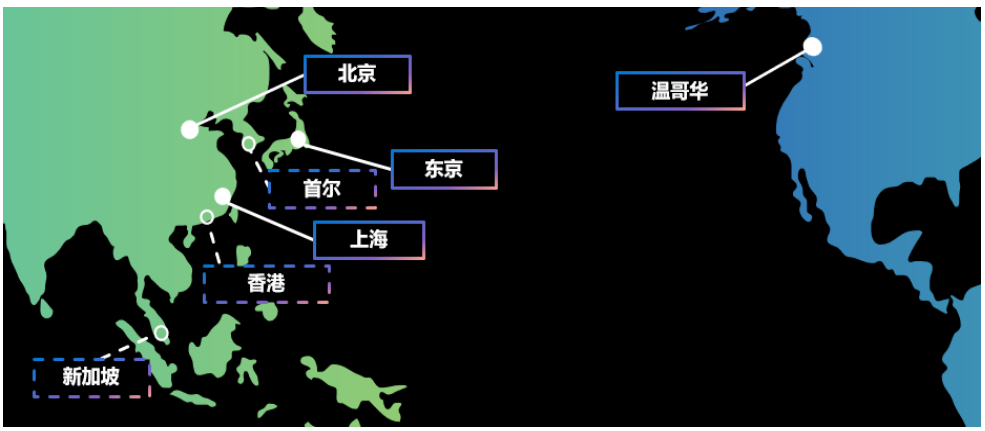
微软亚洲研究院成立于1998年，是微软公司在亚太地区设立的研究机构，在北京、上海、温哥华、东京、首尔、新加坡和香港设有实验室及研究岗位，研究方向涵盖计算基础创新、下一代智能交互、多维感知与通信、人工智能与社会福祉、科学发现与行业赋能等。通过来自世界各地不同学科和背景的多元人才的鼎力合作，微软亚洲研究院已经发展成为世界一流的计算机基础及应用研究机构。多年来，从微软亚洲研究院诞生的新技术层出不穷，对微软公司的产品创新以及全球范围的科技发展产生了深远的影响。

作为微软研究院全球体系的一员，微软亚洲研究院拥有广阔的国际视野，同时融合了东西方创新文化的精髓。秉持开放合作的理念，微软亚洲研究院始终与高校和科研机构开展持久而有效的合作，推动跨地区、跨文化和跨学科的交流，激发创新潜力，促进行业发展。

微软亚洲研究院倡导对技术进步怀有远大抱负，推崇富于冒险的极客创新精神，鼓励研究人员拓展研究的深度与广度，跨越计算机领域的界限，把视野拓展到解决具有广泛社会意义的问题上，为未来的计算新范式奠定基础，并为AI和人类发展创造更美好的未来。

扫描二维码观看视频介绍

微软亚洲研究院实验室分布



Microsoft

微信

知乎

电话：86-10-59178888

网址：http://www.msra.cn/

微博：http://t.sina.com.cn/msra