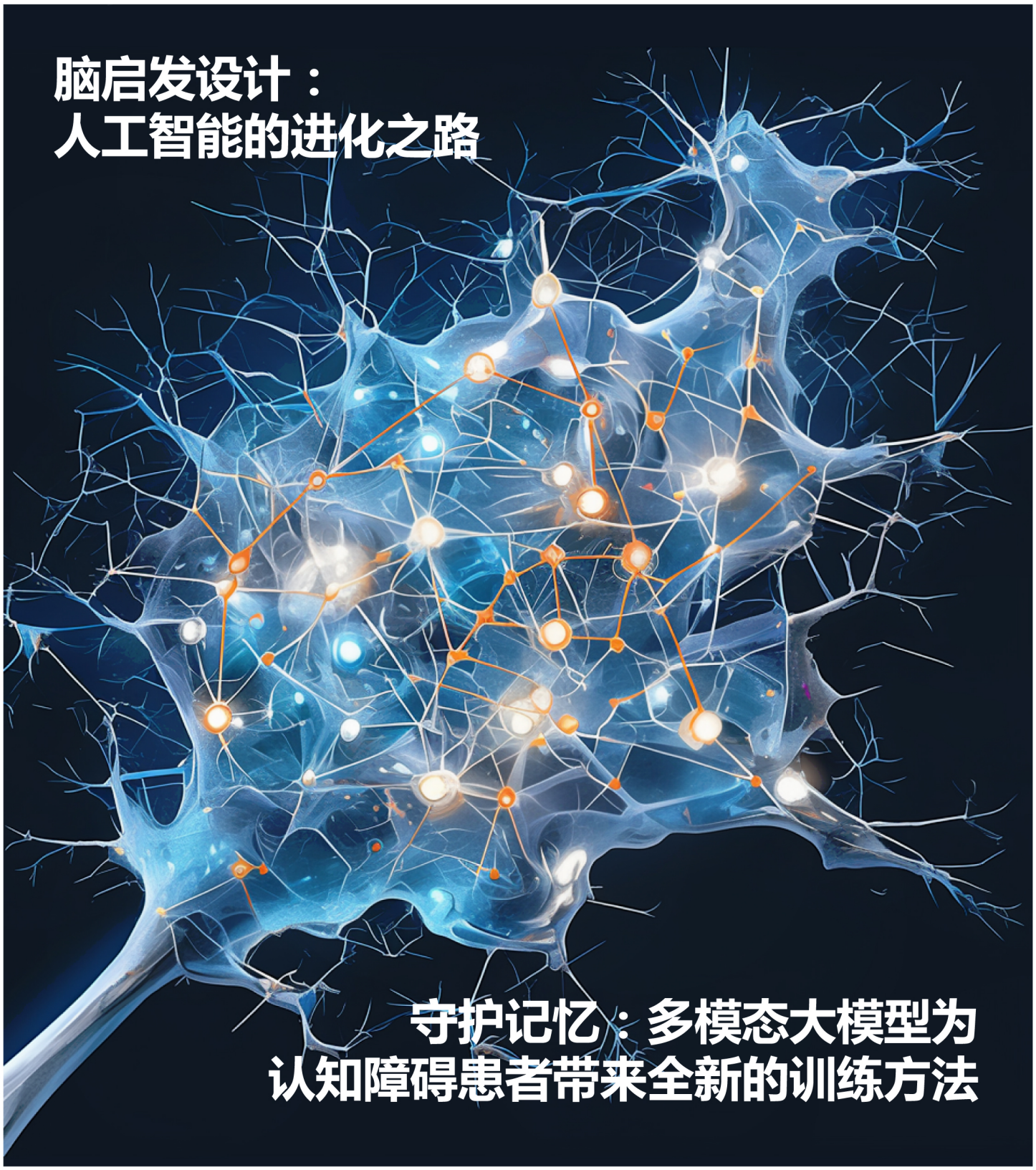
微软亚洲研究院院刊

Matrix

NO.70
2024年 7 - 9月



脑启发设计：人工智能的进化之路

守护记忆：多模态大模型为认知障碍患者带来全新的训练方法

01 焦点

02 前沿求索

03 文化故事

04 媒体报道

数据驱动模型提升电动汽车电池退化预测准确率

在全球向新能源转型的浪潮下，电动汽车的普及率正不断提升。然而，在享受电动汽车便利性的同时，你是否也在担忧电池的续航问题？电池的性能和寿命以及相应的监测、维护、回收等相关问题也同样困扰着电动汽车生产企业。而且如果废旧电池在回收、拆解和再利用的过程中处理不当，可能会对环境造成二次污染。

为了更有效地实现动力电池性能和寿命的精准预测，以及相应的对废旧动力电池的绿色回收和高效重复利用，微软亚洲研究院联手日产汽车针对电池退化问题展开了研究。基于日产汽车特有的电池数据，双方共同开发了一种全新的机器学习预测方法。通过挖掘电池结构的高级特征，该方法将电池退化预测准确率的平均误差控制在0.0094，为日产汽车的电池高效回收提供了有力依据。

近年来，碳排放问题日益严重，对全球社会的可持续发展构成了重大威胁。为了缓解地球面临的危机，全球各界都在积极实施碳减排政策，以实现长期的碳中和目标。而动力电池的回收与再利用，是实现这一目标的关键之一。通过对电池健康状态（SoH）进行评估，然后修复或重组，这些电池可以在小型电动汽车、储能系统和智能微电网等场景中被再次利用，从而延长其使用寿命并充分挖掘其剩余价值。

然而，准确评估电池的剩余价值并非易事。为了解决这一问题，日产汽车与微软亚洲研究院携手合作，共同探索解决方案。

从小处着手，迈向碳中和

作为推出全球首款量产纯电动车型的企业，日产汽车一直积极投身于碳减排行动中。2021年，日产宣布了碳中和目标：计划到2050年，实现汽车全生命周期的碳中和。电池的管理与创新对于日产汽车实现2050年碳中和目标至关重要，电池回收则是实现这一目标的关键组成部分。

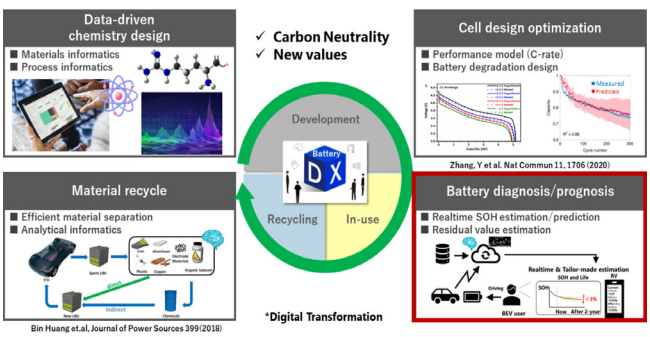


图 1：日产汽车电池生态创新所面临的挑战

日产电动汽车系统研究所负责人大間敦史指出，目前电动汽车和电池的平均生命周期大约为10年，其中材料开采和制造过程产生的二氧化碳排放量占其全生命周期碳排放量的50%。日产汽车的目标是通过将电动汽车和电池的生命周期延长至15年以上，以减少二氧化碳的排放。为了实现这一目标，日产汽车希望利用人工智能和大数据等前沿技术推动电池和电动汽车设计开发的创新。

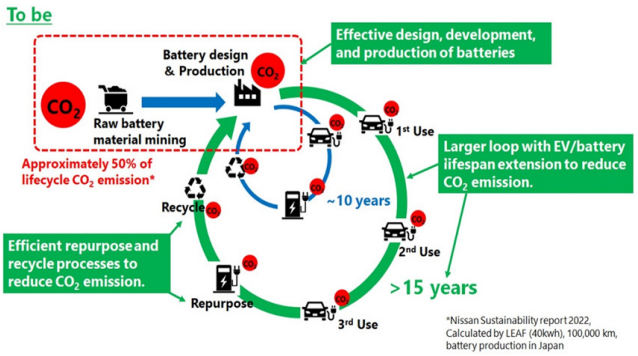


图 2：日产汽车在电动汽车全生命周期的碳减排愿景

携手合作促进电动汽车碳减排

自2020年微软宣布可持续发展承诺及实施计划以来，微软亚洲研究院也积极展开行动，致力于通过跨学科研究以及与相关领域的行业合作，应对可持续发展的挑战。此前，微软亚洲研究院已经开源了可用于电池性能分析和预测的一站式机器学习工具 BatteryML，并持续研究预测和管理电池健康及寿命的方法。

共同的碳中和愿景和对锂离子电池性能预测研究的一致追求，促成了日产汽车与微软亚洲研究院的合作。双方致力于通过深入电池性能退化研究来提升锂离子电池的性能预测，以实现碳中和目标。

大間敦史，
日产汽车电动汽车系统研究所负责人

“通过与微软亚洲研究院合作，我们一起研发了创新性的电池退化预测方法，提升了电池回收的效率并促进了资源的再利用。这是日产汽车实现长期碳中和目标的关键一步，我们正在以‘大处着眼，小处着手’的策略向着目标前进。”

创新方法实现更快、更准确的电池预测

电池健康状态是电池有效回收的关键。不过，可用容量不能完全代表电池健康状态，更重要的因素包括电池使用寿命内化学物质的完整性，如锂、钴和镍的含量。传统上，电池退化预测依赖基于化学、电化学和机械原理的数学模型。但这种方法需要通过重复实验来不断调整参数，而每次实验都要对电池进行拆解和化验分析，这可能会耗费长达半年甚至一年的时间。而且，只要电池化学成分发生变化，就需要进一步的实验并调整参数。为了解决这些问题，日产汽车希望借助机器学习技术，基于外部信号来预测电池健康状态，以最大限度地减少对物理实验的需求。

然而采用机器学习方法预测电池性能面临两个挑战。一是由于电池充放电周期长，难以收集到足够的数据。二是因为电池运行条件和外部因素大相径庭，导致信号采集非常复杂，很多信号虽然与电池容量相关，但又不能直接反映电池健康状态。

为了排除干扰噪音，找到准确反映电池内部状态的信号规律，微软亚洲研究院的研究员们设计了一系列特殊特征，来分析在不同电压和电流条件下锂离子电池内部的化学变化。通过将这些关键特征与日产汽车的真实数据结合，以此提升机器学习模型的预测精度。

郑顺，
微软亚洲研究院高级研究员

“我们发现学术公开数据集与企业真实数据之间存在较大的差异。由于企业的数据模式、测试条件和预测目标等等都与学术数据不同，所以现有学术论文中的方法难以直接适配于企业场景。只有深入到企业的真实场景中，发掘行业独特的数据特色，并与前沿人工智能方法相结合，才能开发出高效实用的电池模型。”

数据驱动模型的准确性在仿真数据中提高了80%，在实验数据上提升超30%

以预测电池的长程健康状态为例，研究员们首先重定义了整个特征空间。这个空间包含了对电池退化原因的统一表征。如图3所示，研究员们采用高阶特征工程分析了充电和放电循环期间电压-容量曲线退化模式产生的各种特征。研究员们通过区分高电压和低电压间隔之间的信息，包括作为电池健康有效指标的一阶和高阶差异，增强了模型的预测能力，并提供了对电池性能和寿命的深入洞察。

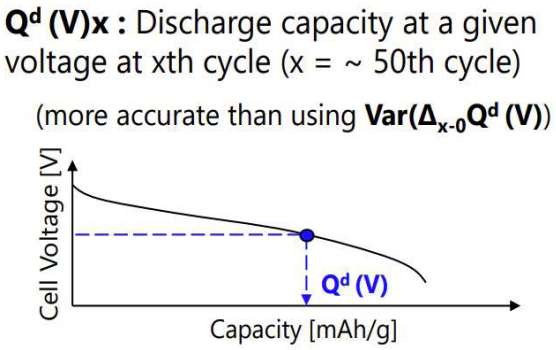


图 3：电压随放电容量变化的特征工程

与学术界 SOTA 的电池预测方法相比，微软亚洲研究院与日产汽车的联合创新方法在使用日产汽车的仿真数据进行测试时，准确性提升了约80%；而在使用实验数据时，准确性提升超过30%。如图4所示，新的方法仅使用电池的前50个循环数据就能预测到200个循环时电池的健康状态，平均绝对误差（MAE）控制在了0.0094。这一结果表明数据驱动模型在电池健康预测上潜力巨大，非常有助于更加高效、准确、智能的电池监测与管理。

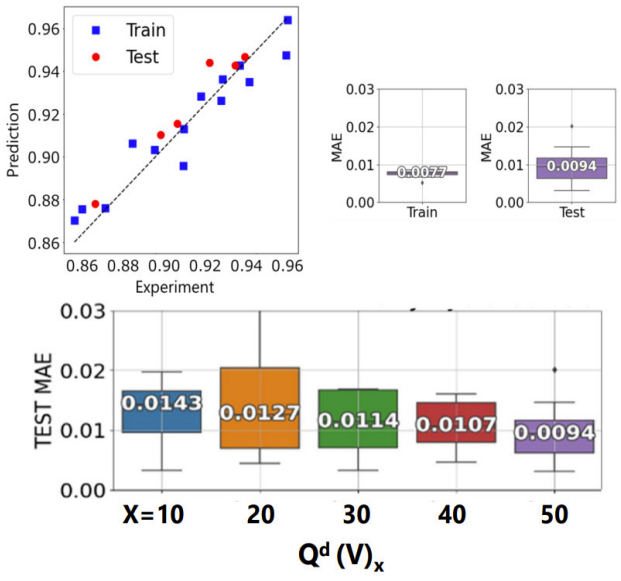


图 4：在使用 Qd(V)50 作为衡量标准进行测试时，模型在预测电池在第 200 个周期时的健康状态方面的平均绝对误差（MAE）为 0.0094

利用数据驱动的方法，研究员们还发现，电压在3.9伏时的电池状态与 NMC（镍锰钴氧化物）晶体结构的变化（从 M 相到 H2 相）呈现出一致性。这一特征与电化学研究领域的发现一致，证

实了数据驱动方法在识别影响电池退化的关键特征方面具有重要的现实意义。

文贞媛，
日产汽车电动汽车系统研究所工程师

“这项研究从两方面延长了动力电池的寿命。一是通过判断动力电池的剩余寿命，更好地实现了电池的再利用价值；二是，找到了一条更有效的退役电池回收策略。这一独特的方法不仅能预测电池健康状态，还能预测正极（NMC）健康状态，提高了预测模型的可靠性。更令人惊讶的是，数据驱动的正极（NMC）电池健康状态预测模型对特定电压（3.9V）表现出的高灵敏度，与基于物理模型的预测结果相吻合。与微软亚洲研究院的合作表明，人工智能技术可以应用于电池的制造阶段，包括材料选择和流程优化等环节。”

人工智能助力可持续发展还将大有可为

日产汽车和微软亚洲研究院的合作成果突显了机器学习、深度学习等人工智能技术在电动汽车领域应用的巨大潜力。除了电池回收的健康状态预测，人工智能还可以为驾驶员提供电池寿命预测服务，优化驾驶体验，实现更智能的驾驶。不仅如此，人工智能在新材料和新物质的发现上也将大有可为，有望推动电池和电动汽车技术的进一步创新。

边江，
微软亚洲研究院资深首席研究员

“锂电池目前仍然存在一些问题亟待解决，我们需要一种具有更高能量密度、更安全、循环寿命更长且对环境影响最小的电池。通过与日产汽车的深入合作，我们看到人工智能在电动汽车领域还有更多的应用可能，例如，通过电池材料组合优化提升电池性能、帮助发现新材料、电池电极工艺优化等。未来，我们希望与更多行业伙伴合作，释放人工智能更大的行业潜能。”

基于合作取得的初步成果，日产汽车与微软亚洲研究院计划进一步深化、拓展双方的合作关系，在推动技术进步的同时，助力双方在实现可持续发展和环境保护方面迈出更坚实的步伐。

守护记忆：多模态大模型为认知障碍患者带来全新的训练方法

阿尔茨海默病作为最常见的认知障碍，一直以来都备受关注。研究证明，科学的认知训练可以起到对该疾病的预防和延缓。为此，微软亚洲研究院与上海市精神卫生中心携手展开联合研究，基于微软 Azure OpenAI 服务中的多模态大模型，开发了利用多种模态数据（如语音、文字、图片、音乐等信息）的个性化认知训练框架，为认知障碍患者的认知训练带来了新的可能。

随着全球人口老龄化的加剧，以阿尔茨海默病为代表的认知症被认为是二十一世纪最大的健康危机之一。近期，微软亚洲研究院与上海市精神卫生中心展开联合研究，借助微软 Azure OpenAI 服务中的多模态大模型以及智能代理技术，开发了个性化认知训练框架“忆我”（ReMe），扩展了自动化认知训练的训练范围，为数字化认知训练提供了新方法，有望帮助延缓认知下降。这项创新工具将助力推进认知训练研究，为各类认知障碍，包括轻度认知障碍的早期预防和非药物干预提供新的方法。

多模态大模型带来更全面的认知训练方法

认知症，也称认知障碍，是一类影响思维、记忆、注意力、理解、判断力和语言等认知功能的疾病，包括阿尔茨海默病、帕金森病、路易体痴呆、脑血管性认知障碍等。认知症的发展是一个渐进且不可逆的过程，虽然目前临床上还没有完全能够治愈该疾病的药物，但及早诊断和提前干预有助于延缓认知功能的衰退，而认知训练已被证实是缓解病程进展的有效手段。

伴随多模态大模型的不断发展，其新能力也在不断涌现，如提供即时的多模态分析、实现基于世界知识的开放式问答、提供充满情感的语音交互、整合处理多种感知数据等。微软亚洲研究院的研究员们由此提出了一个设想：多模态大模型或许能在认知训练方面发挥重要作用。

带着这一想法和诸多疑问，微软亚洲研究院与上海市精神卫生中心展开了深入探讨与合作。上海市精神卫生中心（俗称"600号"）系上海交通大学医学院附属医院，有着雄厚的技术实力和丰富的临床经验。开发个性化认知训练框架"忆我"主要与老年精神科团队合作，该科室作为上海交通大学阿尔茨海默病诊治中心，是全国最早从事阿尔茨海默研究的科室，2022年被评为国家核心高级认知障碍诊疗中心（全国唯一获批精神专科医院），并获得"全国优秀记忆门诊"、"记忆门诊培训基地"称号。围绕阿尔茨海默病发病机制这一课题，该科室展开了卓有成效的工作，形成了一定的研究特色。团队对于认知障碍的诊治，倡导"从不太早，永不言迟"，并创新地提出"老年认知障碍防治上海防治模式"，希望为健康老龄提供切实有用的帮助。

通过与中心专家的交流，研究员们了解到，当前临床上的认知训练主要依靠照护者或治疗师自行开展，或使用软件驱动的数字疗法开展。前者依赖大量人力，成本高昂，给照护者带来了沉重的负担；后者则往往类似于标准化考试，缺乏灵活性，患者仅能解答预设的谜题，与提高认知所需的记忆训练相去甚远。此外，现有的数字疗法在软件应用的便捷性、互动性和直观性方面也存在一定不足，影响了患者的依从性（病人按照医生规定进行治疗的行为）。就认知障碍患者而言，持之以恒的训练对于缓解症状至关重要。

上海市精神卫生中心老年科副主任医师岳玲指出："认知障碍影响的记忆类型多样，包括工作记忆、情景记忆和语义记忆等，但现有训练任务的设计较为单一，限制了训练内容的广度，很多老人也曾反映游戏内容'枯燥'、'没意思'。在多模态大模型出现之前，软件驱动的认知训练主要集中于工作记忆训练，而对于与患者日常生活密切相关的情景记忆和语义记忆训练则鲜有涉及。特别是对于阿尔茨海默病患者，最早受损的往往是与自我相关的记忆。多模态大模型的智能化和个性化特点，为认知训练提供了新的可能性，使其更贴近日常生活。"

涨知识：

- 工作记忆：指进行临时信息处理的能力，如心算或复述新闻内容。当这部分记忆受损时，患者可能在对话中忘记早先提及的信息。
- 情景记忆：涉及自我的记忆，例如记得早餐吃了什么，到过某地或与某人的对话内容等。
- 语义记忆：也称为知识记忆，涵盖广泛的事实性知识，例如知道法国的首都是巴黎。

基于微软Azure OpenAI服务，让认知训练既通用又个性化

综合上海市精神卫生中心专家的建议与患者的需求，微软亚洲研究院的研究员们从易用性、界面友好性、功能专业性和入门难易度等多个维度出发，开发了个性化认知训练框架"忆我"（ReMe）。该工具以微软 Azure OpenAI 服务为基础模型，具备即时交互响应功能，支持文字、图像、语音等多种模态的输入输出，以对话机器人的形式为用户提供全新的认知训练体验。

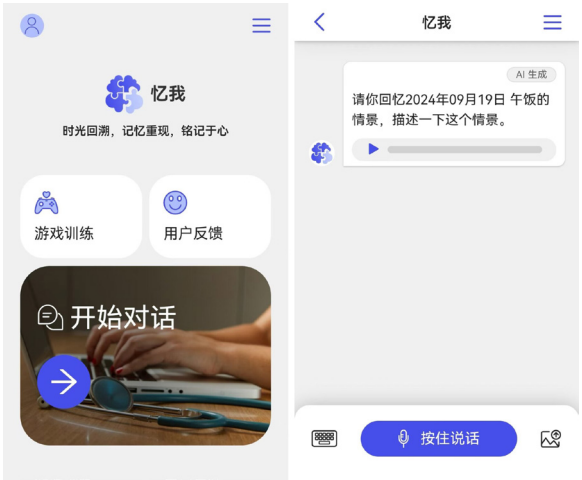


图 1:"忆我"应用界面

"将多模态大模型的通用能力与专业知识相结合，应用到认知训练领域，打造一个专业的领域智能体，是我们面临的一个挑战。"微软亚洲研究院高级研究员王子龙表示，"此前，我们已经开展了一系列研究，探索如何通过思维链（Chain of Thoughts）、检索增强生成（Retrieval Augmented Generation, RAG）等技术，使多模态大模型能够更好地整合专有领域知识，从而优化智能代理的行为逻辑和性能。这些研究确保了智能代理能够使用更新和更全面的领域知识，保持其在特定领域中发挥最前沿的能力，更好地进行交互和训练任务。"

从使用流程和功能来看，个性化认知训练框架"忆我"首先通过手机或可穿戴设备上传个人记忆内容，随后采用更贴近生活的方式，进行个体化的情景记忆或世界知识的开放性记忆训练。研

究员们还提供了一个训练框架，让该领域的医生等研究者可以利用内置的认知游戏模板快速创建个性化的训练游戏。除此之外，该工具还包括交互式的评估，以便追踪患者的认知水平，并根据需要调整训练任务。

不仅如此，微软亚洲研究院与上海市精神卫生中心还在不断向该训练框架扩展更多训练功能和任务，除了开放性的和个体化的训练任务外，通过语音对话的自然交互体验，也兼容并实现已被验证有效的成熟认知训练任务。例如锻炼思维灵活性的颜色识别、锻炼工作记忆的数学运算等。同时，微软亚洲研究院也在尝试基于传感器技术，通过可穿戴设备以机会感知（opportunistic sensing）的方式记录日常生活中的重要节点和事件，促进更便捷的个性化个人记忆训练。

推进临床证据获得，释放人工智能在医疗领域的更大价值

“在此次合作中，微软亚洲研究院所展现出的开放合作精神、专业技术实力和快速创新能力，给我们留下了深刻的印象。这些特质对于推动技术在医疗健康领域的应用至关重要。接下来，上海市精神卫生中心计划以此框架为基础，设计认知训练的新方法，开展一系列严谨的临床实验，验证这种个性化认知训练框架的有效性，以获得更多的临床证据。”岳玲医生表示。

随着证据的逐步积累，微软亚洲研究院将持续对个性化认知训练框架“忆我”进行改进和迭代，期望该工具能够逐渐扩展应用范围，最终能够在社区和家庭环境中帮助提升认知障碍患者的健康水平，减轻看护人员和家属的负担。

除了认知训练，基于多模态大模型的代理干预技术未来有望应用于更广泛的领域，例如情绪调节、习惯改善以及孤独症等辅助干预。微软亚洲研究院副院长邱锂力博士表示，微软亚洲研究院将继续通过研究与创新，推动AI技术在临床决策、疾病预测、药物发现等医疗健康领域的应用，进而帮助医疗机构提高医疗服务的质量和效率，为患者带来更精准、更个性化的治疗方案。

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。

微软亚洲研究院多项创新技术，弥合大模型低比特量化与终端部署间鸿沟

在人工智能领域，模型参数的增多往往意味着性能的提升。但随着模型规模的扩大，其对终端设备的算力与内存需求也日益增加。低比特量化技术，由于可以大幅降低存储和计算成本并提升推理效率，已成为实现大模型在资源受限设备上高效运行的关键技术之一。然而，如果硬件设备不支持低比特量化后的数据模式，那么低比特量化的优势将无法发挥。

为了解决这一问题，微软亚洲研究院推出了全新的数据编译器 Ladder 和算法 T-MAC，使当前只支持对称精度计算的硬件能够直接运行混合精度矩阵乘法。测试结果表明，Ladder 在支持 GPU 原本不支持的自定义数据类型方面，最高提速可达14.6倍；T-MAC 在搭载了最新高通 Snapdragon X Elite 芯片组的 Surface AI PC 上，使 CPU 上运行的大模型吞吐率比专用加速器 NPU 快两倍。此外，研究员们还设计了 LUT Tensor Core 硬件架构，这种精简设计使硬件能够直接支持各种低比特混合精度计算，为AI硬件设计提供了新思路。

大模型已经越来越多地被部署在智能手机、笔记本电脑、机器人等端侧设备上，以提供先进的智能及实时响应服务。但包含上亿参数的大模型对终端设备的内存和计算能力提出了极高的要求，也因此限制了它们的广泛应用。低比特量化技术因其能显著压缩模型规模，降低对计算资源的需求，成为了大模型在端侧部署和实现高效推理的有效手段。

随着低比特量化技术的发展，数据类型日益多样化，如 int4、int2、int1 等低比特数据，使得大模型在推理中越来越多地采用低比特权重和高比特权重计算的混合精度矩阵乘法（mixed-precision matrix multiplication，mpGEMM）。然而，现有的 CPU、GPU 等硬件计算单元通常只支持对称计算模式，并不兼容这种混合精度的矩阵乘法。

混合精度矩阵乘法与传统的矩阵乘法有何不同？

在传统的矩阵乘法中，参与运算的两端数值是对称的，例如 FP16*FP16、int8*int8。但大模型的低比特量化打破了这种对称性，使乘法的一端是高比特，另一端是低比特，例如在 1-bit 的 BitNet 模型中实现的 int8*int1 或 int8*int2，以及浮点数与整数的混合乘法 FP16*int4。

为了充分发挥低比特量化的优势，让硬件设备能够直接支持混合精度矩阵乘法，确保大模型在端侧设备上的高速有效运行，微软亚洲研究院的研究员们针对现有 CPU、GPU 计算算子和硬件架构进行创新：

- 推出了数据类型编译器 Ladder，支持各种低精度数据类型的表达和相互转换，将硬件不支持的数据类型无损转换为硬件支持的数据类型指令，在传统计算模式下，使得硬件能够支持混合精度的 DNN（深度神经网络）计算；
- 研发了全新算法 T-MAC，基于查找表（Lookup Table，LUT）的方法，实现了硬件对混合精度矩阵乘法的直接支持，软件层面在 CPU 上的计算相比传统计算模式取得了更好的加速；
- 提出了新的硬件架构 LUT Tensor Core，为下一代人工智能硬件设计打开了新思路。

Ladder：自定义数据类型无损转换成硬件支持的数据类型

当前，前沿加速器正在将更低比特的计算单元，如 FP32、FP16，甚至 FP8 的运算集成到新一代的架构中。然而，受限于芯片面积和高昂的硬件成本，每个加速器只能为标准的数据类型提供有限类型的计算单元，比如 NVIDIA V100 TENSOR CORE GPU 仅支持 FP16，而 A100 虽然加入了对 int2、int4、int8 的支持，但并未涵盖更新的 FP8 或 OCP-MXFP 等数据格式。此外，大模型的快速迭代与硬件升级的缓慢步伐之间存在差距，导致许多新数据类型无法得到硬件支持，进而影响大模型的加速和运行。

微软亚洲研究院的研究员们发现，尽管硬件加速器缺乏针对自定义数据类型的计算指令，但其内存系统可以将它们转换为固定位宽的不透明数据块来存储任意数据类型。同时，大多数自定义数据类型可以无损地转换为现有硬件计算单元支持的更多位的标准数据类型。例如，NF4 张量可以转换成 FP16 或 FP32 以执行浮点运算。

基于这些发现，研究员们提出了一种通过分离数据存储和计算来支持所有自定义数据类型的方法，并研发了数据编译器 Ladder，以弥合不断出现的自定义数据类型与当前硬件支持的固

有精度格式之间的差距。

Ladder 定义了一套数据类型系统，包括数据类型之间无损转换的抽象，它能够表示算法和硬件支持的各种数据类型，并定义了数据类型之间的转换规则。当处理低比特算法应用时，Ladder 通过一系列优化，将低比特数据转译成当前硬件上最高效的执行格式，包括对计算和存储的优化——将算法映射到匹配的计算指令，并将不同格式的数据存储到不同级别的存储单元中，以实现最高效的运算。

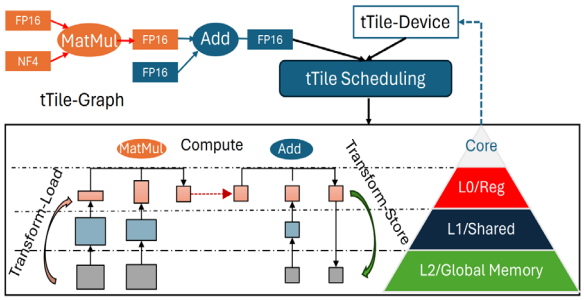


图1：Ladder 的系统架构

在 NVIDIA A100、NVIDIA V100、NVIDIA RTX A6000、NVIDIA RTX 4090 和 AMD Instinct MI250 GPU 上运行的 DNN 推理性能评估显示，Ladder 在原生支持数据类型上超越了现有最先进的 DNN 编译器，并且在支持 GPU 原本不支持的自定义数据类型方面表现出色，最高提速可达14.6倍。

Ladder 是首个在现代硬件加速器上运行 DNN 时，可以系统性地支持以自定义数据类型表示低比特精度数据的系统。这为模型研究者提供了更灵活的数据类型优化方法，同时也让硬件架构开发者在不改变硬件的情况下，支持更广泛的数据类型。

T-MAC：无需乘法的通用低比特混合精度矩阵乘计算

为了让现有硬件设备支持不同的数据模式和混合精度矩阵乘法，在端侧部署大模型时，常见的做法是对低比特模型进行反量化。然而，这种方法存在两大问题：首先，从性能角度来看，反量化过程中的转换开销可能会抵消低比特量化带来的性能提升；其次，从开发角度来看，开发者需要针对不同的混合精度重新设计数据布局和计算内核。微软亚洲研究院的研究员们认为，在设备上部署低比特量化的大模型，关键在于如何基于低比特的特点来突破传统矩阵乘法的实现。

为此，研究员们从系统和算法层面提出了一种基于查找表（LUT，Look-Up Table）的方法 T-MAC，帮助低比特量化的大模型在 CPU 上实现高效推理。T-MAC 的核心思想在于利用混合精度矩阵乘法的一端为极低比特（如1比特或2比特）的特点。它们的输出结果只有2的1次方和2的2次方种可能，这些较少的输出结果完全可以提前计算并存储在表中，在运算时，只需从表中读取结果，避免了重复计算，大幅减少了乘法和加法的运算次数。

具体而言，T-MAC 将传统的以数据类型为中心的乘法转变为基于位的查找表操作，实现了一种统一且可扩展的混合精度矩阵乘法解决方案，减小了表的大小并使其停留在最快的内存单元中，降低了随机访问表的成本。这一创新为在资源受限的边缘设备上部署低比特量化大模型铺平了道路。

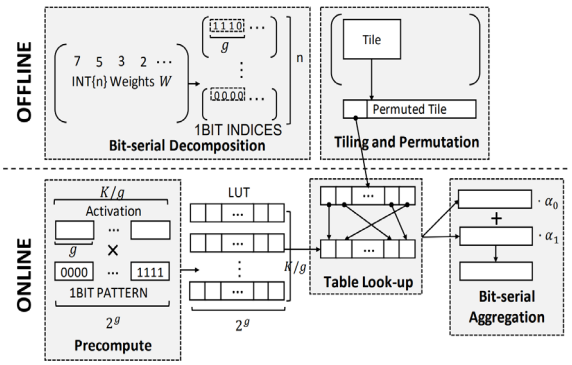


图2：T-MAC 示意图

在针对低比特量化的 Llama 和1比特的 BitNet 大语言模型的测试中，T-MAC 展现出了显著的性能优势。在搭载了最新高通 Snapdragon X Elite 芯片组的 Surface Laptop 7 上，T-MAC 让 3B BitNet-b1.58 模型的生成速率达到每秒48个 token，2bit 7B Llama 模型的生成速率达到每秒30个 token，4bit 7B Llama 模型的生成速率可达每秒20个 token，这些速率均远超人类的平均阅读速度。与原始的 Llama.cpp 框架相比，其提升了4至5倍，甚至比专用的 NPU 加速器还快两倍。

即使是在性能较低的设备上，如 Raspberry Pi（树莓派）5，T-MAC 也能使 3B BitNet-b1.58 模型达到每秒11个 token 的生成速率。T-MAC 还具有显著的功耗优势，在资源受限的设备上可以达到相同的生成速率，而它所需的核心数仅为原始 Llama.cpp 的1/4至1/6。

这些结果表明，T-MAC 提供了一种实用的解决方案，使得在使用通用 CPU 的边缘设备上部署大语言模型更为高效，且无需依赖 GPU，让大模型在资源受限的设备上也能高效运行，从而推动大模型在更广泛的场景中的应用。

LUT Tensor Core：推动下一代硬件加速器原生支持混合精度矩阵乘法

T-MAC 和 Ladder 都是在现有 CPU 和 GPU 架构上，实现对混合精度矩阵乘法的优化支持。尽管这些软件层面的创新显著提升了计算效率，但它们在效率上仍无法与能够直接实现一个专门

查找表的硬件加速器相比。研究员们认为，最理想的方法是重新设计硬件加速器，让 CPU、GPU 等能够原生支持混合精度矩阵乘法，但这一目标面临三大挑战：

- 效率：设计和实现方式必须具有成本效益，通过优化芯片的利用面积，最大限度地提高低比特数据的计算效益。
- 灵活性：由于不同的模型和场景需要不同的权重和激活精度，因此硬件中的混合精度矩阵乘法设计必须能够处理各种权重精度（如 int4/2/1）和激活精度（如 FP16/8、int8）及其组合。
- 兼容性：新设计必须与现有的 GPU 架构和软件生态系统无缝集成，以加速新技术的应用。

为了应对这些挑战，微软亚洲研究院的研究员们设计了 LUT Tensor Core，这是一种利用查找表直接执行混合精度矩阵乘法的 GPU Tensor Core 微架构。一方面，基于查找表的设计将乘法运算简化为表预计算操作，可直接在表中查找结果，提高计算效率。另一方面，这种方法也简化了对硬件的需求，它只需用于表存储的寄存器和用于查找的多路选择器，无需乘法器和加法器。同时，LUT Tensor Core 通过比特串行设计实现了权重精度的灵活性，并利用表量化实现了激活精度的灵活性。

此外，为了与现有 GPU 微架构和软件堆栈集成，研究员们扩展了 GPU 中现有的 MMA 指令集，加入了一组 LMMA 指令，并设计了一个类似于 cuBLAS 的软件堆栈，用于集成到现有的 DNN 框架中。研究员们还设计了一个编译器，用于在具有 LUT Tensor Core 的 GPU 上进行端到端的执行计划。这些创新方法可以让 LUT Tensor Core 被无缝、快速地采用。

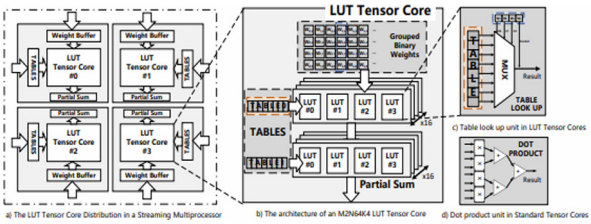


图3：LUT Tensor Core 微架构概述

在 Llama 和 BitNet 模型上的测试显示，LUT Tensor Core 可以提供高达6.93倍的推理速度，且只占传统 Tensor Core 面积的38.7%。在几乎相同的模型精度下，这相当于20.7倍的计算密度和19.1倍的能效提升。随着人工智能大模型规模和复杂性的不断增长，LUT Tensor Core 有助于进一步释放低比特大语言模型的潜力，推动人工智能在新场景中的应用。

“查找表方法引领了计算范式的转变。在过去，我们依赖于矩阵乘法和累加运算，而在大模型时代，得益于低比特量化技术，查找表方法将成为主流。相较于传统的浮点运算或矩阵乘法，查找表方法在计算上更轻便高效，而且在硬件层面上更易于扩展，能够实现更高的晶体管密度，在单位芯片面积上提供更大的吞吐量，从而推动硬件架构的革新。”微软亚洲研究院首席研究员曹婷表示。

低比特量化的长尾效应：为具身智能带来新可能

低比特量化技术不仅优化了大模型在端侧设备上的运行效率，还通过减少单个参数的“体积”，为模型参数的扩展（Scale up）提供了新的空间。这种参数扩展能力，使模型拥有了更强的灵活性和表达能力，正如 BitNet 模型所展示的，从低比特模型出发，逐步扩展至更大规模的训练。

微软亚洲研究院的 T-MAC、Ladder 和 LUT Tensor Core 等创新技术，为各种低比特量化大模型提供了高效能的运行方案，使得这些模型能够在各种设备上高效运行，并推动科研人员从低比特角度设计和优化大模型。其中部分技术已经在微软必应（Bing）搜索及其广告业务等搜索大模型中发挥作用。随着对内存和计算资源的降低，低比特大模型在机器人等具身智能系统上的部署也将成为可能，可以使这些设备更好地实现与环境的动态感知和实时交互。

目前，T-MAC 和 Ladder 已经在 GitHub 上开源，欢迎相关研发人员测试应用，与微软亚洲研究院共同探索人工智能技术的更多可能。

相关链接：

Ladder 论文链接：
https://www.usenix.org/conference/osdi24/presentation/wang-lei

BitBLAS/Ladder GitHub 链接：
https://github.com/microsoft/BitBLAS

T-MAC 论文链接：
https://arxiv.org/abs/2407.00088

T-MAC GitHub 链接：
https://github.com/microsoft/T-MAC

LUT Tensor Core 论文链接：
https://arxiv.org/abs/2408.06003

BitDistiller 论文链接：
https://arxiv.org/abs/2402.10631

BitDistiller GitHub 链接：
https://github.com/DD-DuDa/BitDistiller

脑启发设计：人工智能的进化之路

你可以用左手（不常用的那只手）的小指与食指拿起一件物品么？试完你是不是发现自己竟然可以毫不费力地用自己不常用的手中，两根使用频率相对较低的手指，做一个不常做的动作。这就是人类大脑不可思议之处——无需经过特别的训练，大脑就能够在短时间内以低功耗的方式控制身体完成各种复杂行为，甚至是全新的动作。相比之下，人工智能虽然是人类智慧的产物，但在很多方面还远不及人类大脑。

为此，微软亚洲研究院（上海）团队的研究员们从理解大脑结构与活动中获得灵感，开发了一系列涵盖大脑学习、计算过程不同层级的创新技术，包括模仿脑神经回路连接方式，可高效处理众多任务的 CircuitNet 神经回路网络；可应用于时间序列预测，更适配神经拟态芯片的新型 SNN（脉冲神经网络）框架和策略；以及可为具身智能提供理论指导的贝叶斯行为框架。这些探索为未来的人工智能技术发展提供了新的可能。

从能耗的角度来看，人类大脑只需要大约20瓦的功率即可维持运转，这约等于一个节能灯泡的功耗。但随着人工智能大模型参数和规模的增大，其能源需求远高于传统的数据中心。主流的大语言模型训练过程预计会消耗上千兆瓦的电力，相当于数百个家庭一年的用电量。这种能源消耗的增长趋势显然不利于人工智能技术的可持续发展。那么如何通过新的处理机制解决能耗问题，就成了信息科学领域一个紧迫且前沿的挑战。

《千脑智能》一书为我们提供了启示："要创造出真正智能的机器，我们首先需要对大脑进行逆向工程。我们研究大脑，不仅是为了理解它的工作原理，更是为了探索智能的本质。"

其实，人工智能本身就是人类对大脑探索的产物，在计算机诞生之初，人们就已经利用神经连接模式+数字计算的方式模拟大脑。但受限于当时的算力和人们对大脑粗浅的认知，人工智能发展非常缓慢，甚至一度被束之高阁。近几十年来，随着神经科学家对大脑结构的深入理解和计算资源及相关技术的增强，以脑启发为核心的"人工智能文艺复兴"也掀起了新一轮热潮，促使科研人员重新定位大脑机制对人工智能的作用。

来自微软亚洲研究院（上海）的研究员们跨越计算机和脑科学专业知识，深入理解大脑的结构与行为活动，针对大脑学习和计算过程，从神经元、网络层和更高级别的系统层出发，分别设计研发了高性能的脉冲神经网络（SNN）、参数效率更高的回路神经网络（CircuitNet），以及提升决策效率的贝叶斯行为框架，促进了人工智能网络向着更低功耗、更高效率、更好性能的方向良性发展，同时也为具身智能发展提供了理论和方法。

CircuitNet：模拟大脑神经元连接，实现更低功耗与更高性能

人工神经网络（ANN）已经被广泛应用于人工智能的众多领域，包括自然语言处理、机器学习、语音识别和控制系统等。这些应用的成功，很大程度上得益于它们对大脑神经元工作模式的模仿。神经元是大脑最基本的单元，它们之间通过复杂的连接模式相互作用来传递和处理信息。但早期的人工神经网络设计相对简单，仅能模拟一两种连接模式。

随着神经科学的发展，人们发现大脑神经元的连接方式多种多样，其中有四种常见模式：前馈激励和抑制、反馈抑制、侧抑制和相互抑制。然而，现有的许多人工神经网络，如具有残差连接的网络，只能模拟前馈激励和抑制模式。即便是能够模拟循环模式的递归神经网络（RNN），在信息传入前也无法处理上游神经元间的复杂相互作用，从而影响了神经网络在不同机器学习任务中的表现。

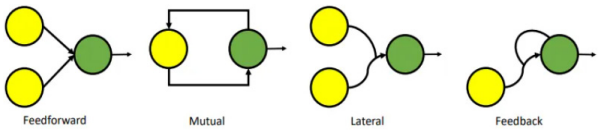


图1：大脑神经元的四种连接模式

生物神经网络与人工神经网络的整体连接模式也大不相同。生物神经网络的一个显著特点是局部密集连接与全局稀疏连接的结合。尽管单个神经元可以有数千个突触，但它们大多数位于一个脑区内，形成针对特定任务的功能集群。只有少数突触作为不同脑区之间的桥梁，延伸到其它功能集群，而人工神经网络通常不具备这样的特性。此外，人工神经网络中的许多参数也被证实是冗余的，增加了网络的复杂性。

基于对大脑神经连接的新理解，研究员们提出了新的回路神经网络 CircuitNet，它能够模拟包括反馈和侧向模式在内的多种神经元连接模式。CircuitNet 的设计还借鉴了大脑神经元局部密集和全局稀疏连接的特性，通过不同电路模式单元（Circuit Motif Unit，CMU）的输入端和输出端的稀疏连接，实现了信号在不同 CMU 之间的多轮传输。

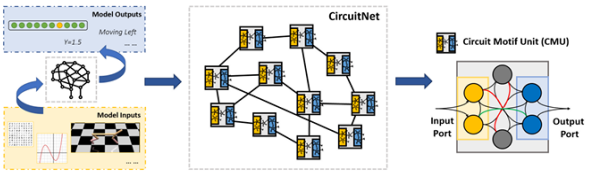


图2：CircuitNet 架构

实验结果表明，CircuitNet 在函数逼近、强化学习、图像分类和时间序列预测等任务中的表现超越了当前流行的神经网络架构。而且，在各种类型的任务中，CircuitNet 在达到与其它神经网络相同性能的同时，具有相当或更少的参数，展示了其在机器学习任务中的有效性和强大的泛化能力。

让SNN更适用于时间序列预测任务的新框架

脉冲神经网络（SNN）因其能效高、事件驱动范式和生物学上的合理性，正逐渐受到业内的重视。SNN 的设计灵感来源于生物神经网络中神经元间的信息传递方式——神经元不是在每次迭代传播中都被激活，只有膜电位达到特定阈值时才被激活，进行信号传递。这种事件驱动机制使得 SNN 只在接收到有效刺激时才进行信息处理，从而避免了无效计算，极大地提高了运算效率和能效比。

但研究员们发现，现有的 SNN 设计大多聚焦于其离散的事件驱动特性，有的会忽略其时间属性，有的则为了适应事件驱动范式过程，过度简化序列数据模式。这些方法虽然让 SNN 在图像分类、文本分类和序列图像分类任务上实现与人工神经网络接近的性能，但并未充分发挥 SNN 在处理时间信号方面的潜力。

研究员们认为，时间序列预测是 SNN 一个理想的应用场景。作为现实数据分析的重要组成部分，时间序列预测广泛应用于交通、能源、医疗等领域，旨在基于按时间顺序排列的历史数据来预测未来。但是，将 SNN 应用于时间序列预测还面临两大挑战：

- SNN 中脉冲值的离散特性与时间序列数据的浮点属性之间存在巨大的差异，需要一种有效的机制来减少在将浮点值转换为脉冲序列时的信息丢失和噪声。
- 如何选择用于时序数据的 SNN 标准化模型目前还缺少一个指导方针，进而加剧了任务的复杂性，这就需要对 SNN 架构及其参数进行深入探索，以适应不同时间序列数据的特定特征。

研究员们提出了一个用于时间序列预测任务的 SNN 框架。该框架充分利用了脉冲神经元在处理时间序列信息上的高效性，成功实现了时间序列数据与 SNN 之间的时间同步。研究员们还设计了两种编码层，可以将连续时间序列数据转换为有意义的脉冲序列。这之后，研究员们又利用多种脉冲化的时间序列模型对脉冲序列进行了建模，得到了最终的预测结果。

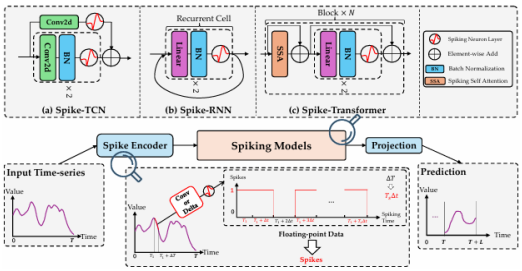


图3：SNN 用于时间序列预测的新框架

通过在多个时间序列预测基准集上的测试，研究员们证实了SNN方法在时间序列预测中的有效性。该方法不仅展现出与传统时间序列预测方法相媲美或更优的性能，且显著降低了能耗。

此外，在分析实验中，研究员们还展示了 SNN 如何捕获时间序列数据中的时间依赖性，并发现 SNN 确实能够模拟时间序列数据的内在动态。这项研究为 SNN 领域提供了一个既节能，又符合生物学原理的时间序列预测新方案。

大脑中枢模式发生器与位置编码双加持，让SNN序列预测更上一层楼

尽管 SNN 在多个领域取得了显著进展，但它们在适应不同类型任务时仍面临挑战。SNN 作为事件驱动的系统，缺乏有效机制来捕获索引信息、节奏模式和周期性数据，从而限制了它们处理自然语言和时间序列等数据模式的能力。而且，SNN 依赖于脉冲形式的通信，这使得并非所有适用于人工神经网络的深度学习技术都能直接迁移到 SNN 上。

为了克服这些限制，研究员们进一步从生物神经学机制中汲取灵感，基于人类大脑中枢模式发生器（Central Pattern Generator, CPG）和位置编码（Positional Encoding，PE）技术，开发了针对 SNN 的新型位置编码技术 CPG-PE。

中枢模式发生器（CPG）：在神经科学中，CPG 是一组能够在不需要节奏输入的情况下，产生有节奏的模式输出的神经元。这些神经回路位于脊髓和脑干中，负责产生控制运动、呼吸和咀嚼等重要活动的有节奏信号。

位置编码（PE）：PE 是人工神经网络中的一项关键技术，尤其在序列处理任务中尤为重要。通过为输入序列的每个元素赋予位置信息，PE 使神经网络能够识别序列中元素的顺序和相对位置。

CPG 和 PE 都能产生周期性输出，CPG 是相对于时间的输出，而 PE 则是相对于位置的输出。研究员们将两者类比，使 CPG-PE 可以编码时间或空间的位置信息，预测神经信号的来源或位置。

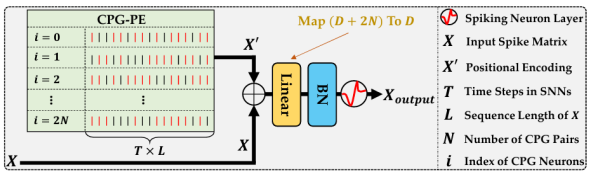


图4：CPG-PE 在 SNN 中的应用。X、X′ 和 Xoutput 是脉冲矩阵

在 Metr-la（洛杉矶高速公路平均交通速度数据）、Pems-bay（湾区平均交通速度数据）、Electricity（以千瓦时 kWh 测量的每小时电力消耗数据）和 Solar（太阳能发电数据）四个真实世界数据集上进行的时间序列预测实验表明，采用 CPG-PE 策略的 SNN 在时间序列分析方面显著优于没有 PE 特性的神经网络。同时，CPG-PE 可以无缝集成到任何能够处理序列的 SNN 中，理论上可以实现与 SNN 硬件的兼容，适配各类神经拟态芯片。

贝叶斯行为框架：为具身智能提供理论指导

在心理学和认知神经科学领域，以人类为代表的智能生物群体被认为会执行两类行为：习惯性行为和目标导向行为。习惯性行为是指为了最大化利益而自动执行的动作，无需意识思考或意图的参与，例如寻找食物和避免危险。目标导向行为是指为了实现特定目标而执行的动作，例如有计划地前往某个地点。传统上认为，在认知科学和机器学习中，习惯性行为和目标导向行为由两套独立的系统控制，因此在建模时，研究人员通常会为这两种行为设计独立的模型。

然而，微软亚洲研究院的研究员们认为，这两种系统应该更紧密地结合，实现协同学习和工作。尽管在大脑中这两种系统之间的相互作用尚未完全明了，但习惯性行为和目标导向行为共享着诸如脑干这样的下游神经回路。两种行为共享低级运动技能，且每个系统都可能利用对方学习到的高级动作。例如，习惯性行为虽然缺乏灵活性，但通过练习可以提供熟练的运动技能，这些技能可以被目标导向行为用于更复杂的任务规划。那么如何在保持两种行为差异的同时实现协同？

为此，研究员们提出了一个基于变分贝叶斯方法的理论框架——贝叶斯行为（Bayesian Behavior）框架，用于理解感知运动学习中的行为。其核心创新在于引入了一个贝叶斯“意图”（intention）变量，从而有效连接习惯性行为与目标导向行为。

习惯性行为由感官输入计算的意图先验分布驱动，无需具体目标。目标导向行为则由一个通过最小化变分自由能推断（active inference）的目标条件意图后验分布引导。

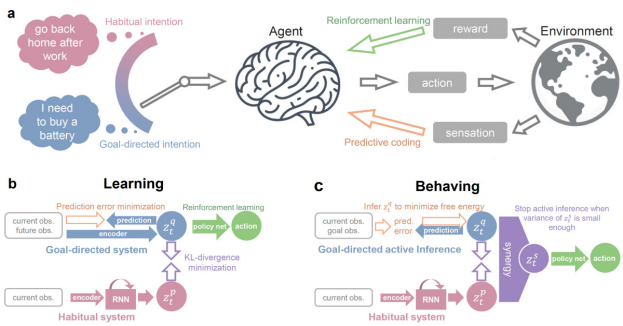


图5：（a）贝叶斯行为框架概述；（b）和（c）学习过程和行为过程框架图

在视觉引导的感知运动任务中进行模拟实验的测试结果显示，贝叶斯行为框架所得出的结论与神经科学和心理学实验的观察数据相吻合。这一发现不仅为认知科学中“行为”的理解提供了新的视角，也为具身智能的构建提供了理论基础。例如，人类能够轻松地用左手食指和小指拿起东西，或者原地转圈，未来的具身智能也可能完成这种未曾学习过的动作，展现出更高的适应性和灵活性。

跨领域研究让人工智能向节能高效进化

从达尔文进化论的角度来看，现在的主流人工智能模型在未来可能会面临淘汰。在生物进化的过程中，物种的基因变异是繁殖下一代时的常态。那些有利于生物适应环境的变异，将通过环境的筛选，以“适者生存”的原则被保留下来。然而，将这一概念应用于人工智能时，我们会发现能耗问题并不利于人工智能的发展和“进化”。

借鉴人脑的工作原理，构建脑启发的人工智能，不失为促进人工智能技术向节能高效方向发展的有效途径。这一趋势已经引发了新的研究热潮，包括对大脑理解的研究、基于神经元构建新的语言模型、根据不同脑区功能设计的 MoE 架构等脑启发人工智能正蓬勃发展。

在微软亚洲研究院进行脑启发式人工智能研究的过程中，研究员们更加体会到跨学科、跨领域专家协作支持的重要性。CircuitNet、SNN 时间序列框架、贝叶斯行为框架等创新成果的背后，凝聚了来自复旦大学、上海交通大学及日本冲绳科学技术大学院大学等机构的神经科学和脑科学专家的专业知识和贡献。

未来，随着对大脑机理的深入理解和技术的不断创新，我们有望增进对智能本质的理解，构建出更加智能、高效且环保的人工智能技术，更好地服务于人类社会。

相关链接：

CircuitNet: A Generic Neural Network to Realize Universal Circuit Motif Modeling
https://openreview.net/pdf?id=Fl9q5z40e3

Efficient and Effective Time-Series Forecasting with Spiking Neural Networks
https://arxiv.org/pdf/2402.01533

Advancing Spiking Neural Networks for Sequential Modeling with Central Pattern Generators
https://arxiv.org/pdf/2405.14362

Synergizing Habits and Goals with Variational Bayes
https://www.nature.com/articles/s41467-024-48577-7
*该论文已在《Nature Communications》杂志上发表。

集成大语言模型与产业数据智能，迈向“产业基础模型”

作者：机器学习组

随着数据量和模型规模的增加，大语言模型在指令执行、知识存储、逻辑推理和编程技能等方面展现出了突破性的能力。然而，大语言模型在产业领域的潜能尚未得到充分挖掘，特别是在满足产业数据分析、推理、预测、决策等数据智能需求方面。如何有效地变革各行业的数据模型及智能的构建方法与应用范式，仍然面临诸多挑战。为应对这些挑战，微软亚洲研究院提出了构建产业基础模型的倡议，其核心理念在于通过持续预训练，将产业数据智能相关的知识与技能融入到大语言模型中。基于这一理念，微软亚洲研究院开发了生成式表数据学习（Generative Tabular Learning，GTL）框架，展示了如何在表数据这一广泛使用的数据表征上，构建具有跨行业、跨数据模式、跨任务的产业基础模型。

尽管大语言模在新闻撰写、文档总结、客服助理和虚拟助手等以语言为中心的任务上表现出色，但在深入理解和处理特定的行业数据时仍存在局限。为了应对大模型在产业界应用中所面临的挑战，微软亚洲研究院提出了构建产业基础模型（Industrial Foundation Models）的创新思路，并在表数据上成功验证了实现跨领域通用数据智能的可行性及其巨大潜力。研究员们设计的生成式表数据学习（Generative Tabular Learning，GTL）框架，成功地将多行业数据智能相关的知识融入大语言模型中，使其具备在新领域、新数据及新任务上的直接迁移和泛化能力，更加敏捷地响应不同的产业需求。现在，微软亚洲研究院正式开源这一技术范式，并希望通过此范式推动数据科学在各行业中的广泛应用，促使复杂的数据智能技术变得人人可及。

产业数据的巨大潜力亟待挖掘

微软亚洲研究院的研究员们发现，大语言模型在利用产业数据这一关键资源方面，尚未充分发挥其潜力。产业数据通常以特定结构存储在不同行业和部门的数据仓库中，比如用于关系结构的表数据、记录时变信号的时间序列数据，以及用于复杂相互关联的图数据。这些结构中蕴含的丰富数据知识往往难以通过自然语言捕捉，因此当前以语义知识为核心的大语言模型在掌握数据智能相关的知识与能力方面存在不足。

更重要的是，产业数据及其蕴含的智能，为多个领域的重要应用奠定了基础。这种智能不仅来源于数值和结构化信息，还包

括特定任务的需求和领域专有知识。例如，在医疗健康领域，来自患者的基本信息、生理信号和治疗历史的数据，可用于辅助精确诊断和预后分析。在能源存储领域，分析电池循环数据中的模式，可以加速材料筛选、优化充放电协议、指导电池回收中的价值评估。在商业领域，历史销售和需求数据可以辅助预测未来的市场趋势并制定定价策略。传统的数据智能方法通常依赖于特定的数据模式与任务需求，具体表现为各个垂直领域中独立开发及优化的小模型。

为应对这些挑战与机遇，微软亚洲研究院提出构建产业基础模型的新思路。其核心策略是以统一的方式表征产业数据，并在此基础上对大语言模型进行持续预训练，从而将通用的数据智能知识与能力整合到大语言模型中，创造出在新场景上可直接应用的产业基础模型。这种模型不仅能够在各行业部门中执行以语言指令为中心的任务，还可以提取跨任务和跨部门的产业知识，并进行数据驱动的预测和逻辑推理。

此外，通过提供一个以语言为中心、无需参数调优和编写代码的用户界面，产业基础模型还有潜力改变传统的数据科学应用范式。这个用户友好的界面将使各行业的领域专家具备全面的数据科学技能，推动先进数据分析技术的普及。

同时，产业基础模型强大的跨领域能力，也使其能够有效地进行知识迁移与技能泛化。这对在数据有限的领域进行有效地少样本上下文学习尤为关键。

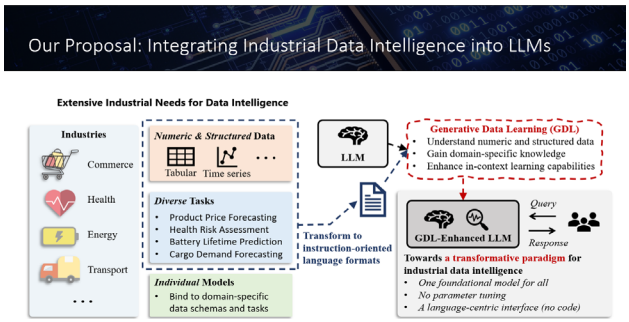


图1：产业基础模型的架构概览

基于表数据开发产业基础模型

表数据（Tabular Data）通常存储于关系型数据库中，是众多产业领域中最普遍的数据格式之一，也是预测建模的基础。因此，微软亚洲研究院的研究员们从表数据着手，构建能够横跨不同产业领域的基础模型。

研究员们收集了来自不同产业领域的各种表数据集及其相应的预测任务，并将这些数据转换为面向指令的语言格式。这种转换使得大语言模型能够适应多样化的数据模式，例如不同特征的语义和数值含义，支持数值和类别特征的任意组合。此外，通过将大语言模型与数据样本及可选的背景信息结合，模型不仅能够处理回归和分类任务，还能够支持零样本（Zero-Shot）学习和少样本上下文学习（In-Context Learning）的场景。

然而，将大语言模型的语言处理能力融入表数据的学习中仍面临巨大的挑战。最主要的问题在于，大语言模型通常在自然语言数据上进行预训练，因此在处理格式化表数据的精细差别时显得力不从心，并且缺乏对特定领域知识的深入理解，而这些知识对于有效的表数据学习至关重要。

为了解决这些挑战，研究员们引入了一个持续预训练阶段，即生成式表数据学习（Generative Tabular Learning，GTL）。通过对特征和标签标记进行自回归式生成建模，GTL 框架可以将数据知识与统计学习能力有效整合到大语言模型中。经过 GTL 框架增强的大语言模型，可以通过调整指令提示，直接应用于新的产业数据和任务需求。这意味着，模型能够在无需复杂参数调优的情况下，实现高效的数据处理，并且在不同领域知识、数据模式和任务之间进行广泛迁移，从而推动大语言模型向产业模型的方向进化。

实验结果：GTL显著增强了LLaMA模型对表数据的理解能力

为了验证 GTL 的有效性，研究员们收集了来自超过400个不同领域的表数据集，经过严格的去重过滤和筛选，最终保留了384个独立的数据集。其中，44个数据集被用于模型评估，其余的数据集用于构建1000多个不同的预测任务，以支持 GTL 的持续预训练。研究员们选择 LLaMA 2 作为基础大语言模型，并将其与开源和私有的大语言模型，以及传统表数据机器学习算法进行了比较。

如图2所示，实验结果表明，GTL 显著增强了 LLaMA 模型对表数据的理解能力。这表明，表数据中所蕴含的行业知识尚未被开源的 LLaMA 模型充分掌握，而 GTL 则有效弥补了纯语言数据训练出的语言模型在产业数据智能上的不足。值得一提的是，尽管 GTL 增强的 LLaMA 模型参数规模较小，但其性能与 GPT-4 等更大规模的模型相比仍具有竞争力，甚至在某些情况下表现更为优异。不过需要注意的是，与 GPT-4 在公开表数据上的对比结果可能因其私有训练数据中潜在的"数据污染"问题而产生偏差。

此外，GTL 增强的 LLaMA 模型不仅在少样本学习场景中通过无须调参的上下文学习，超越了传统表数据机器学习方法的统计学习能力，还具备了这些方法所缺乏的零样本学习能力。

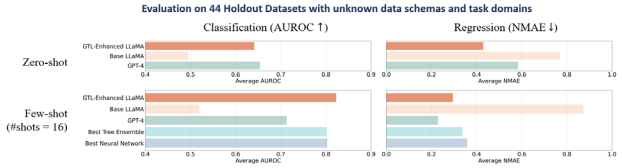


图2：GTL 增强的 LLaMA-2-13B 与其他基线模型的对比（更多详细结果请参阅论文）

研究员们还初步探究了 GTL 的规模定律。如图3所示，数据的多样性和模型参数规模都以幂律方式提升了新数据和新任务上的性能。这一发现表明了产业基础模型在跨多样任务和领域的广泛泛化潜力，有望使复杂的数据智能技术变得更加普及，即便在数据可得性有限的行业中也能发挥重要作用。

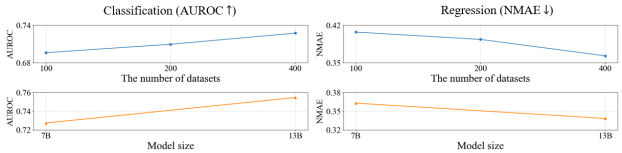


图3：初探 GTL 的规模定律

多维度拓展产业基础模型的潜力

生成式表数据学习（GTL）为会话式表数据深度学习打开了大门，使用户能够通过与模型对话来实现数据智能相关的分析、预测、推理和决策。通过将 GTL 与语言模型集成，模型不仅能够生成预测结果，还可以提供对相应结果的解释，从而为表数据学习的可解释性带来了新的机遇。基于这一范式所展现出的巨大潜力，微软亚洲研究院从两个角度展望了产业基础模型未来的研究和应用前景。

首先，产业基础模型本身的多维度扩展蕴含着巨大的潜力。这包括扩展数据集的种类和规模、增加模型规模、延长上下文长度，以及整合多样化的数据格式，如时间序列和图数据等。全面的扩展将使产业基础模型能够以更高的精度和更强的适应性，处理更多领域的更广泛任务。同时，产业数据知识与大语言模型生态系统的前沿进展相结合，如工具使用、智能体和对话交互，将进一步拓展产业基础模型的能力边界。这种协同作用可以打造更鲁棒和多功能的模型，将产业数据智能与大语言模型的复杂功能无缝融合。

其次，从用户视角来看，产业基础模型的发展将彻底革新产业数据智能的实现方式，重新定义数据科学的用户界面和工具链，进而催生出创新性的产品和服务。例如，领域专家无需掌握深厚的编程和数据科学知识，即可借助数据科学助手获得先进的数据分析和预测能力，从而推动前沿数据科学工具的普及。另外，产业基础模型可以作为决策支持工具，为行业领导者和从业者提供深刻的数据洞察和个性化分析，帮助企业做出更明智的战略决策，优化运营流程，并发掘新的增长机遇。

将大语言模型与产业数据智能相结合，是迈向产业基础模型的关键一步。通过持续扩展和创新，创建以用户为中心的工具，使前沿的数据智能技术更易于获取，能够释放出产业基础模型在各个行业中的更多潜能。微软亚洲研究院将持续推动这一进程，不断突破界限，让前沿的数据智能技术惠及更多的行业领域。

相关链接：

From Supervised to Generative: A Novel Paradigm for Tabular Deep Learning with Large Language Models（已收录于KDD 2024）
https://dl.acm.org/doi/10.1145/3637528.3671975

项目链接：
https://github.com/microsoft/Industrial-Foundation-Models

跨越模态边界，探索原生多模态大语言模型

当前多模态模型大致分为两类，一类是专用多模态模型，如文本生成图像、文本生成视频等；另一类则是通用型多模态大语言模型，这类模型的目标是让人工智能具备自然语言理解和生成、图像识别，以及语音和视频的交互能力。近日，微软亚洲研究院又提供了一个新的选择——原生多模态大语言模型。它能够更深入地理解物理世界并执行多模态推理和跨模态迁移，其在不同模态的数据学习中还涌现出了新的能力。

随着人工智能技术的持续发展，大模型已经从单一模态向多模态演化，多模态模型的应用也开始逐渐进入人们的视野。然而，终端用户现在所接触到的多模态模型还不是多模态模型的“完全体”。目前，多模态模型主要有三种实现方式：

多模态接口：在系统层开发统一的用户界面，具备多种模态数据输入和多种模态输出的能力，但是实现上则可以通过调用不同模态的模型甚至是 API，在终端实现多模态能力；

多模态对齐与融合：在技术框架层将语言模型、视觉模型、声音模型等进行连接，这些模型相互独立学习，使用不同模态的数据进行训练，然后将拼接好的模型在跨模态数据上继续预训练以及在不同任务数据上进行微调；

原生多模态大语言模型：从训练阶段开始，模型就利用大量不同模态的数据进行预训练，技术上实现紧密的耦合，可以在输入和输出端实现多模态，且还具备强大的多模态推理能力以及跨模态迁移能力。通常，这一类型才被认为是真正的多模态模型。

原生多模态大语言模型

在微软亚洲研究院全球研究合伙人韦福如看来，真正的原生多模态大语言模型不仅要在输入输出端支持多模态，还必须是具有实现多模态推理和跨模态迁移能力的端到端模型。而且基于多模态数据原生训练的每一种单模态能力，都应该超越只在单模态数据上训练的模型的性能。更重要的是，在不同模态数据学习的过程中，模型应该能够涌现出新的能力。

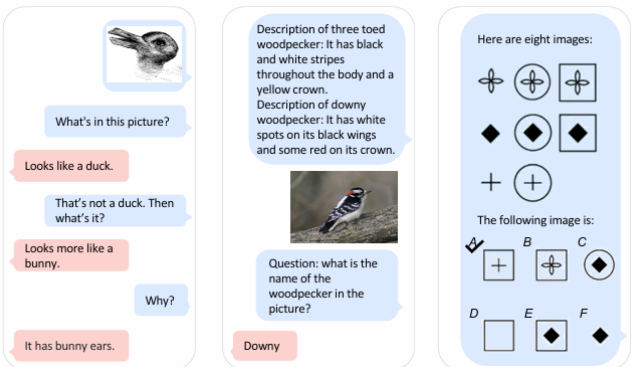
沿着这一思路，微软亚洲研究院通用人工智能组的研究员们先后研发了多模态大语言模型 KOSMOS-1 (opens in new tab)、KOSMOS-2 (opens in new tab)、KOSMOS-2.5 (opens in new tab)。现在，在这些工作的基础上，研究员们持续探索原生多模态语言模型，希望能够在输入和输出端都实现对原生多模态数据的支持，从而更深入地理解物理世界，并执行多模态推理和跨模态迁移。

KOSMOS的不断发展得益于前代模型的研究成果：KOSMOS-1 实现了语言与感知的对齐，为大语言模型支持多模态任务奠定了基础。KOSMOS-2 引入了 Grounding 能力，增强了模型的空间想象力，解锁了多模态大语言模型的细粒度理解和推理的能力。KOSMOS-2.5 通过统一框架来处理文本密集图像的多模态阅读和理解任务，为文本丰富图像的应用提供了通用接口。

语言是多模态模型的基础

“语言是所有多模态模型的基础。在人工智能和计算机科学领域，我们的目标是让机器理解人类的语言，而不是迫使人类去学习机器的语言。所以，从模型的最终应用形态来看，语言是最直接的交互方式。此外，语言及文本具有独特的优势，能够促进模型上下文理解、指令遵从以及推理能力的训练，这是其他单一模态数据难以提供的”。韦福如表示。

基于这些思考，微软亚洲研究院的研究员们在 KOSMOS 项目的早期研究中，就将语言模型原生支持多模态数据作为目标。在 KOSMOS-1，研究员们实现了大语言模型与感知能力的对齐，使 KOSMOS-1 模型能够原生支持语言、感知-语言和视觉任务，涵盖了广泛的感知密集型任务，包括视觉对话、简单数学方程求解、OCR，以及带描述的零样本图像分类等。



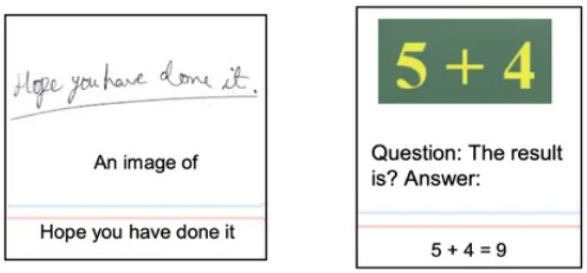


图1: KOSMOS-1 涵盖广泛的感知密集型任务：视觉对话、带描述的零样本图像分类、非语言推理、OCR、数学计算

与此同时，KOSMOS-1 在大语言模型推理能力的基础上，可以进行非语言推理。研究员们根据瑞文推理测验（Raven's Progressive Matrices）建立了 IQ 测试基准，来评估 KOSMOS-1 模型在非语言任务上的推理能力。结果表明，KOSMOS-1 能够感知非语言上下文中的抽象概念模式，并可以从多个选项中推导出下一个元素。这标志着 KOSMOS-1 可有效地完成部分零样本瑞文推理测验。

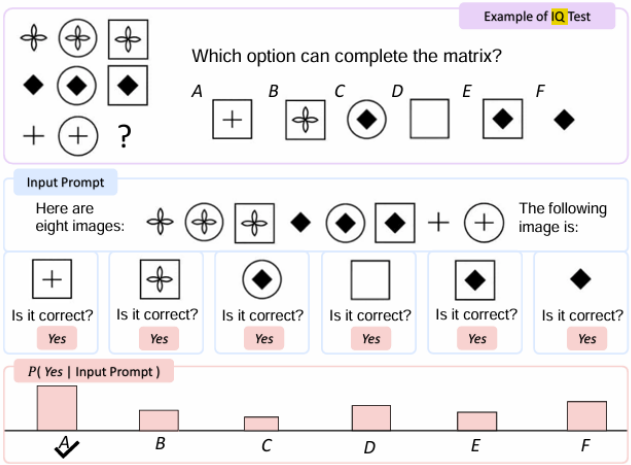


图2: 上图展示了瑞文推理测验的一个例子，下图则展示了将 KOSMOS-1 在瑞文推理测验中进行评测

在 KOSMOS-1 的基础上，KOSMOS-2 采用了相同的模型架构，并引入了 Grounding 能力，赋予模型"空间想象力"。KOSMOS-2 允许用户直接选择图像中的对象或区域作为输入，无需输入详细的文本描述，模型便能够理解该图像区域及其空间位置。Grounding 能力还使模型能够以视觉答案（例如边界框）的形式进行回应，并将生成的自由形式文本响应中的名词短语和指代表达链接到图像区域，有效解决了指代歧义问题，从而提供了更准确、信息丰富且全面的答案。

KOSMOS-2.5 在 KOSMOS-2 的基础上，进一步增强了对文本密集图像的多模态阅读和理解能力，包括信息提取、布局检测和分析、视觉问答、截图理解、用户界面自动化（UI Automation）等。KOSMOS-2.5 能够无缝处理视觉和文本数据，实现对文本丰富图像的深入理解，并生成结构化的文本描述。

图3: KOSMOS-2 可以将文本回答同图像中对应的区域进行连接，用户也可以通过边界框表明多模态的指代

通过统一的框架，KOSMOS-2.5 可处理两个紧密协作的任务。第一个任务是根据文本密集图像生成具有空间感知的文本块，即同时生成文本块的内容与其在文本密集图像中对应的坐标框。第二个任务是以 Markdown 格式生成结构化的文本输出，同时捕捉各种样式和结构。KOSMOS-2.5 将基于 ViT（Vision Transformer）的视觉编码器与基于 Transformer 架构的解码器相结合，并通过一个重采样模块将它们连接起来，实现了高效的多模态数据处理。

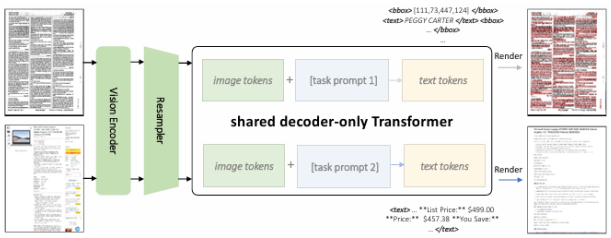


图4: KOSMOS-2.5 模型架构由一个预训练视觉编码器和一个与重采样器模块连接的语言解码器组成

这种统一的模型接口显著简化了下游任务的训练，并使模型能够在现实世界的应用中有效遵循指令。

声音与视频模态也可以基于语言模型的方法建模

尽管声音和视频是一种连续信号，但它们也可以被转换为类似文字的离散词元（token），这样声音模型可以与语言模型无缝融合。微软亚洲研究院首次基于语言建模的方法设计了文本到语音合成框架 VALL-E，首次实现了零样本TTS合成。

作为一个神经编解码语言模型，VALL-E 利用神经音频编解码模型的离散代码表示声音，并基于离散代码将 TTS 视为一个条件语言建模任务，而非传统的连续信号回归。与以往的语音合成流程，如音素 → 梅尔频谱图 → 波形不同，VALL-E 的处理流程是音素 → 离散代码 → 波形。VALL-E 基于目标文本对应的音素和用户声学提示生成离散的音频编解码代码，这些代码解码后可以得到对应目标文本内容的声音，并且具有和用户声学提示一样的音色。

VALL-E 还展现出了类似于文本语言模型的上下文学习能力。仅需一段3秒钟未见过说话者的录音作为声学提示，VALL-E 就能合成高质量的个性化语音。目前升级版的 VALL-E X 支持包括英文、德文在内的多种语言的 TTS 合成。

在原生多模态数据的学习过程中，VALL-E X 模型展示了一种新的、有趣的能力。即便没有经过专门的数据训练，VALL-E X 也能合成不同口音的语音，比如英伦风格、日韩式口音的英语，或者外国人说汉语时的特殊腔调。值得一提的是，为了确保模型使用的安全性，研究员们还给 VALL-E 多模态语音模型添加了水印功能，以确保输入的声音数据得到本人授权，防止滥用现象发生。

而视频则是多模态大语言模型的基础能力。“从数据形式上看，视频是融合不同模态数据的最佳数据类型，它包含了文字、图像、声音等多种元素，并且天然就是流式的（streaming）数据。而且，对于世界模型的构建来说，视频能够提供最丰富的数据，帮助模型学习物理世界的规律。因此，无论是从训练学习的角度，还是从最终能力的角度来看，视频都是多模态模型不可或缺的要素”。韦福如表示。

从算法和架构上推动原生多模态模型发展

韦福如认为，当前多模态模型的发展将经历几个主要阶段。第一阶段，大语言模型将调用其他模型或服务，来完成多模态的输入或者输出。例如，读取图片内容时，可以通过调用 OCR 功能提取文本信息或者利用 ASR 模型把语音转换成文本，进而作为语言大模型的输入。这将使得多模态模型在输入端具备视觉和听觉能力，然而这一阶段通常不包含多模态推理。同样，通过调用文本到图像、文本到语音或文本到视频的模型，多模态模型在输出端也能生成不同模态的内容。

而在第二阶段，模型需要实现多模态融合和推理。例如，当谈到“如何将大象装入冰箱”时，模型需要像人脑一样自然地联想并用到不同模态的相关知识（例如大象和冰箱的概念）和步骤（把一个物体放入冰箱的流程）。

“要想实现原生多模态模型的终极形态，我们还面临几个关键问题”，韦福如说，“首先，我们需要决定模型输入和输出端数据的表示方式，其本质是离散数据（例如文本）和连续数据（例如图像和语音）的统一建模、表示和学习。是直接使用原始图像或视频等数据以保留尽可能多的信息？还是将连续数据转换成离散的词元以实现不同模态数据类似自然语言的统一表示和学习？其次，如何有效地融合不同模态的数据？这需要设计新的模型架构，以便模型能够在理解和整合来自不同源的信息的同时不会相互冲突。最后，也是最具挑战性的问题，如何构建一个支持多模态原生的学习目标和范式？比如一个开放的问题要怎么统一语言模型（LLM）和扩散模型（Diffusion Model），来实现深度多模态对齐、推理和跨模态迁移，并促进新的能力涌现。我们相信这些方面近期都会取得重大研究成果。”

面对这些问题，微软亚洲研究院将持续探索。研究员们已经在应对技术和算法上的挑战，希望能够为未来原生多模态模型的研究和开发提供基础技术的创新突破。

注：本文中提及的所有人工智能技术，均为科研层面的探索和实验性成果，旨在利用人工智能技术为人类社会带来更多的可能性和价值。微软亚洲研究院在进行这些研究的同时，始终遵守微软负责任的人工智能流程的指导，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。微软始终致力于打击虚假信息，并尽其所能提供最新技术来检测被人为操纵的内容，帮助人们识别“深度伪造”（deepfake）的信息（欲了解微软为打击虚假信息所做的努力，请访问：https://blogs.microsoft.com/on-the-issues/2020/09/01/disinformation-deepfakes-newsguard-video-authenticator/）。

USENIX ATC 2024最佳论文 | 微软如何提升云AI基础设施可靠性

作者 : SuperBench团队组

来自微软亚洲研究院的研究员们和来自微软 Azure 云平台的工程师们联合发布了一项开创性的、聚焦云 AI 基础设施高可靠性的研究工作 : SuperBench 系统。该系统通过主动验证的手段解决了云 AI 基础设施中难以捉摸的"灰色故障"问题。这一研究工作已被全球计算机系统领域的顶级学术会议 USENIX ATC 2024 接收 , 并荣获最佳论文奖。SuperBench 不仅引起了业界的广泛关注 , 还有望改变未来云服务提供商确保 AI 基础设施高可靠性的方式 , 从而为行业树立新的标准。

随着云 AI 工作负载变得越来越复杂和大规模 , 维护系统的高可靠性变得至关重要。传统的系统高可靠性保障方法 , 如冗余组件 , 不经意间引入了一个新的问题——隐性性能退化 , 又被称为灰色故障。灰色故障由冗余组件的逐渐失效引起 , 前期主要表现为不明显的性能逐渐下降 , 并且难以被系统管理者察觉。当后期冗余组件完全失效时 , 系统才会显现出明显的性能退化。这使得识别和解决系统故障的任务变得十分复杂。

传统的系统可靠性保障方法往往依赖于被动的故障排除手段 , 比如硬件预检和故障后修复 , 这些方法无法有效解决灰色故障问题。微软亚洲研究院的研究员们与微软 Azure 云平台的工程师们意识到 , 仅靠被动的故障排除并不足以应对这一挑战。于是 , 他们提出了一种创新的主动验证解决方案——SuperBench 系统。SuperBench 通过引入全面的基准测试和主动验证技术 , 能够在故障发生之前识别潜在的性能问题 , 从而显著提升系统的整体可靠性。相关论文已被全球计算机系统领域的顶级学术会议 USENIX ATC 2024 接收 , 并荣获最佳论文奖。

SuperBench 的设计理念是主动验证而非被动反应，它能够在系统出现显著性能退化之前，及时检测并修复潜在的问题。这种方法不仅提高了系统的稳定性，也减少了维护成本和用户遭遇的性能问题。

为了有效缩短平均故障间隔时间，主动验证必须满足以下要求：首先，它需要全面覆盖各种 AI 工作负载，以确保检测到在新集群中可能被忽视的问题；其次，验证必须具有明确的标准，以区分正常性能和渐进性性能退化的问题，确保测试结果的一致性；最后，验证过程必须具备成本效益，以确保验证开销远低于处理故障所带来的费用。

然而，实现这些要求面临着不少显著的挑战：工作负载和节点组合的数量庞大，使得验证过程中无法涵盖所有场景；缺乏对缺陷组件的可靠评估标准，所以硬件规格无法准确预测负载性能；AI 硬件的变化性加大了问题的复杂性；此外，验证时间和平均故障间隔时间之间存在相互影响，让优化验证成本与延长平均故障间隔时间的平衡成为一项复杂的任务。

SuperBench 的核心是一套全面的基准测试套件，用于评估单个硬件组件和各种真实的 AI 工作负载，其能够确保系统检测到在正常操作过程中可能隐匿的问题。

SuperBench 包括：

- 全面的基准测试套件：包括对典型 AI 工作负载的端到端基准测试和针对单个硬件组件的微基准测试，能够更全面、更彻底地对系统进行测试并及早发现潜在问题。
- 选择器模块：采用实时概率模型来确定最有效的基准子集，能够在验证时间和事件相关成本之间取得平衡，从而确保验证的高效和影响力。
- 验证器模块：利用先进的机器学习技术分析基准数据，并精准定位缺陷硬件。通过关注累积分布指标而非平均值，SuperBench 可以清晰地区分功能正常和故障的组件。

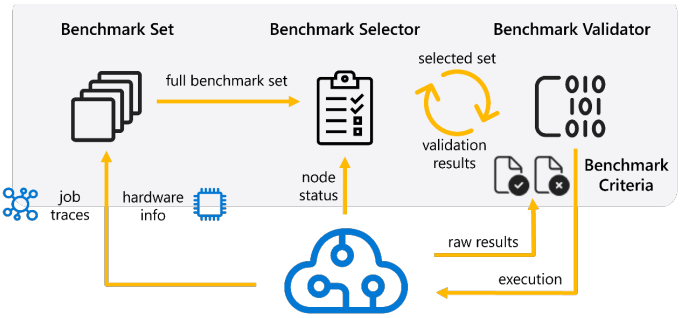


图1 : SuperBench 的工作原理概览

通过在 Azure 生产环境中两年的成功部署 , SuperBench 充分展示了其有效性。在此期间 , SuperBench 验证了数十万块 GPU , 识别出了10.36%的节点存在缺陷 , 并显著提高了系统的可靠性。

模拟结果表明 , 与未进行验证和未选择基准的全套验证相比 , SuperBench 可以将平均故障间隔时间 (MTBI) 提高至22.61倍 , 并将用户 GPU 利用率增加4.81倍 , 同时将验证时间成本降低92.07%。

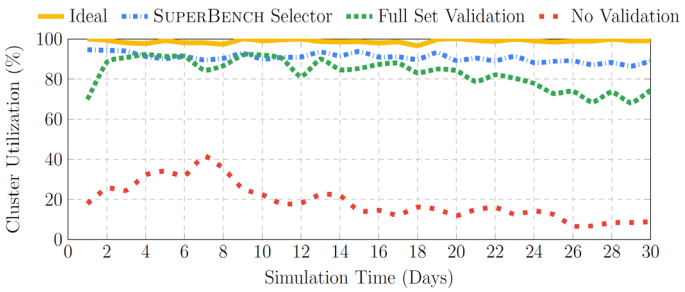


图2：30天内采用不同基准模拟的平均节点利用率，SuperBench 实现了90.70%的高集群利用率，将无验证基线提高了4.81倍，全集基线提高了1.09×

SuperBench 的推出标志着主动系统验证的重大进展，其通过解决灰色故障问题提高了云 AI 基础设施的可靠性，同时还带来了成本节约和运营效率的提高。该研究不仅深入探究了云 AI 基础设施中的灰色故障问题，还分析了包括硬件故障、性能倒退等问题的来源和根本原因，对相关领域的研究做出重要贡献。在未来，微软亚洲研究院将继续探索如何提升云 AI 基础设施性能，完善云 AI 高效可靠的服务。

相关链接：

SuperBench: Improving Cloud AI Infrastructure Reliability with Proactive Validation
论文链接：https://www.microsoft.com/en-us/research/publication/superbench/
GitHub 链接：https://github.com/microsoft/superbenchmark

为什么你的LLMs玩不转外部知识？RAG分类学助你诊断！

作者：系统组（上海）

大语言模型在教育、医疗、金融等多领域的应用已展现出其不可忽视的价值。如何更好地结合外部数据，如何提升模型处理专业领域问题的可靠性，是大语言模型应用开发中值得不断思考的问题。针对此，微软亚洲研究院的研究员们提出了一种基于查询需求分层的 RAG 任务分类法，从显式事实、隐式事实、可解释的推理、隐式推理4个层级出发，直指大模型应用在不同认知处理阶段所面临的难点和定制化的解决手段。该研究可以使大模型更好地整合专有领域知识，保证其在特定领域中发挥最前沿的能力，在微软亚洲研究院与上海市精神卫生中心针对个性化认知训练展开的联合研究中发挥了关键作用。

随着人工智能的快速发展，结合外部数据的大语言模型（LLMs）在完成真实世界任务时展现出了卓越的性能。这些外部数据既能够提升 LLMs 的专业性和时效性，还降低了模型产生幻觉的风险，同时增强了其可控性和可解释性。尤其是当模型结合了无法纳入初始训练语料库的私有数据或特定场景数据时，这些优势更加明显。

然而，要将结合外部数据的 LLMs 有效地应用于不同的专业领域，目前仍面临重大挑战。这些挑战范围广泛，不仅包括构建数据管道以及准确捕捉用户查询的真实意图，还涉及到如何充分挖掘 LLMs 的潜力以实现复杂的智能推理。

在与领域专家和开发者进行深入讨论，并细致分析当前的挑战后，微软亚洲研究院的研究员们认识到，数据增强型 LLM 应用并非是一劳永逸的解决方案。由于实际需求复杂多变，尤其是在专业领域，数据与所需推理难度之间的关系可能存在显著差异。所以，在实际应用大模型时，表现不佳往往是因为未能准确把握任务的重点，或者任务本身需要多种能力的融合，要拆分处理后才能更有效地解决。

因此，研究员们提出了一种“RAG 任务分类法”，根据用户对外部数据查询的需求类型，将其分为四个级别，并对这四个级别查询的主要难点以及解决这些难题的技术手段进行归纳总结。

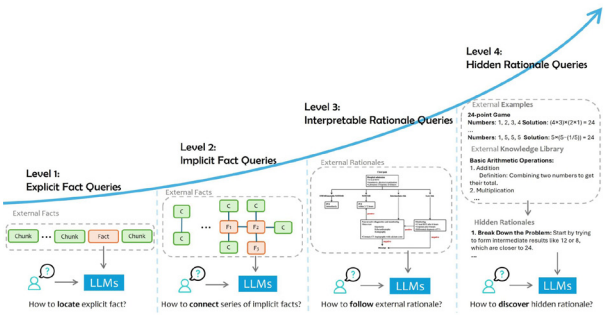


图1：四种级别查询的主要难点

研究员们认为，只要深入理解各个层面的需求以及与之相伴的独特挑战，就能顺利构建出应用，并通过不断改进来满足最终的任务需求。为此，他们精心编制了一份详尽的调研报告，旨在精确界定不同层次的查询需求，识别每个层次所面临的独特挑战，并详述解决这些挑战的相关工作与努力。该调研报告希望帮助读者构建对数据增强型 LLM 应用的全面认识，成为开发此类应用的实用指南。

将查询分层为四个级别，明确LLMs任务复杂度

在数据增强型 LLM 应用的框架下，研究员们首先根据复杂性和所需数据的交互深度，对查询进行了系统化的分层。这种分层方法有助于更深入地理解 LLMs 在生成准确且相关性强的响应时所经历的不同认知处理阶段。从基础的事实检索到深层次的隐含知识解释，每层都标志着 LLMs 任务复杂度的逐级提升。

具体的层次划分如下：

Level-1 显性事实：此类查询涉及直接从数据中提取明确存在的事实，无需进行任何形式的额外推理。这构成了最基础的查询类型，其中模型的主要任务是精确定位并提取相关信息。例如"2024年夏季奥运会在哪里举行？"

Level-2 隐性事实：这些查询要求揭示数据中隐含的事实，可能需要一些常识推理或简单的逻辑推断。信息可能分散在不同的数据片段中，或者需要通过简单的推理过程来获取。例如，"目前哪个国家正在举办堪培拉所在国家的执政党会议？"这个问题可以通过结合堪培拉位于澳大利亚的事实和当前执政党的信息来解答。

Level-3 可解释的推理：在这一层级，查询不仅要求对事实的掌握，还要求模型能够理解并应用与数据背景密切相关的领域特定推理依据。例如，在制药领域，LLMs 需要解读美国食品药品监督管理局（FDA）的指导文件，以评估药品申请是否符合监管要求。在客户支持场景中，LLMs 必须遵循预定义的工作流程来有效响应用户查询。在医学领域，LLMs 可以开发成一个专门管理胸痛的专家系统，遵循权威的诊断手册和标准化指南。这种能力确保了 LLMs 的输出不仅在事实上正确，而且在上下文中也相关，且严格遵守监管和操作规范。

Level-4 隐式推理：这一级别的查询进入了一个更具挑战性的领域，其中推理依据并未明确记录，而是需要通过分析历史数据中的模式和结果来推断。例如，在 IT 运营领域，LLMs 需要从云运营团队解决的历史事件中挖掘隐性知识，识别成功的策略和决策过程。在软件开发中，LLMs 必须从以往的调试错误记录中提取出指导性原则。通过整合这些隐含的推理依据，LLMs 提供的回答不仅准确，而且能够反映出经验丰富的专业人士的隐性知识和问题解决技巧。

将查询划分为不同层次，既体现了 LLMs 需要理解的复杂性和多样性，也指明了各个层次的关注点，如图2所示。前两个层级——显性事实和隐性事实，主要聚焦于事实信息的检索，无论是直接呈现的还是需要基本推理得出的。这些层级考验的是 LLMs 从数据中提取和综合信息以形成连贯事实的能力。与此相对，后两个层级——可解释的推理和隐式推理，则将重点转向了 LLMs 学习和应用数据背后逻辑的能力。这些层级要求更高层次的认知介入，LLMs 必须要么与专家的思维方式保持一致，要么从非结构化的历史数据中提炼出洞见。

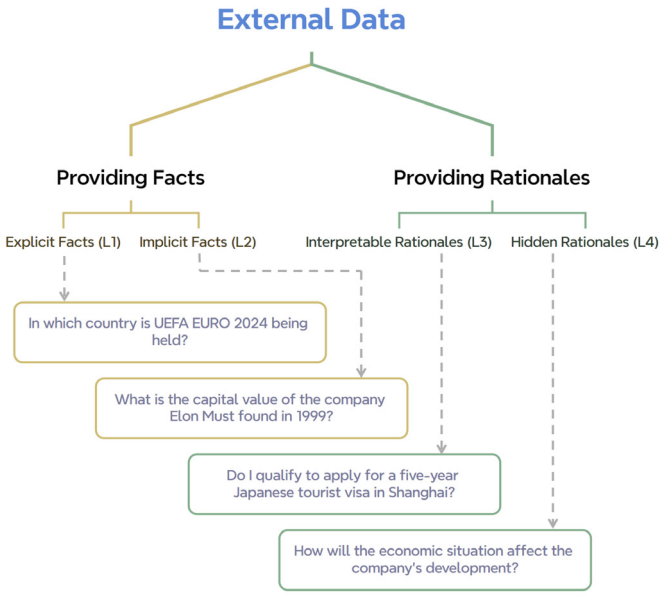


图2：数据增强的大语言模型应用中的查询级别总结

对于显性事实查询，准确的回答依赖于在庞大的外部数据库中精确定位具体的外部数据信息；而对于隐性事实查询，答案通常需要从多个相互关联的事实中综合得出，因此，全面检索并整合有效信息成为了这一类查询的关键挑战。可解释的推理查询任务是将多样的外部逻辑关系输入大语言模型，并确保其精确遵循这些逻辑指导来生成回应；对于隐式推理查询，从外部示例或知识库中提炼并识别出解决问题的策略是至关重要的任务。

针对不同层级查询的定制化解决方案

数据增强型大语言模型应用的四个层级各具特点，面临的挑战也各有不同，因此，每个层级都需要量身定制解决方案，如图3所示。对于涉及静态常识的查询，采用链式推理的通用型大语言模型能够有效应对。

在处理显性事实查询时，其关键挑战在于如何在数据库中精确地定位事实，因此，基础的 RAG（Retrieval-Augmented Generation，检索增强生成）方法成为了首选策略。对于隐性事实查询，这类查询要求整合多个相关事实，所以采用迭代式的 RAG 方法或基于图结构、树结构的 RAG 实现更为适宜，因为它

们能够同时检索独立事实并建立数据点之间的联系。在需要广泛数据互联的情况下，Text-to-SQL 技术则显得尤为重要，它可以通过数据库工具来增强外部数据的搜索能力。

针对可解释推理查询，运用提示调优和链式推理提示技术可以增强 LLMs 对外部指令的遵循度。而最具挑战性的隐藏推理查询，则需要从大量数据中自动提炼出问题解决策略。在这种情况下，离线学习、上下文学习以及模型的微调就成为了解决问题的关键手段。

总体而言，研究员们认为，开发者作为领域专家在着手开发特定的大语言模型应用之前深入洞察预期任务，明确相关查询的复杂性，并选取恰当的技术手段来解决问题十分必要。这些方法主要可以通过以下三种机制向 LLMs 注入知识，如图3所示：a) 根据查询需求，从领域数据中提取部分内容作为 LLMs 的上下文输入；b) 训练一个规模较小的模型，该模型在特定领域数据上训练后，用于引导外部信息的整合，并最终输入至 LLMs；c) 直接利用外部的领域知识对通用大语言模型进行微调，从而将其转化为领域专家模型。

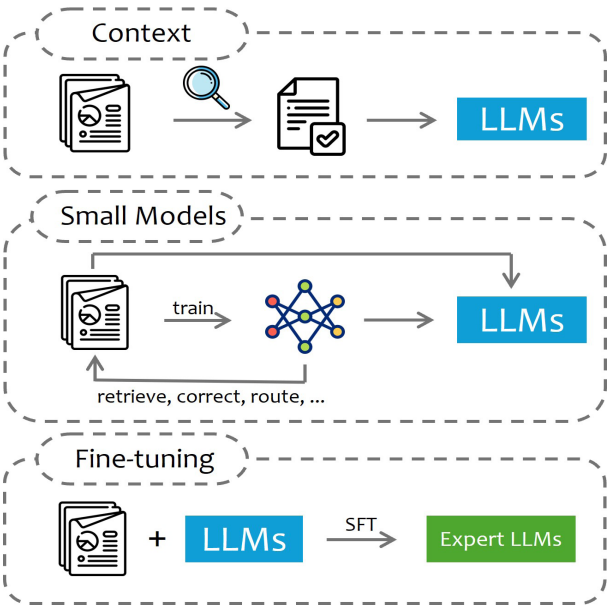


图3：将特定领域数据注入大语言模型的三种方式

这三种策略在数据量、训练时间和计算资源的需求上各有不同，且需求逐渐增加。通过上下文进行知识注入的方法在解释性和稳定性方面表现更为优异，但受限于上下文窗口的大小和潜在信息丢失，尤其是中间信息的缺失，这一方法也面临一定的局限性。因此，该方法更适用于可以通过简短文本解释的数据场景，对模型的检索和知识提取能力提出了较高要求。小型模型方法的优势在于训练时间短，且能够处理大量数据，但效果依赖于模型的能力，对于复杂任务，LLMs 的性能可能会受到限制，且随着数据量的增加，可能需要额外的训练成本。微调方法能够利用大模型处理大量特定领域数据的能力，但对 LLMs 的影响在很大程度上取决于所使用的数据质量。使用领域外的事实数据进行微调可能会无意中导致 LLMs 产生更多错误输出，甚至可能使其丧失原有的领域知识，并在微调过程中忽视未遇到的任务。

因此，在将数据注入 LLMs 时，选择合适的策略需要对数据源有深刻的理解，并基于此做出明智的选择。并且，在实际应用场景中，数据增强的 LLM 应用通常涉及多种类型的查询，这就要求开发者设计出一个合适的架构，将多种方法融合在一起，以有效地应对这些复杂挑战。

微软亚洲研究院的研究员们提出的 RAG 任务分类学方法已经在实际大语言模型应用中体现出其价值。在近期微软亚洲研究院和上海市精神卫生中心的联合研究中，研究员们开发了个性化认知训练框架"忆我"（ReMe），为帮助认知障碍患者进行认知训练带来更具便捷性、互动性、直观性和个性化的工具。该工具的设计正是结合了 RAG 等技术，使多模态大模型更好地整合了相关领域的专业知识，并优化了智能代理的行为逻辑和性能，展现出了该方法的巨大贡献。

相关链接：

Retrieval Augmented Generation (RAG) and Beyond: A Comprehensive Survey on How to Make your LLMs use External Data More Wisely
论文链接：https://arxiv.org/abs/2409.14924

开源工具RD-Agent：让研究与开发过程更智能

作者：机器学习组

研究与开发（R&D）是推动社会进步、经济增长和技术创新的核心动力。在人工智能时代，如何充分激发大语言模型的潜力，通过自动化手段提升研究与开发效率，实现跨领域知识迁移与创新，已成为 R&D 智能化转型的关键。为应对这一挑战，微软亚洲研究院推出了自动化研究与开发工具 RD-Agent，依托大语言模型的强大能力，开创了以人工智能驱动 R&D 流程自动化的新模式。RD-Agent 不仅提高了研发效率，还利用智能化的决策和反馈机制，为未来的跨领域创新与知识迁移提供了无限可能，赋能 R&D 迈向全新高度。

在现代工业中，研究与开发（R&D）是推动数字化转型和提升生产力与生产效率的关键。然而，随着人工智能技术的快速发展，传统 R&D 自动化方法的局限性逐渐显露。尤其是在提供高效、精准的自动化解决方案时，这些方法缺乏足够的智能，难以满足创新型研究和复杂开发任务的需求，远未能达到"像人类专家那样创造显著产业价值"的水平。相比之下，经验丰富的人类专家能够基于深厚的知识提出新想法、验证假设，并通过反复试验不断优化流程。

大语言模型（LLMs）的出现，为这些问题带来了全新的解决方案，并将为数据驱动的 R&D 场景的自动化产生巨大的推动作用。通过在各个领域的海量数据中进行训练，大语言模型积累了丰富的知识，能够提供传统方法所缺乏的智能性。凭借从数据中提取逻辑推理的能力，大语言模型可以支持复杂的决策过程，帮助自主执行任务，并在多种工作流程中作为智能代理（AI agent）发挥作用。

大语言模型为R&D注入新智能

微软亚洲研究院的研究员们认为，大语言模型在推动创新性研究方面具有巨大的潜力和价值，其广泛的知识覆盖面不仅有助于提出全新的想法和假设，还能够通过强大的推理能力为研究设计新的实验路径和方法，进而促进持续创新。在开发环节，大语言模型在数据处理和分析方面表现出色，能够高效提炼信息、总结规律。此外，凭借对代理工具（agentic tools）的灵活运用或创建能力，大语言模型可以自动处理重复且复杂的任务，从而显著加快开发进程。

为此，研究员们设计了一个基于大语言模型能力的自动化研究与开发工具 RD-Agent。通过整合数据驱动的 R&D 系统，RD-Agent 可以借助强大的人工智能能力驱动创新与开发的自动化。

RD-Agent 的核心是一个自主代理（autonomous agent）框架，由研究（R）和开发（D）两个关键模块构成。研究模块负责提出新想法，积极探索新的可能性；开发模块则专注于实现这些想法。两者相辅相成，在实际应用中通过反馈循环不断优化。随着时间推移，这些模块的能力将逐步提升，以应对日益复杂的研发需求。

图1：用 AI 驱动 AI

在实际应用中，RD-Agent 可以发挥众多作用，它既可以作为高效的研发助手，遵循指示完成日常繁琐的研发工作，也可以作为具有高度自主性的智能代理，主动提出创新性想法并自动进行探索研究。

以下是 RD-Agent 可支持的部分场景演示，包括了从通用研究助理到辅助特定专业领域的数据智能研发：

- 作为通用科研助理，自动阅读研究论文或报告，并实现模型结构。
- 自动探索和实现模型结构，挖掘数据规律：如金融、医疗等领域。
- 作为自动化 Quant 工厂，在复杂的真实系统中，自动化完成大量耗时的特征工程工作。

目前，RD-Agent 工具已在 GitHub 上开源，微软亚洲研究院的研究员们正不断更新和扩展 RD-Agent 的功能，以适应更多的方法和场景，进一步优化研发过程，提高生产率。

RD-Agent的关键挑战与技术创新

在数据驱动的 R&D 自动化领域，大语言模型的应用带来了革命性的创新机遇。然而，实现这一愿景的关键挑战在于如何获取并持续进化专业知识。

具体来说，现有的大语言模型在完成初始训练后，其能力很难持续增长。因为大语言模型的训练过程更侧重于通用知识学习，所以对于高度专业化知识的理解并不透彻，而这些专业知识需要从行业内的深度实践中获得，这成为了解决领域内复杂研发问题的一大难题。

微软亚洲研究院的研究员们认识到，只有深入探索研发阶段，并持续获得深度领域知识，才可能让大语言模型的研发能力不断增长。因此，研究员们从研究、开发、测试基准三个层面展开了研究，进而设计了 RD-Agent 工具，实现了在真实世界的实践和反馈中的动态学习。

研究层面：探索新的想法并通过反馈对其优化。在 R&D 过程中，提出和验证新想法是研究的核心环节。数据挖掘专家会首先提出假设，例如循环神经网络 RNN 能够捕捉时间序列数据中的模式；然后设计实验，如在包含时间序列的金融数据场景中验证该假设；随后将实验想法转化为代码，例如 PyTorch 模型结构；最后执行代码以获取反馈，诸如指标、损失曲线等。专家们会从反馈中学习，并在下一次迭代中改进。

受这些理念的启发，研究员们建立了一个基本的方法框架，支持自动提出和验证假设，并从实践反馈中积累知识。RD-Agent 是首个将科学研究自动化和实践验证相连接的框架，并融入了知识管理机制，使其在探索中能够像人类专家一样不断地验证和积累知识。随着 agent 对场景的理解逐步加深，它还能提出更优的解决方案。

图2：研究层面的基本方法

开发层面：高效实现并执行想法。开发过程的关键在于高效实现研究成果，同时通过合理的任务优先级调度来最大化效益。研究员们在 RD-Agent 框架中提出了面向数据中心任务开发的解决方案 Co-STEER。这一方法旨在处理从简单任务入手，通过学习不断提高的开发策略，并利用持续反馈优化整体开发效率。

Co-STEER agent 通过不断进化的策略，积累特定领域的开发经验，不仅提高了任务调度的效率，还加速了开发能力的提升。Agent 开发水平不断增强，其反馈质量也随之提升，从而进一步优化调度算法，实现开发与调度的协同进化。

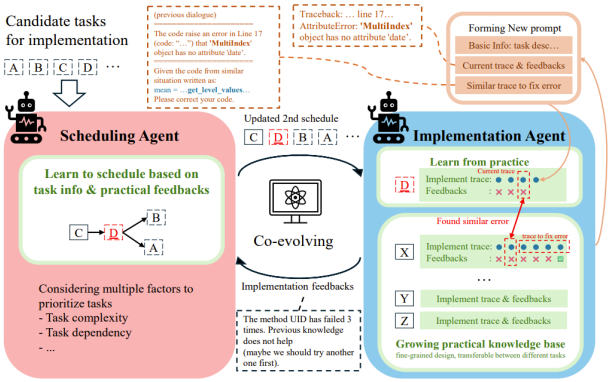


图3：Co-STEER 设计细节

基准测试（Benchmark）层面：构建新的基准测试体系，评测 agent 的 R&D 能力。研究员们还开发了一个全新的基准测试集：RD2Bench。该基准测试涵盖了从数据构建到模型设计的一系列任务，用于评估大语言模型代理（LLM-Agent）在数据和模型研发方面的能力。

在评估模型开发能力时，研究员们从专注于模型结构设计的论文中抽取关键信息，并将实现细节通过数学公式和文本描述结合的方式提供给agent。在数据开发能力的评估中，研究员们选择了金融特征（因子）作为典型的高知识密集型场景，从公开的研究报告中抽取因子的实现公式和描述，用于研发agent的输入。针对所有任务，研究员们都实现了对应的正确版本，以此作为评估模型和数据构建结果的基础。

大语言模型的创新潜力有待进一步激发

展望未来，如何更高效地开展自动化数据科学研究仍是一个开放性问题，而如何充分激发大语言模型的创新潜力，实现跨领域乃至跨学科的知识迁移、融合与创新，更是当前面临的重要挑战。在开发过程中，如何自动化地理解反馈信息，并将其与现有的开发水平紧密结合，同时智能地调度任务、择优执行，以提升基础模型作为 agent 的能力，都是极具挑战且具有深远意义的研究方向。

要解决这些挑战，关键在于通过实践反馈促进研究与开发能力的同步提升，实现二者的协同进化。这种有机结合的方法将极大地提升大语言模型的创新能力，推动跨领域和跨学科的知识转移与创新，从而进一步提升研发效率与质量，真正实现自动化研究与开发的飞跃。

相关链接：

RD-Agent GitHub链接：
https://github.com/microsoft/rd-agent

Towards Data-Centric Automatic R&D
https://arxiv.org/abs/2404.11276

Collaborative Evolving Strategy for Automatic Data-Centric Development
https://arxiv.org/abs/2407.18690

RD-Agent 体验链接：
https://aka.ms/RD-Agent

ProbTS：时间序列预测的统一评测框架

作者：机器学习组

如今，时间序列预测在健康、能源、商业、气候等多个行业发挥着至关重要的作用。它不仅影响着相关资源的分配和调度，还影响着行业的管理和运营决策。但是现有的时间序列预测方法通常缺乏对基础预测需求的全面考虑，无论是经典的时序预测模型还是近期涌现的时序基础模型，都存在方法设计上的"偏见"。

为此，微软亚洲研究院的研究员们联合香港科技大学（广州）和清华大学的科研人员合作开发了 ProbTS 框架，希望对现有时序预测模型进行统一的基准评测。在 ProbTS 框架下，研究员们通过在点估计/分布估计、长程/短程、自回归/非自回归等多维度上的预测效果比较，揭示了各模型在关键方法论上的"抉择"难题和差异，并对各模型进行了全面的优劣势辨析。ProbTS 的分析结果可以帮助业界反思当前时间序列预测模型在底层方法论上遭遇的挑战，更重要的是为未来预测模型的发展梳理出了更加清晰的研究方向。

时间序列预测（Time-series Forecasting）对众多行业都至关重要，包括健康、能源、商业、气候等。在不同预测长度上的准确性，对这些领域中服务短期和长期的规划和决策需求来说极其重要。例如，在疫情爆发这种公共卫生的紧急情况下，预测一到四周内的感染病例和死亡人数对于有效分配医疗和社会资源非常重要。在能源领域，准确预测每小时、每天、每周甚至每月的电力需求也对电网管理和可再生能源调度十分关键。同样，在物流行业，准确预测短期和长期的货物量能有效帮助企业合理安排运输路线以及高效管理供应链。

除了涵盖各种预测长度，面向规划和决策的精准预测不仅要考虑到点估计（Point Estimation），更要支持分布估计（Distribution Estimation），以衡量估计的不确定性。因为期望下的预测值及其相关的不确定性可以为随后的规划和优化提供一个全面的视角来引导更好的决策。

鉴于不同预测长度对点预测和分布预测的迫切需求，来自微软亚洲研究院的研究员们对现有不同研究领域开发的最先进的模型进行了回顾，这些模型包括：

- 经典时间序列模型：这些模型通常需要在每个数据集上从头开始训练，包括专门用于长程点预测（例如，PatchTST、iTransformer）以及专注于短程分布预测的方法（例如，CSDI、TimeGrad）。

- 近期的时间序列基础模型：这些模型涉及在广泛的时间序列数据集上进行通用预训练，包括由工业实验室（例如，TimesFM、MOIRAI、Chronos）和学术机构（例如，Timer、UniTS）开发的方法。

研究员们发现，尽管目前的预测模型有着可观的进展，但现有的方法通常缺乏对基础预测需求的全面考虑。这种局限性将导致现有模型方法在设计上存在"偏见"，而且这些模型能力尚未在更广泛的预测场景中得到验证。

基于此，研究员们开发了 ProbTS 框架。ProbTS 是一个统一的基准评测框架，旨在评估当前方法在满足基本预测需求方面的表现。研究员们通过 ProbTS 工具，不仅对预测研究的关键方法论差异进行了探讨，还对各类时间序列预测的经典模型和基础模

型进行了评测，揭示了现有时间序列预测研究中存在的问题，以及各模型的优劣势，进而对该领域未来的研究方向进行了梳理。

范式差异：时间序列预测的方法论辨析

研究员们通过 ProbTS 进行的基准研究发现，目前的时间序列预测关键方法论存在两方面的差异——点估计和分布估计的预测范式，以及多步预测的解码方案。点估计和分布估计的预测范式：

- 点预测：只支持点预测的方法，提供预期估计值而不进行不确定性量化。
- 预定义的分布函数预测头：使用预定义的分布函数预测头生成分布预测的方法，提供了一定的不确定估计，但缺乏对复杂数据分布的建模能力。
- 神经分布估计模块：采用基于神经网络的模块来估计数据分布，允许更灵活且可能更准确的不确定性量化。

多步预测输出的解码方案：

- 自回归（Autoregressive，简称 AR）方法：这些方法逐步生成预测，使用先前的预测作为未来时间步的输入，适用于序列依赖性至关重要的场景。
- 非自回归（Non-autoregressive，简称 NAR）方法：这些方法同时为所有时间步生成预测，提供更快的预测速度，并且可能在长程预测中表现更好。

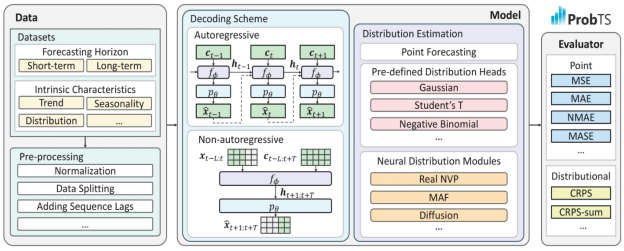


图1：ProbTS 概览，展示了其在不同预测场景中的覆盖范围，包括不同研究分支中开发的典型模型和全面的评估指标。

在 ProbTS 框架下的研究结果显示：首先，在长程及短程预测中，长程点预测的方法因定制化的神经架构在长程场景中表现出色，但在短程案例和复杂数据分布中表现不佳，并且因为缺乏对预测不确定性的量化评估，导致其与概率模型相比在应对复杂数据分布情况下存在显著的性能差距。而短程概率预测方法仅在短程分布预测方面表现专业，但在长程预测场景中就会出现性能下降以及计算效率的问题。

其次，针对解码器设计，长程点预测模型主要采用非自回归解码，而在短程概率预测模型设计中则没有出现这种偏向性。并且，尽管自回归解码在长程预测中容易受到误差累积的影响，但在具有强周期性模式的场景下可能表现更好。

最后，在当前涌现的时间序列基础模型中，部分采用自回归解码的基础模型在长程预测中同样面临误差累积的挑战，且尚未有较好的应对措施。同时，当前基础模型对分布预测的支持有限，突显了提升复杂数据分布建模能力的需求。

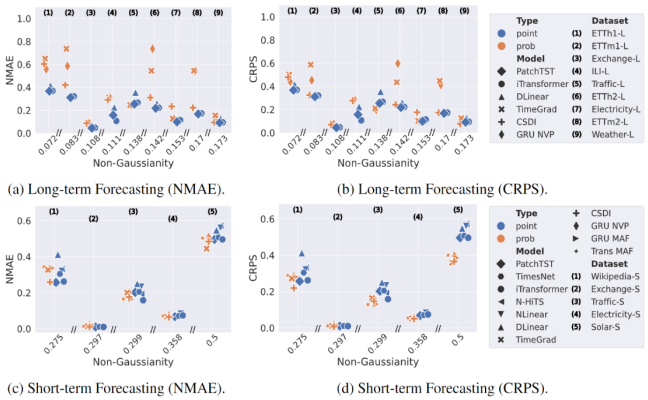
经典时间序列模型的评测结果与分析

研究员们使用 ProbTS 框架对广泛的预测场景中的各种经典时间序列模型进行了基准评测，涵盖短程和长程预测。具体评测指标包括点预测指标 NMAE (Normalized Mean Absolute Error) 和分布预测指标 CRPS (Continuous Ranked Probability Score)。此外，研究员们还通过计算一种非高斯性的评分，量化了每个预测场景中数据分布的复杂性。

根据 ProbTS 的评测结果，研究员们发现：

- 长程点预测模型的局限性：针对长程点预测所设计的时间序列神经架构，在长程场景中表现出色，然而，它们在短程预测任务中的架构优势显著降低（见图 2(a) 和 2(c)）。而且，这些模型无法衡量预测的不确定性，导致其与概率模型相比在分布预测上存在更大的性能差距。这一差距在数据分布复杂时会更加显著（见图 2(c) 和 2(d)）。

- 短程概率预测模型的弱点：当前的概率预测模型虽然在短程分布预测方面表现出色，但在长程场景中面临挑战，表现为显著的性能下降（见图 2(a) 和 2(b)）。此外，随着预测长度的增加，一些模型会遭受严重的计算效率问题。



(a) Long-term Forecasting (NMAE). (b) Long-term Forecasting (CRPS).

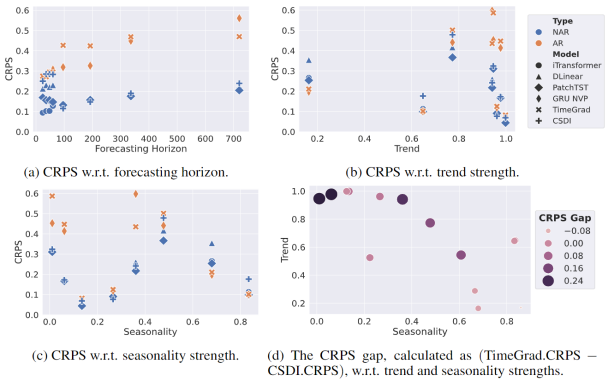
(c) Short-term Forecasting (NMAE). (d) Short-term Forecasting (CRPS).

图2：使用 ProbTS 对经典时间序列模型进行基准评测

这些观察结果表明，当前已有的预测模型中仍然缺乏适合短程预测的有效架构设计；另外刻画复杂数据分布的能力对于这些预测模型的能力而言及其重要。同时，目前的长程分布预测在性能和效率方面都面临着重大挑战。

随后，研究员们在各种预测场景中比较了自回归（AR）和非自回归（NAR）解码方案，以突出它们在预测长度，以及面对不同

趋势性和周期性时序数据方面的优势与劣势。



(a) CRPS w.r.t. forecasting horizon. (b) CRPS w.r.t. trend strength.

(c) CRPS w.r.t. seasonality strength. (d) The CRPS gap, calculated as (TimeGrad.CRPS − CSDI.CRPS), w.r.t. trend and seasonality strengths.

图3：使用 ProbTS 比较自回归和非自回归解码方案

研究员们发现，目前几乎所有的长程点预测模型都在使用非自回归解码方案进行多步预测输出，而概率预测模型则更平衡地使用自回归和非自回归方案。从数据特性视角出发，两种方案的差异可能源于：

- 预测长度影响：图 3(a) 显示，随着预测长度的增加，AR 解码与 NAR 方法相比表现出更大的性能差距，表明 AR 可能受到错误累积的影响。

- 趋势性强度影响：图 3(b) 将性能差距与趋势性的强度联系起来，表明强烈的趋势效应可能导致 NAR 和 AR 模型之间的显著性能差异。当然也存在例外情况，即使趋势性强，基于 AR 的模型也未必出现大幅度性能下降。

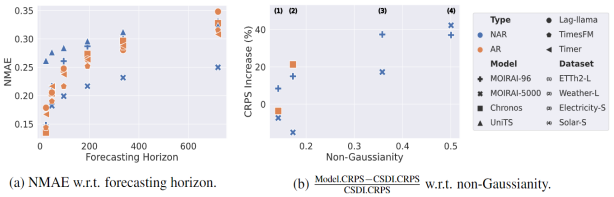
- 周期性强度影响：图 3(c) 通过引入周期性强度作为另一个因素来解释这些例外。令人惊讶的是，基于 AR 的模型在具有强周期性模式的场景中表现更好，这很可能是由于它们在这种情况下具有更高的参数效率。

- 趋势性和周期性的综合影响：图 3(d) 展示了趋势性和周期性对性能差异的综合影响。

基于此，研究员们指出，不同研究分支选择 AR 和 NAR 解码方案主要是由它们所关注的预测场景中特定的数据特性所驱动的，这也解释了大多数长期预测模型对 NAR 解码范式的偏好。然而，这种对 NAR 的偏好可能忽略了 AR 的优势，特别是 AR 在处理强周期性方面的有效性。由于 NAR 和 AR 各自拥有独特的优势，未来的研究可以探索两者的平衡之道，并改善它们的弱点。

时间序列基础模型的评测结果与分析

研究员们还使用 ProbTS 框架将分析扩展到最新涌现的时间序列基础模型上（参见图4），不仅评估了这些模型在各种预测长度内的表现，还检验了它们的分布预测能力。



(a) NMAE w.r.t. forecasting horizon. (b) $\frac{\text{Model.CRPS}-\text{CSDI.CRPS}}{\text{CSDI.CRPS}}$ w.r.t. non-Gaussianity.

图4：使用 ProbTS 进行时间序列基础模型评测

评测结果表明：

- AR 解码在扩展预测长度时存在局限性（图 4(a)），这可能是由于时间序列具有数值连续的特性，与语言建模中 AR 方法操作在离散空间中的情况不同，AR 解码方法在时序预测上会遇到更加严重的误差累积问题。
- AR 和 NAR 模型在短程场景中可以提供相当的性能，有时基于 AR 的模型甚至会优于它们的 NAR 对手。
- 当前的基础模型在分布预测方面支持有限，通常使用预定义的分布预测函数（例如：MOIRAI）或在值量化空间中用离散分布来近似建模（例如：Chronos）。这一点可以通过其与经典概率模型 CSDI 在比较捕捉复杂数据效果时发现（图 4(b)）

总而言之，虽然当前基于 AR 的基础模型在短程场景中表现优异，但它们的性能在面临更长的预测长度时显著降低，这表明时间序列数据预测，尤其在长程预测场景中，需要独特的处理来优化 AR 解码。同时，上述研究再次证明了准确刻画复杂数据分布的能力仍然是时间序列基础模型中亟需提升的关键领域。

未来方向：视角、模型和工具的三重更新

基于对现有方法的评测与分析，研究员们提出了在时间序列预测模型领域未来最重要的若干研究方向，在这些方向上的深入探索将会对各个行业的关键预测场景产生重大影响。

方向1：采用全面视角。研究员们指出，在开发新模型时有必要采用全面的视角审视前述的核心预测需求。这将帮助我们重新思考不同模型的方法选择，迭代它们的优势和劣势，并促进更多样化的研究探索。

方向2：创建通用模型。ProbTS 的研究引发了一个基本问题，即能否开发出一个满足所有核心预测需求的通用模型？或者是否应该分别处理不同的预测需求，为每种需求引入特定技术？

研究员们认为，虽然很难给出一个明确的答案，但当前的发展趋势可能倾向于创建一个通用模型。在设计该模型时，需要考虑输入表示、编码架构、解码方案和分布估计模块等问题。此外，未来还需要研究该通用模型如何应对高维数据和嘈杂场景中的分布预测（特别是对于长程预测），并探索如何利用 AR 和 NAR

解码方案的不同优势，同时规避它们各自的弱点。

方向3：开发研究工具。未来应进一步加强对相关研究工具的开发。目前研究中使用的 ProbTS 框架已经开源，研究员们希望通过这一框架吸引并凝聚研究社区的集体力量，从而促进时间序列预测领域的进步。

通过解决这些问题，微软亚洲研究院的研究员们将不断探索时间序列预测研究的边界，致力于研发出更加稳健、多功能且能够应对各种实际工业预测场景和挑战的时间序列模型。未来，完善的预测模型将极大激发多个行业的潜力，推动资源的高效利用、优化决策流程以及提升运营效率，从而加速产业智能化发展，并提升人们的生活品质。

相关链接：

ProbTS: Benchmarking Point and Distributional Forecasting across Diverse Prediction Horizons
论文链接：https://arxiv.org/abs/2310.07446v4
GitHub链接：https://github.com/microsoft/ProbTS

科研第一线

nnScaler：重塑深度学习并行策略，大幅提升训练效率（OSDI 2024）

深度学习在图像、语音识别、自然语言处理等领域已展现出巨大的应用价值，但随着模型规模的不断扩大，训练变得耗时且昂贵。设计最优的并行策略组合以提高其在多设备上的执行性能是目前该领域的一大挑战。对此，微软亚洲研究院提出了nnScaler 技术，通过并行化原语和策略限定搜索的方法寻找最佳并行策略组合。这一尝试为寻求深度学习并行策略最优化提供了方案和工具，有效应对了当前的深度学习训练效率难题。

扫描二维码查看文章

大语言模型应用如何实现端到端优化？（OSDI 2024）

当前大语言模型应用大多依赖公共 LLMs 服务提供的 API，但这些以请求为中心的 API 难以有效优化整个应用流程，进而影响任务的端到端性能。微软亚洲研究院为此开发了一个专注于 LLMs 应用端到端体验的服务系统 Parrot，它具有减少网络延迟、提高吞吐量、减少冗余计算等优势。Parrot 可以通过引入语义变量，向公共LLMs 服务公开请求关系，从而开辟了 LLMs 应用端到端性能优化的空间。

扫描二维码查看文章

如何理解和探索大模型的多语言能力？（ACL 2024）

大语言模型在未使用多语言平行语料库进行预训练的情况下，依然表现出了卓越的多语言能力。但大模型如何处理多语言文本的底层机制仍是一个具有挑战性的问题。对此，微软亚洲研究院联合中国人民大学提出了语言激活概率熵，用于识别大模型中的语言特定神经元。该研究为理解和探索大模型的多语言能力提供了重要依据。

扫描二维码查看文章

VALL-E 2，大幅提升语音大模型的稳健性与自然度

文本到语音合成（Text-to-Speech，TTS）是一种将书面文字转化为自然语音的技术，在提高无障碍性、增强跨语言交流等方面发挥着重要作用。微软亚洲研究院此前推出了第一个离散编码的语音大模型 VALL-E，并在此基础上通过重复感知采样和分组编码建模技术将其升级为 VALL-E 2 版本。新版本突破了语音稳健性、自然度和说话人相似度方面的界限，让零样本 TTS 性能在 LibriSpeech 和 VCTK 数据集上与人类水平相近。

扫描二维码查看文章

代码摘要、生成、翻译、修复全覆盖... WaveCoder开启代码智能新篇章

代码大语言模型(Code LLMs)通过自动生成和补全代码,可以帮助开发者加速实现功能。但目前针对代码大语言模型的指令微调方法主要集中在传统的代码生成任务上,忽略了模型在处理复杂多任务场景中的表现。为此,微软亚洲研究院开发了 WaveCoder 模型,其使用包含19,915个指令、涵盖4个代码任务的数据集 CodeSeaXDataset 进行训练,在代码摘要、生成、翻译和修复等多个代码任务的基准测试中显著优于其他开源模型,具有更强的泛化能力。目前WaveCoder已开源,希望助力开发者更高效的编程。

扫描二维码查看文章

MedVTAB:大规模医学视觉任务适应基准

综合性的医学视觉任务适应性基准数据集 Med-VTAB 涵盖了168万张医学图像,包括10个重要器官和5种在真实世界医学场景中具有挑战性的模态,使其成为最广泛的同类基准之一。

对齐视觉模型与人类美学:算法与评估

研究员们评估了各种提示词下的大语言模型和美学模型,证明了大语言模型带来美学理解的有效性和美学模型所包含图像先验的有效性和互补性。

GLC:基于生成式特征编码的极低码率图像编解码器

GLC是一个可以在生成式 VQ-VAE 的特征空间进行编码的模型,在多个测试基准中实现了最高的压缩性能。通过利用特征空间,GLC 在压缩图像的同时还能实现图像恢复、风格迁移等功能。

MH-MoE:多头混合专家网络

研究员们提出了多头混合专家网络 MH-MoE。通过将每个输入的令牌分割成多个子令牌的方法,MH-MoE具有更高的专家激活效率、更精细的理解能力。

扫描二维码查看文章

基于微调大语言模型的生成式推荐系统(ACL 2024)

微软亚洲研究院和深圳大学合作开发了以用户为中心的新一代推荐系统。该系统由大语言模型驱动,能更自然地理解用户需求的动态变化并提供更加精准的个性化服务。

大语言模型驱动的数据科学代理的基准测试(ACL 2024)

研究员们提出了新型基准框架DSEval,通过引入新的注释过程和语言,显著地提高了基准的可扩展性和覆盖范围。目前该框架和数据集已开源。

BitDistiller:通过自蒸馏释放低于4比特大模型的潜力(ACL 2024)

BitDistiller是一个基于自我蒸馏的 QAT 框架,采用了定制的非对称量化和 Clipping 技术来提升量化效果。该方法不仅在资源受限的设备上实现了高效部署,且只需较少的训练数据和资源。

PIN：使用强化学习优化得到可解释提示词（ACL 2024）

研究员们提出了使用 Tsallis 熵来约束强化学习过程 PIN，从而在采样和价值函数估计阶段关注出现概率最高的候选提示词，加快了对提示词价值的评估。

提高大语言模型在事件关系逻辑预测中的表现（ACL 2024）

研究员们提出了几种提升大语言模型逻辑推理能力的策略，包括：生成式方法、检索式方法、微调式方法，希望可以为未来设计有效的方法以及如何将大模型应用到实际任务中提供了新的思路和解决方法。

E5-Mistral：大语言模型增强的文本嵌入（ACL 2024）

该研究从两个方面挖掘了大语言模型在文本嵌入方面的潜力。在嵌入模型定制化方面，E5-Mistral 支持通过自然语言来描述当前的嵌入任务，可以在不更改模型参数的前提下，定制化嵌入模型的行为。相关模型已开源并受到广泛关注。

扫描二维码查看文章

通过扩展式模态，对齐推动多模态感知

为了对齐多种感知模态，，研究员们提出了 BABEL 框架，包括神经网络架构、数据准备与处理，以及训练策略。作为一个可扩展的预训练多模态感知神经网络，BABEL 目前对齐了六种广泛应用的感知模态。

基于扩散模型引导的元智能体，实现可控金融市场生成

研究员们提出了可控金融市场生成问题，并构建了一个名为扩散模型引导的元智能体模型 DiGA。DiGA能够有效地进行可控金融市场生成，使生成的金融市场贴近控制目标，且由 DiGA 模型生成的金融市场具有优越的保真度。

NutePrune：采用多个教师模型对大语言模型进行高效渐进式剪枝

NutePrune是一种高效的结构化剪枝方法。其利用多种不同稀疏度的教师模型逐步指导学生模型学习，从而缩小教师和学生之间的能力差距，提高剪枝效果。

T10：分布式内存AI芯片的计算新模式（SOSP 2024）

T10是首个针对分布式内存架构 AI 芯片的深度学习编译器，充分利用了核心间通信带宽。在真实的分布式内存架构芯片 Graphcore IPU 上，相比于现有深度学习编译器和计算库，T10取得了最高3.3倍的加速，并可以支持更大的模型规模或者数据规模。

扫描二维码查看文章

执业医师转型人工智能研究员，王子龙说“跨”才是关键

生命健康是人类永恒的探索主题，也是医疗工作者不懈追求的使命。迈入全新的人工智能时代，如何让机器学习算法和人工智能大模型助力医疗健康行业发展，是学术界和产业界共同关注的议题。然而，从计算机领域看医学行业，与从医学需求出发寻找技术突破之间必然存在着认知偏差。身为具有执业医师资格的医学博士，微软亚洲研究院（上海）高级研究员王子龙对此有哪些独到的见解？他又将如何在人工智能与医疗健康之间架起创新的桥梁？

人工智能是一个充满活力的领域，每隔几年就会涌现一些引人注目的新技术，引领着产品和应用的新潮流。过去的十几年中，多次的技术革新浪潮给各行各业都带来了大量机遇。

“作为一名医学生，我会从医学角度出发，探索创新技术在医疗领域的应用潜力，如卷积神经网络可以推动医学影像处理的进步，基础模型能够扩展医疗产品的功能并增强人机交互。然而，医疗行业是一个严谨且相对保守的领域，对于新技术的应用和融合有着更高的标准和要求。这就需要跨领域的研究者将两者融合起来，我希望自己可以成为这样一座桥梁。”这是王子龙决定从技术应用回归学术研究，并加入微软亚洲研究院的主要原因。

图1: 微软亚洲研究院（上海）高级研究员王子龙

“在我之前的职业生涯中，无论是在大型企业开发医疗产品，还是在初创公司带领团队开发前沿应用，产品研发首要考虑的是短期时效性与利益，这限制了对更深层次研究课题的探索。我更喜欢探索医学与人工智能的交叉领域，在研究机构中，我能更自由地专注于解决长期性的重大问题。”

正因为此，2023年王子龙加入了微软亚洲研究院（上海），开始了他的跨领域研究之旅。从执业医师到算法研究员，王子龙带着探索性视角，致力于将人工智能的前沿技术应用于医疗健康领域，寻找创新的解决方案。王子龙的研究不仅关注技术突破本身，更着眼于如何将这些技术与医疗实践相结合，以实现更精准的诊断、更有效的治疗和更广泛的健康监护。

从医疗领域审视技术应用：洞见症结，推动人工智能大模型创新

王子龙深知，医疗行业对人工智能技术有着迫切的需求，尤其是在提高诊断准确性和临床效率方面。尽管人工智能在视觉和自然语言处理领域取得了显著进展，但在医疗领域的应用仍有很大的发展空间。例如，大语言模型在问答机器人中的应用较为普遍，但在特定疾病领域的深入应用还需进一步探索。王子龙认为，其中的关键在于如何让这些模型更好地辅助医生进行诊断和实践，尤其是在医学报告的生成中，模型需要对临床内容有深刻的理解和准确的表达，而对于聊天对话或生成文字的优美性与流畅度只是次级需求。

之前的图像算法大多以分类的形式，表示在 X 光片中的肺炎或骨折等阳性发现。然而，随着大模型，特别是多模态大模型的发展，医学影像分析迎来了新的变革。这些模型能够将 X 光片等医学影像转化为自然语言或类似报告的形式，使得结果更易于医生阅读。王子龙和同事们认识到，这种新方法需要更为复杂的评估手段以确保其临床意义和准确性，基于分类标签的或者传统的文本匹配的评估手段已不能满足对新方法的评估。

为此，他们设计了一系列数据集和方法，来判断生成内容的质量，尤其是临床意义上的准确性。同时也设计了能够增强模型交互性的数据集和方法，可以让医生通过自然语言指令修改既往报告、对比历史报告，并添加相关信息，从而使模型能够整合其他病理检查信息，方便医生重新审核病患图像和报告，辅助医生进行全面的诊断。

再比如，在自然图像领域表现出色的视觉模型，可能在医学影像分析中的表现并不尽如人意。医学影像的特殊性在于其对细节的要求极高，图像上面积占比极小的病灶决定了整个图像的类

别，这与自然图像分析中常见的任务，如识别图像中的主要物体及其位置，有着很大的不同。例如，需要在包含数百万像素的胸片中有能力识别出仅有10*10像素大小的病灶，并确定其性质和属性。在医学影像分析的过程中，图像还常常需要被缩放（resize）到特定尺寸以适应模型的输入和输出需求。这是一项对细节处理能力要求极高的任务，但是当前的大规模采样方法在处理医学图像时可能会丢失关键的诊断信息，直接影响了诊断准确性。针对这一问题，王子龙和同事们正在探索改进医学图像领域的视频和图像编码器技术，以期把高分辨率的二维与三维的医学影像以一种更合适的方式引入图像处理。

从技术角度看医疗行业：让人工智能更有的放矢

尽管此前身处医疗领域，但王子龙对计算机科学也有着深入的研究，在加入微软亚洲研究院之后，他对计算机领域的前沿技术有了更深入的理解。这让他能够洞察到这些技术在医疗领域的应用潜力，从而促进技术与医疗难题之间的有效对接，释放出创新技术的真正价值。

在微软亚洲研究院，王子龙了解到音频处理技术带来的更多可能性，经过与医学专家的深入讨论，他认为这项技术在心脏和血管健康管理方面具有巨大的应用潜力。例如，通过监测和分析血液流动的声音，音频技术可以用于无创地检测血管状态，为相关疾病的早期发现和跟踪提供了新的可能性。

王子龙还注意到了微软亚洲研究院在无线通信与无线感知领域的研究成果，一旦应用于移动设备或可穿戴设备，将极大地推动远程健康监测的发展。患者可以在家中自行监测多种疾病的变化，及时识别潜在的健康风险，实现更加主动的健康管理，进而提高医疗服务的覆盖范围和效果。

王子龙认为，微软亚洲研究院自由开放的研究环境、多元化的技术路线，以及汇聚了世界一流人才，为他在人工智能与生命科学和医疗健康领域的交叉研究提供丰富的资源和合作机会。

跨领域研究与跨学科学习的精髓在于“跨”

高中时，王子龙就对多学科学习有着浓厚的兴趣，在生物、物理和计算机竞赛中都取得了优异的成绩，并因此被保送至复旦大学上海医学院。在进行医学专业课程学习的同时，复旦大学开放的学习环境也让他可以去“蹭”更广的课程，实现了医学与计算机科学学习的兼顾，为他的跨学科知识储备奠定了坚实的基础。

随着对医学知识不断地深入探索，获得了执业医师资格并即将取得肿瘤学博士学位的王子龙认识到，尽管当时医疗领域已经有了一些成熟的技术和工具，但医学的进步仍需更多创新的模型和技术支持，新的人工智能技术也将在医学实践中形成更大的影响力。在见证了首个基于人工智能的医学影像诊断产品 IDx-D 获得 FDA 批准后，王子龙更加坚定了这一信念。

博士毕业后，王子龙加入了一家知名的企业研究机构，投身于人工智能在医学领域的应用探索，包括图像识别技术在疾病诊断中的应用，并参与开发了相关的辅助诊断工具。此后，他转战初创企业，带领团队开发了针对眼底和胸部 CT 的辅助诊断产品，并在这一时期入选了科学和医疗健康领域2020福布斯中国30岁以下精英榜和2021胡润 U30 中国创业领袖。

图2：王子龙曾入选科学和医疗健康领域2020福布斯中国30岁以下精英榜

无论身处哪一发展阶段，王子龙都在时刻充实自己的跨学科知识。他认为，在人工智能行业落地的过程中，跨领域合作和跨学科人才至关重要。“通过与产业领域专家的深入合作，充分理解行业需求，以开放心态发现并解决行业关键问题，才能共同设计出针对医疗目标和特定场景的学习框架与模型，使 AI、大模型、RAG（检索增强生成）等技术更好地服务于医疗发展。”王子龙说。

跨学科知识的积累非一朝一夕之事，需要兴趣的驱动和勇于跨界的勇气。“兴趣是最好的老师”恰当地说明了兴趣如何深刻地影响我们对学习的态度和效果。当人们投身于自己感兴趣的领域时，他们往往会更加投入和专注，从而加速学习过程，更容易进入“心流”状态。王子龙认为跨学科的精髓在于“跨”，需要以兴趣为导向，勇于跨越学科界限，不设限地探索未知领域。虽然初入新领域可能会感到陌生与不安，但当找到交叉领域的兴趣点后，就会逐步形成平滑的学习曲线，从而实现跨学科知识的积累与应用。

顶尖高校优秀学子齐聚微软亚洲研究院新星科技节，论道科研！

2024年8月12日，微软亚洲研究院新星科技节成功举办，此次活动汇聚了来自清华大学、北京大学、斯坦福大学、卡耐基梅隆大学、剑桥大学、牛津大学、多伦多大学等国内外多所顶尖高校的优秀学生。除了微软亚洲研究院的实习生和联合培养博士生，其他高校的顶尖学子也在此汇聚一堂，他们通过口头报告和海报展示的形式，分享了自己的最新研究成果和科研洞见，共同呈现了一场学术与思想的盛宴。

微软亚洲研究院新星科技节的独特之处在于，它不仅仅是一次简单的学术交流，更将计算机领域内的新星连接起来，为同学们构建了一个跨校、跨领域、跨文化的学术网络，让他们在思想的碰撞中拓宽视野、激发灵感，在与同侪的并肩前行中，进一步明确未来的发展方向。

图1：海报讲者合影

微软亚洲研究院副院长杨懋博士在开幕致辞中表示："新星科技节既是展示个人科研成果的平台，更是思想交融、创新诞生的摇篮。我们期待通过这种深度交流，连接更多具有全球视野和创新潜力的青年才俊。"这也是本次活动的核心理念——通过连接不同背景、不同领域的年轻学者，推动计算机科学的边界不断拓展。资深学术合作经理孙丽君女士补充："通过新星科技节，我们希望让更多的青年学者在这里找到志同道合的伙伴，开启未来的科研新篇章。"

图2：微软亚洲研究院副院长杨懋博士

图3：微软亚洲研究院资深学术合作经理孙丽君

在主题报告环节，八位同学围绕大语言模型、具身智能、生成式人工智能、社会责任人工智能等主题进行了学术报告，与在场的同学们交流互动。此外，30位同学通过学术海报展示了他们在各自研究领域的创新成果，内容涵盖了广泛的研究方向。

图4：微软亚洲研究院人才项目负责人张津为学生颁发证书

通过此次活动，同学们获益颇多。来自北京大学图灵班博士一年级的严汨同学进行了主题报告和海报展示。谈及来参加分享的原因，她说自己"仰慕微软亚洲研究院已久"，也希望借此机会了解研究院目前的工作。"很惊喜研究院的方向与我非常一致，因此也很期待未来可以有机会来实习。"严汨觉得整个活动非常有价值，"在海报环节，有很多同学的提问很有道理，可以作为很好的follow-up 去弥补现有工作的不足。"

图5：北京大学严汨（左一）

中国科学院大学博士三年级的王鸿钰同学已在微软亚洲研究院实习一年有余，他在本次活动中报告了他在研究院的最新科研工作。这次活动给了他一次极佳的拓宽视野的机会，"大家的科研主题非常多元，有 LLMs、Embodied AI，还有一些 societal 相关的课题，有利于帮助我们打开视野，也能观察到一些更加前沿的东西。"王鸿钰说，"借此机会，我深入了解了很多同学的科研内容，并结识了许多志同道合的伙伴，相信未来会有进一步的交流和合作。"

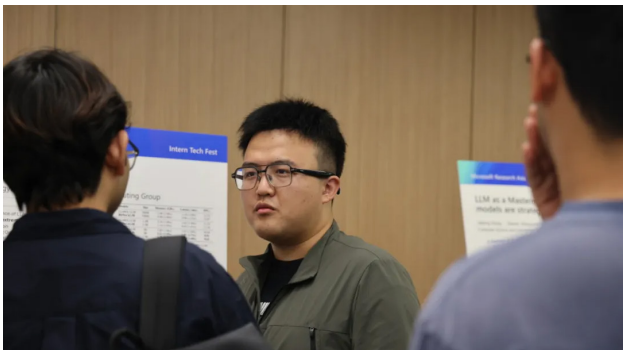


图6：中国科学院大学王鸿钰（左二）

来自北京大学计算机学院国际班的 Ben Redhead 说着一口流利的中文，非常畅快地与同行们交流。Ben 的本科专业是数学，他欣喜地发现微软亚洲研究院不仅有很多理论研究，还有许多研究将数学演算结合其中。他特别提到在午餐会时，他与郑顺研究员交流，发现彼此有相似的研究方法和方向，这次交流让他受益匪浅，获得了不少科研灵感。他也通过准备海报内容，很好地梳理了自己的研究思路，为三天后的研究发表打了一个漂亮的前战。

图7：北京大学 Ben Redhead（左一）

来自牛津大学的林芳如不久前刚在 ICML 2024 介绍了自己的工作，与国际顶会相比，她认为在本次活动中与同龄人交流学术让她收获最大。"我之前就有关注 1-Bit LLM 的研究，今天作者王鸿钰的 poster 就在我旁边，他给我很清楚地讲了他的研究。"微软亚洲研究院浓厚的科研氛围，也是她在实习中印象最深刻的部分之一，"大家在食堂排队的时候还在讨论科研问题。我主要做 LLM 这个方向，在这里我有机会看到不同领域的工作和 LLM 的结合，许多是我没有想象过的研究方向。

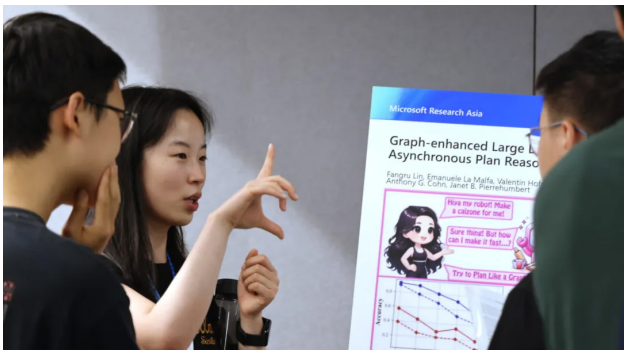


图8：牛津大学林芳如（左二）

图9：微软亚洲研究院（上海）线上联动

量子位 | 只激活3.8B参数，性能比肩同款7B模型！训练微调都能用，来自微软

只需激活60%的参数，就能实现与全激活稠密模型相当的性能。微软亚洲研究院的一项新研究，实现了模型的完全稀疏激活，让推理成本大幅下降。

而且适用范围广泛，无论是从头训练、继续训练还是微调，都能提供有效支持。

该方法名为Q-Sparse，在神经元级别上实现了模型稀疏化，相比于其他方式粒度更细，在相同推理开销下，无论性能还是稀疏率都更好。

名称之中，Q指的是量化（Quantization），意味着它除了普通模型之外，也兼容量化技术，适用于各种量化方式的模型。

作者进一步表示，如果把Q-Sparse与模型量化技术结合，还可以实现更大程度的降本增效。

另外在研究Q-Sparse的同时，团队也对参数规模、稀疏率和模型性能三者之间的关系进行了深入探寻，并发现了适用于模型推理优化的“Scaling Law”。

有网友认为，这项技术确实不错，而且比ReLU要更好。

还有人开启了许愿模式，表示如果（AMD的）ROCm能比英伟达更快支持这项技术就好了。

用Top-K函数实现稀疏化

Q-Sparse所做的最核心的操作，是对输入的张量应用Top-K稀疏化函数。

具体来说，Transformer架构在注意力层和前馈层中都使用nn.Linear线性层（矩阵乘法）进行投影，可以表示为Y=X·W^T。（其中X就是输入张量，W代表其权重，Y为输出张量）

Q-Sparse中，对于一个输入激活张量X，首先会计算其绝对值|X|并进行排序，找出其中绝对值最大的K个元素。

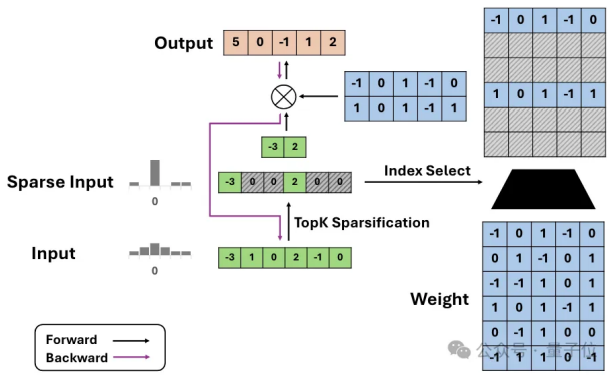
这里的K是预先设定的超参数，决定了稀疏化的程度。

之后Q-Sparse会创建一个与X形状相同的二进制掩码张量M，对于一系列|X|中绝对值最大的K个元素对应的位置，将M中的相应位置设置为1，其余位置设置为0。

接着，将输入张量X与掩码张量M进行Hadamard积（逐元素相乘）运算，就得到了稀疏化的张量X_sparse。

在前向传播过程中，稀疏化后的张量X_sparse将代替原始的输入张量X参与后续的计算（如矩阵乘法）。

由于X_sparse中大部分元素已经被设置为零，因此可以显著减少计算量和内存带宽需求。



在反向传播过程中，Q-Sparse使用了直通估计器（Straight-Through Estimator，STE）来计算Top-K函数的梯度。

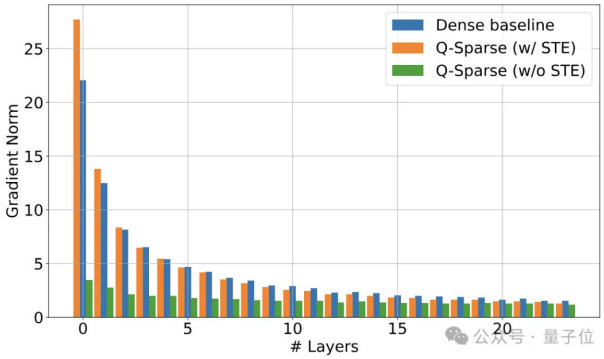
传统的训练方式中，通常需要计算损失函数对网络参数的梯度，并使用梯度下降法更新参数以最小化损失。

但当网络中存在量化、Top-K等一些不可微的操作时，梯度的计算就会遇到问题，因为这些操作的输出对输入的梯度在大多数点上都是0，导致梯度无法有效传播。

STE通过直接将梯度传递给稀疏化之前的张量，避免了梯度消失的问题。

一般的反向传播中，损失函数L对x的梯度∂L/∂x=∂L/∂y□∂y/∂x，但由于不可微分无法直接计算。

STE的解决方案是只计算损失函数对稀疏化张量y的梯度，然后将其直接复制给原始张量x，也就是直接将∂L/∂y作为∂L/∂x的估计。



对于前馈层，Q-Sparse使用平方ReLU函数代替常规的ReLU激活函数，平方运算可以进一步提高激活的稀疏性（⊙表示Hadamard积）。

$$\mathrm{ReLU}^2\mathrm{GLU}(\mathbf{X}) = \mathbf{X}\mathbf{W}_{\mathrm{up}}^T \odot \mathrm{ReLU}^2(\mathbf{X}\mathbf{W}_{\mathrm{gate}}^T)$$

另外，为了适配量化模型，Q-Sparse在应用Top-K稀疏化之前，会先对输入张量进行量化，以确保稀疏化操作与量化表示兼容，其函数表示如下：

$$\mathrm{Q}(X) = \mathrm{RoundClip}(\frac{\mathbf{X}}{\gamma+\epsilon}, -128, 127)$$

$$\gamma = \max(|\mathbf{X}|)$$

$$\mathrm{RoundClip}(X, a, b) = \min(\max(\mathrm{round}(X), a), b)$$

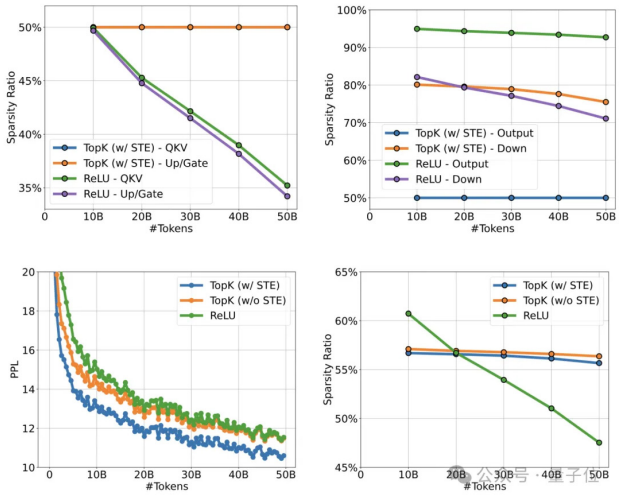
其中，ε是一个小常数，用于避免出现分母为零的情况。

特别的，对于1-bit量化的权重，Q-Sparse使用以下量化函数，其中α是权重张量W的平均绝对值。

$$\mathrm{Q}_w(W) = \mathrm{RoundClip}(\frac{\mathbf{W}}{\alpha+\epsilon}, -1, 1)$$

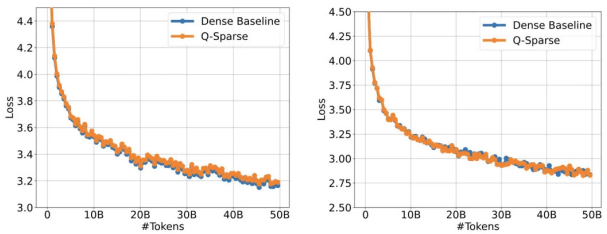
60%激活参数达到相同效果

对比实验表明，无论是稀疏率还是模型表现，Q-Sparse都显著优于此前的ReLU方法。



针对Q-Sparse的具体效果，作者对其在从头训练、继续训练和微调三项任务上的性能进行了评估。

从头训练实验使用的模型为Llama，结果在700M和7B模型上，使用70% top-K(即40%的整体稀疏率)的Q-Sparse可以达到与密集baseline相当的训练损失。



(a) 700M model size (b) 7B model size

继续训练的目的是将稠密模型稀疏化，这里的实验对象是Mistral-7B。

结果，在激活参数为2.9B和3.8B的情况下，模型在ARC、MMLU等数据集中的得分均未发生明显下降。

Models	Activated	ARC	HS	MMLU	WG	TQA	Avg.
Dense Baseline	7.0B	61.8	81.4	59.8	77.5	42.7	64.6
ReLUfication [MAM+23]	5.0B	57.2	78.8	54.7	74.7	38.8	60.8
dReLU Sparsification [SXZ+24]	5.4B	59.2	78.0	54.0	75.8	38.3	61.0
Q-Sparse (this work)	2.9B	59.0	79.0	55.6	74.0	41.0	61.7
	3.8B	60.5	80.7	58.0	75.9	43.5	63.7

Table 1: The results of the continue-training for Q-Sparse and the baselines on the end tasks.

在微调实验中，对于Qwen-7B和Mistral-7B两种模型，Q-Sparse显示出了与继续训练相似的结果，用60%左右的激活参数实现了与密集模型十分接近的表现。

Models	Activated	ARC	HS	MMLU	WG	TQA	Avg.
Qwen1.5-4B	3.2B	42.8	68.2	53.6	67.1	47.9	55.9
Qwen1.5-7B	6.5B	47.7	74.6	61.5	71.4	50.7	61.2
Q-Sparse	3.6B	46.3	72.6	59.1	67.5	50.3	59.2
	4.1B	47.9	73.2	59.2	69.4	51.1	60.1
Mistral-7B	7.0B	62.5	82.6	61.2	77.6	50.3	66.8
Q-Sparse	3.8B	60.5	81.5	60.0	77.1	50.5	65.9
	4.3B	61.4	81.6	60.6	77.6	50.7	66.4

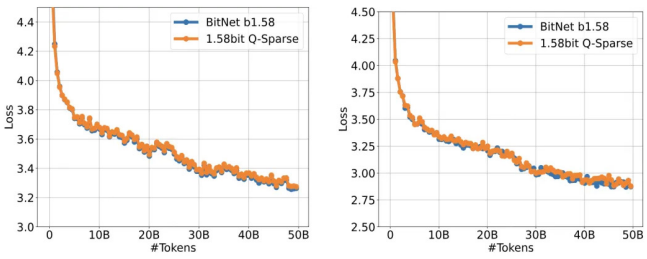
这些结果意味着，在相同的性能下，与密集模型相比，稀疏激活模型在推理过程中可以显著减少激活参数，进而降低消耗FLOPS的数量。

对于量化模型，团队在自研的BitNet b1.58模型上应用了Q-Sparse，并在多个数据集上进行了训练和评估。

可以看到，在700M和7B两种规模下，使用Q-Sparse的量化模型的收敛速度和最终损失函数值与未使用Q-Sparse的量化模型(BitNet b1.58)相当。

这说明Q-Sparse可以无缝集成到量化模型中，而不会显著影响模型的训练和收敛。

据此作者认为，将Q-Sparse与量化技术相结合，可以进一步提高大语言模型在推理阶段的效率。



(a) 700M model size (b) 7B model size

发现推理优化新“Scaling Law”

除了测评这些模型采取稀疏激活时的表现，作者也对模型性能、规模和稀疏率三者之间的关系进行了探究，并有了一些新的发现。

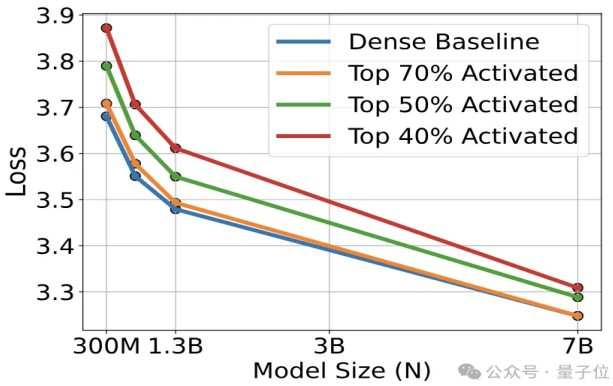
稀疏激活模型的性能缩放定律：作者发现，与密集模型类似，稀疏激活模型的性能也遵循一个幂律缩放关系。

具体来说，给定稀疏率S，模型在收敛时的损失函数值L(N,S)可以用以下公式近似：

$$L(N,S) \triangleq E + \frac{A(S)}{N^{\alpha}}$$

其中，N是模型参数的数量；E是一个常数，表示模型在无限大时的损失；A(S)是一个与稀疏率S有关的缩放因子。

这个缩放定律表明，稀疏激活模型的性能随着模型规模的增大而提高，但提高的速度会逐渐变慢。



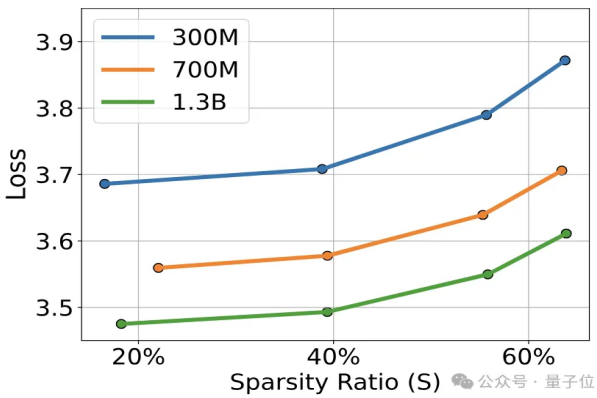
同时作者发现，模型的性能也会受到稀疏率的影响。

在参数规模与性能之间关系的部分提到，A(S)是一个与稀疏率S有关的缩放因子，可以用以下公式近似：

$$A(S) = B + C\exp\left(\frac{\beta}{1-S}\right)$$

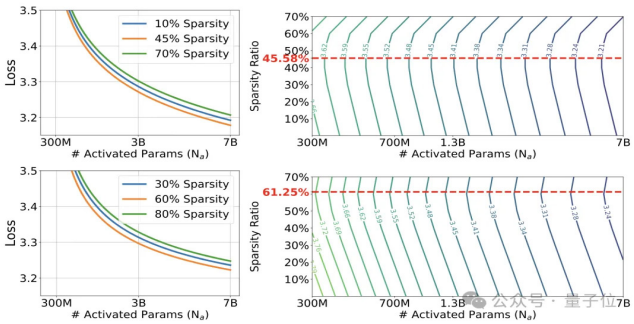
其中B和C是常数，β是一个控制指数衰减速度的参数。

这个公式表明，当稀疏率S增大（模型变得更稀疏）时，意味着更高的稀疏率会导致性能的下降，下降的速度是指数级的。



基于上述发现，作者得出了一个推理最优的稀疏率S*，能在预算（推理时的浮点操作数）一定时，实现模型损失函数值的最小化。

对于全精度（FP32）模型，最优稀疏率约为45.58%；而低精度（如1.58-bit）模型的最优稀疏率则更高，约为61.25%。



作者观察到，随着模型规模的增大，稀疏激活模型与密集模型之间的性能差距逐渐缩小。

这可以从缩放定律中得到解释：当模型规模N趋于无穷大时，稀疏激活模型的损失函数值趋于L(∞,S)=E，而密集模型的损失函数值趋于L(∞,0)=E。

这意味着，在极大规模下，稀疏激活模型有可能达到与密集模型相当的性能，为设计和训练大规模稀疏激活模型提供了一个有用的参考。

相关链接：

Q-Sparse: All Large Language Models can be Fully Sparsely-Activated
论文链接：https://arxiv.org/abs/2407.10969

AI前线|大模型端侧 CPU 部署最高提效 6 倍！微软亚研院新开源项目 T-MAC 技术解析

为增强设备上的智能性，在边缘设备部署大型语言模型（LLMs）成为了一个趋势，比如微软的 Windows 11 AI + PC。目前部署的大语言模型多会量化到低比特。然而，低比特 LLMs 在推理过程中需要进行低精度权重和高精度激活向量的混合精度矩阵乘法（mpGEMM）。现有的系统由于硬件缺乏对 mpGEMM 的原生支持，不得不将权重反量化以进行高精度计算。这种间接的方式导致了显著的推理开销，并且无法随着比特数进一步降低而获得加速。

为此，微软亚洲研究院的研究员们开发了 T-MAC。T-MAC 采用基于查找表（LUT）的计算范式，无需反量化，直接支持混合精度矩阵乘，其高效的推理性能以及其统一且可扩展的特性为在资源受限的边缘设备上实际部署低比特 LLMs 铺平了道路。

此外，当前大模型的部署普遍依赖于专用加速器，如 NPU 和 GPU 等，而 T-MAC 可以摆脱专用加速器的依赖，仅利用 CPU 部署 LLMs，推理速度甚至能够超过同一片上的专用加速器，使 LLMs 可以部署在各类包括 PC、手机、树莓派等边缘端设备。T-MAC 现已开源。

在 CPU 上高效部署低比特大语言模型

T-MAC 的关键创新在于采用基于查找表（LUT）的计算范式，而非传统的乘累加（MAC）计算范式。T-MAC 利用查找表直接支持低比特计算，从而消除了其他系统中必须的反量化（dequantization）操作，并且显著减少了乘法和加法操作的数量。

经过实验，T-MAC 展现出了卓越的性能：在配备了最新高通 Snapdragon X Elite 芯片组的 Surface AI PC 上，3B BitNet-b1.58 模型的生成速率可达每秒 48 个 token，2bit 7B llama 模型的生成速率可达每秒 30 个 token，4bit 7B llama 模型的生成速率可达每秒 20 个 token。这甚至超越了 NPU 的性能！

当部署 llama-2-7b-4bit 模型时，尽管使用 NPU 可以生成每秒 10.4 个 token，但 CPU 在 T-MAC 的助力下，仅使用两核便能达到每秒 12.6 个 token，最高甚至可以飙升至每秒 22 个 token。都远超人类的平均阅读速度，相比于原始的 llama.cpp 框架提升了 4 至 5 倍。即使在较低端的设备如 Raspberry Pi 5 上，T-MAC

针对 3B BitNet-b1.58 也能达到每秒 11 个 token 的生成速率。T-MAC 也具有显著的功耗优势：达到相同的生成速率，T-MAC 所需的核心数仅为原始 llama.cpp 的 1/4 至 1/6，降低能耗的同时也为其它应用留下计算资源。

值得注意的是，T-MAC 的计算性能会随着比特数的降低而线性提高，这一现象在基于反量化去实现的 GPU 和 NPU 中是难以观察到的。但 T-MAC 能够在 2 比特下实现单核每秒 10 个 token，四核每秒 28 个 token，大大超越了 NPU 的性能。

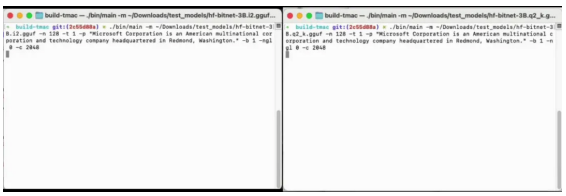


图 1：BitNet on T-MAC vs llama.cpp on Apple M2

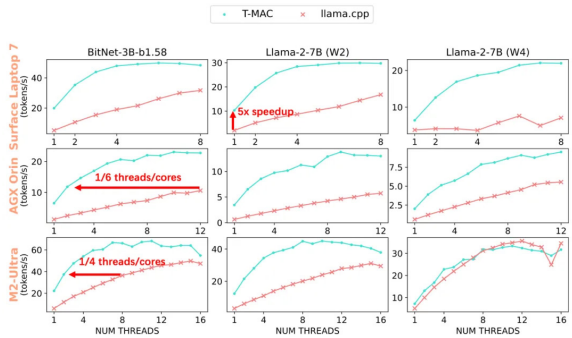


图 2：在不同端侧设备 CPU（Surface Laptop 7, NVIDIA AGX Orin, Apple M2-Ultra）的各核数下 T-MAC 和 llama.cpp 的 token 生成速度可达 llama.cpp 的 4-5 倍。达到相同的生成速率，T-MAC 所需的核心数仅为原始 llama.cpp 的 1/4 至 1/6。

矩阵乘不需乘，只需查表 (LUT)

对于低比特参数 (weights)，T-MAC 将每一个比特单独进行分组（例如，一组 4 个比特），这些比特与激活向量相乘，预先计算所有可能的部分和，然后使用 LUT 进行存储。之后，T-MAC 采用移位和累加操作来支持从 1 到 4 的可扩展位数。通过这种方法，T-MAC 抛弃了 CPU 上效率不高的 FMA （乘加）指令，转而使用功耗更低效率也更高的 TBL/PSHUF（查表）指令。

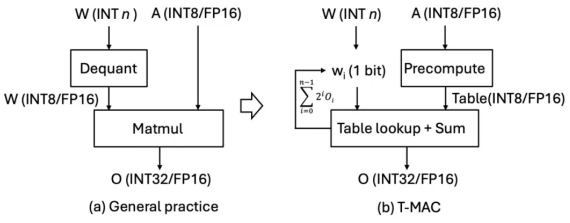


图 3：混合精度 GEMV 基于现有反量化的实现范式 vs T-MAC 基于查找表的新范式

以比特为核心的计算，取代以数据类型为核心的计算

传统的基于反量化的计算，实际上是以数据类型为核心的计算，这种方式需要对每一种不同的数据类型单独定制。每种激活和权重的位宽组合，如 W4A16（权重 int4 激活 ßoat16） 和 W2A8，都需要特定的权重布局和计算内核。例如，W3 的布局需要将 2 位和另外 1 位分开打包，并利用不同的交错或混洗方法进行内存对齐或快速解码。然后，相应的计算内核需要将这种特定布局解包到硬件支持的数据类型进行执行。

而 T-MAC 通过从比特的视角观察低比特矩阵乘计算，只需为单独的一个比特设计最优的数据结构，然后通过堆叠的方式扩展到更高的 2/3/4 比特。同时，对于不同精度的激活向量（float16/float32/int8），仅有构建表的过程需要发生变化，在查表的时候不再需要考虑不同的数据结构

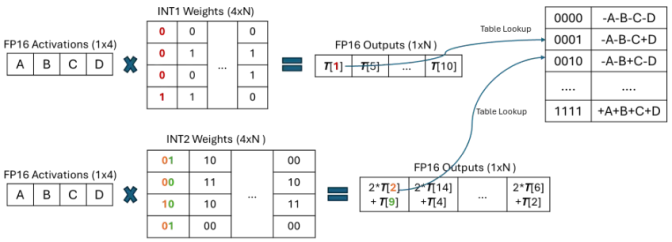


图 4：以比特为核心的查表计算混合精度 GEMV

同时，传统基于反量化的方法，从 4- 比特降低到 3/2/1- 比特时，尽管内存占用更少，但是计算量并未减小，而且由于反量化的开销不减反增，性能反而可能会更差。但 T-MAC 的计算量随着比特数降低能够线性减少，从而在更低比特带来更好加速，为最新的工作 BitNet, EfÞcientQAT 等发布的 1- 比特 /2- 比特模型提供了高效率的部署方案。

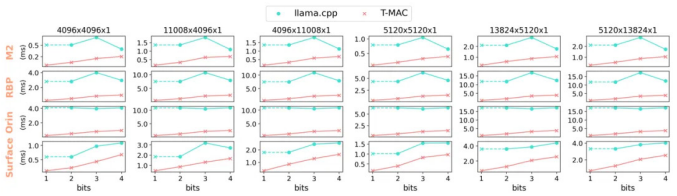


图 5：使用不同端侧设备 CPU 的单核，T-MAC 在 4 到 1 比特的混合精度 GEMV 算子相较 llama.cpp 加速 3-11 倍。T-MAC 的 GEMM 耗时能随着比特数减少线性减少，而基于反量化的 llama.cpp 无法做到（1 比特 llama.cpp 的算子性能由其 2 比特实现推算得到）。

高度优化的算子实现

基于比特为核心的计算具有许多优势，但将其实现在 CPU 上仍具有不小的挑战：(i) 与激活和权重的连续数据访问相比，表的访问是随机的。表在快速片上内存中的驻留对于最终的推理性能

尤为重要，(ii) 然而，片上内存是有限的，查找表（LUT）方法相比传统的 mpGEMV 增大了片上内存的使用。这是因为查找表需要保存激活向量与所有可能的位模式相乘的结果。这比激活本身要多得多。

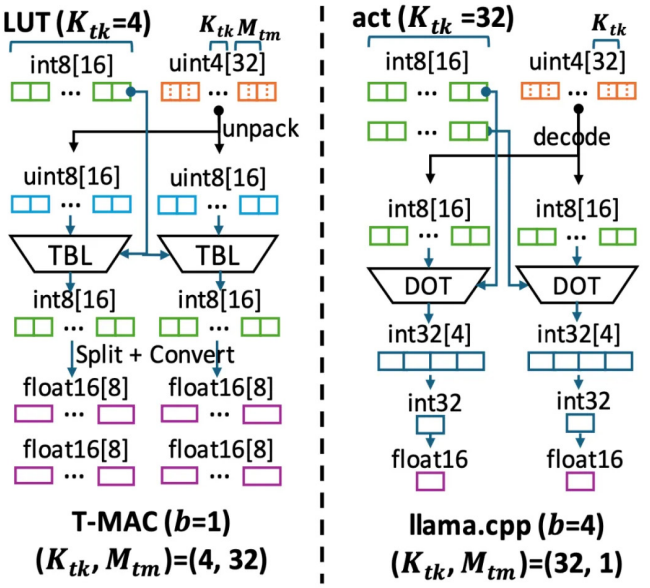


图 6：T-MAC 与 llama.cpp 在计算数据流上的不同

为此，微软亚洲研究院的研究员们深入探究了基于查表的计算数据流，为这种计算范式设计了高效的数据结构和计算流程，其中包括：

1. 将 LUT 存入片上内存以利用 CPU 上的查表向量指令 (TBL/PSHUF) 提升随机访存性能。

2. 改变矩阵 axis 计算顺序，以尽可能提升放入片上内存的有限 LUT 的数据重用率。

3. 为查表单独设计最优矩阵分块 (Tiling) 方式，结合 autotvm 搜索最优分块参数

4. 参数 weights 的布局优化
a.weights 重排，以尽可能连续访问并提升缓存命中率
b.weights 交错，以提升解码效率

5. 对 Intel/ARM CPU 做针对性优化，包括
a. 寄存器重排以快速建立查找表
b. 通过取平均数指令做快速 8- 比特累加

研究员们在一个基础实现上，一步步应用各种优化，最终相对于 SOTA 低比特算子获得显著加速：

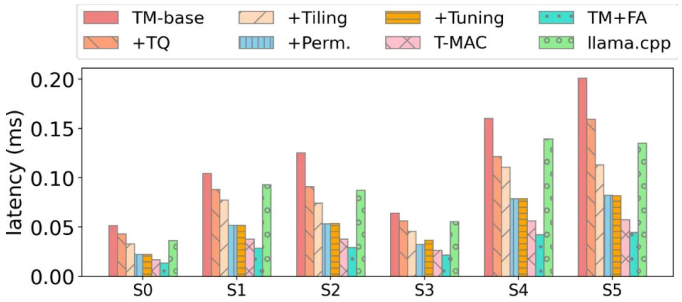


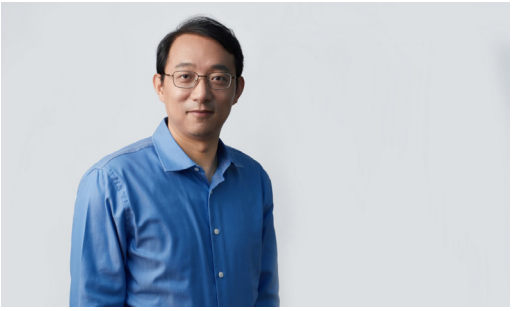
图 7：在实现各种优化后，T-MAC 4- 比特算子最终相对于 llama.cpp 获得显著加速

开源易用的工具

T-MAC 现已开源 https://github.com/microsoft/T-MAC，简单输入几条命令即可在自己的笔记本电脑上高效运行 Llama-3-8B-instruct 模型。

相关链接：

T-MAC: CPURenaissance via Table Lookup for Low-Bit LLM Deployment on Edge
论文链接：https://www.arxiv.org/pdf/2407.00088
GitHub链接：https://github.com/microsoft/T-MAC



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

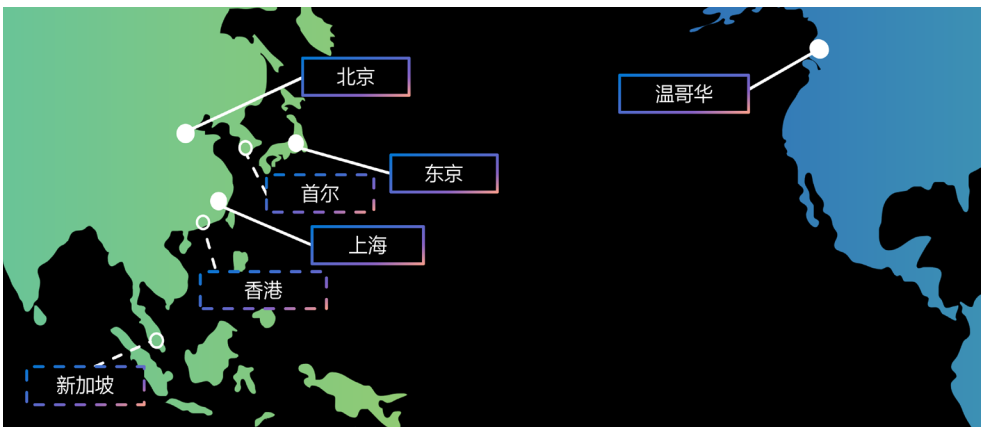
微软亚洲研究院成立于1998年，是微软公司在亚太地区设立的研究机构，在北京、上海、温哥华、东京、首尔、新加坡和香港设有实验室及研究岗位，研究方向涵盖计算基础创新、下一代智能交互、多维感知与通信、人工智能与社会福祉、科学发现与行业赋能等。通过来自世界各地不同学科和背景的多元人才的鼎力合作，微软亚洲研究院已经发展成为世界一流的计算机基础及应用研究机构。多年来，从微软亚洲研究院诞生的新技术层出不穷，对微软公司的产品创新以及全球范围的科技发展产生了深远的影响。

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