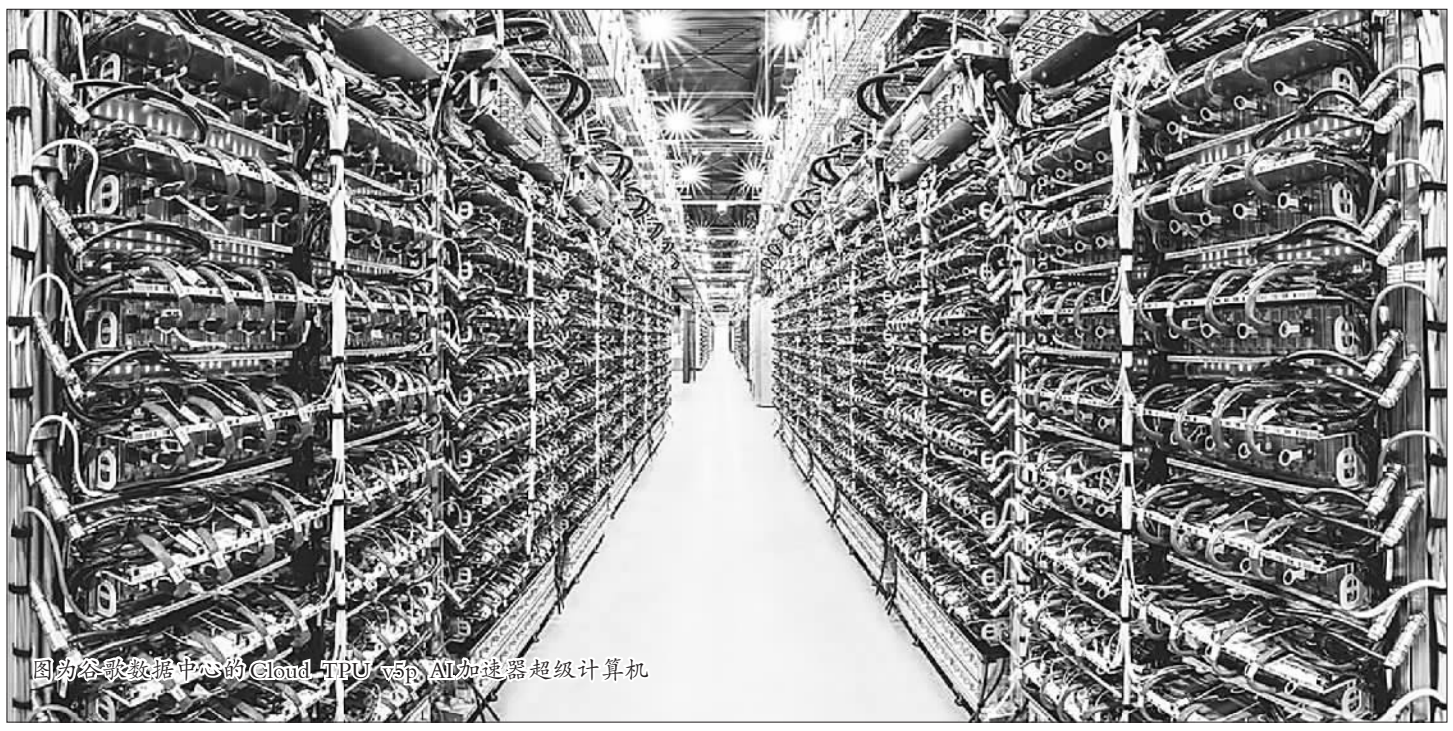
谷歌与OpenAI决战2024

特约撰稿 李佳师


图为谷歌数据中心的Cloud TPU v5p AI加速器超级计算机

谷歌与OpenAI的真正交火会发生在2024年。

2023年的最后一个月，谷歌发布了原生多模态大模型Gemini，称其在大语言模型领域的32个常用测试指标里，有30项领先于GPT-4。Gemini针对不同任务设定了Nano、Pro、Ultra三个版本，目前上线的只是Gemini Pro版，而“顶配”Gemini Ultra将于2024年年初推出。随后，谷歌发布2023年度AI研究总结，给出了“全面碾压OpenAI”的万字长文。

OpenAI首席执行官山姆·奥特曼公布的圣诞愿望清单，可以理解为OpenAI对于谷歌的“应战书”，并透露将在2024年推出GPT-5。

对此，外界认为，2024年全球AI大模型比拼的焦点是GPT-5和Gemini Ultra，虽然目前两个模型都还没有问世，无法直接比较，但两家公司在模型之争背后的数据、算力、生态的全方位较量已经开始。

1月11日凌晨，OpenAI的GPT Store（GPT应用商店）正式上线，首先面向付费用户开放，最终将直接进行创收。这意味着App Store商业模式被引入大模型领域。几乎在同一时间，谷歌DeepMind也宣布了SARA-RT、RT-Trajectory、AutoRT三项具身智能成果。

2024年元旦刚过，OpenAI与谷歌关于AI之争的火药味开始弥漫，一场事关AI核心基础的变革“山雨欲来”。

多模态数据 谁拥有更多？

对于2024年大模型的竞争焦点，业界的共识是：无论是GPT-5还是Gemini Ultra，都主打多模态，这意味着训练这些模型需要海量高质量的视频、音频、图片等数据。

那么，谷歌和OpenAI，谁拥有更多的此类资源？

360集团创始人、董事长周鸿祎在与甲子光年创始人兼CEO张一甲的对话中表示：“人才密度、算力密度和数据质量的高低是决定通用大模型胜负的关键，谷歌的人才不是问题，算力也不是问题，在数据方面拥有搜索、YouTube和Android生态系统。”有消息称，谷歌训练Gemini时所用的数据量是GPT-4的两倍之多。

Open AI同样拥有自己的生态。如果说搜索、短视频是互联网时代的“超级应用”，在大模型时代，杀手级应用很可能是“AI智能体”。从这个角度来看，Open AI的GPT商店有可能建立起强大的“AI超级应用”。据悉，目前用户自定义构建的ChatGPT助手已超过300万个。

接下来，数据竞争的焦点有可能是机器数据。蚂蚁集团副总裁、金融大模型负责人王晓航表示：“有一种说法是，预计在2025年左右，全世界50%左右的数据会来自感知和传感等IoT（物联网）的数据，这部分数据能产生新的能力。”目前的数据来源主要是传统计算平台，如PC、服务器、手机和平板电脑等，嵌入式数据则来源于分布极其广泛的智能设备。

因此，大模型与硬件的结合将成为2024年重要的创新方向，也将创造智能硬件新物种。2023年11月，由山姆·奥特曼投资的智能穿戴设备公司推出一个基于大模型的智能穿戴设备AI Pin，有人称其“有可能取代智能手机”。据悉，AI Pin将于2024年3月批量上市。

基于此就很好理解，为什么谷歌Gemini大模型包括了Nano版本。“当手机上的传感器跟大模型整合，会产生非常多的应用场景，谷歌推出Gemini Nano版本，能够在手机等各种设备上部署，与安卓系统紧密地联系在一起。”周鸿祎表示。

大模型的目标不仅在于理解文本、图片、视频，它还必须真正理解物理世界。谷歌Deep Mind首席执行官德米斯·哈斯比斯表示：“谷歌Deep Mind已经在研究如何将Gemini与机器人技术结合，与世界进行物理交互，真正的多模态需要包括对于触摸和触觉的反馈。”

2024年元旦刚过，谷歌DeepMind又拿出了SARA-RT、RT-Trajectory、AutoRT三项机器人与大模型结合的成果，其中AutoRT是一个机器人数据收集系统，可以一次管理20个机器人。而在此前，OpenAI也投资了一家人形机器人公司1X。谷歌与OpenAI，谁都不愿错失任何一个有可能产生AI爆品的机会。

算力是关键底座 谁的主动权更大？

谷歌在发布Gemini大模型时，特别强调了自家的TPU（AI专用张量处理器）v4和v5e对大规模训练的给力支撑。为何？因为算力资源是AI的关键基础设施，是AI研究、AI模型训练、AI商业应用的利器。有信息透露，谷歌训练Gemini 1.0时所用的算力是OpenAI训练GPT-4的数倍。谷歌除了想凸显自身的算力优势之外，还要做算力服务的生意。当天，谷歌还发布了号称“迄今为止最强大、最高效且可扩展的TPU系统Cloud TPU v5p，用于开发更高层次的AI大模型。

谷歌这样做当然是希望“打脸”OpenAI——就在谷歌发布Gemini前，OpenAI宣布暂停ChatGPT Plus付费新用户的注册，此前还出现了ChatGPT周期性的宕机。直到2023年12月中旬，山姆·奥特曼才发文称：“我们重新启动了ChatGPT Plus付费订阅的注册，感谢您的耐心等待，同时我们找到了更多的GPU。”

山姆·奥特曼没有说明，其GPU究竟来自英伟达还是微软，但至少承认了一件事：OpenAI缺算力。尽管OpenAI的“好伙伴”微软已推出AI专用处理器，OpenAI也用上了，甚至还挖来了前谷歌TPU负责人主管OpenAI的硬件，但短期来看，其算力资源仍无法与谷歌相提并论。有报道称，即便OpenAI在两年内将GPU的总数增加四倍，但依然无法赶上谷歌的算力资源，目前OpenAI、Meta、CoreWeave、甲骨文、亚马逊的GPU总和，仍小于谷歌所拥有的TPU v5的数量。

算力资源的差距有可能在2024年改变谷歌与OpenAI模型的差距，拥有更多的算力资源意味着能进行更多的试验、更快地迭代模型。按照SemiAnalysis的预测，到2024年年底，谷歌模型训练的AI算力资源有可能是OpenAI的20倍。

业界也有人说，缺算力花钱买就好了。OpenAI正在启动新一轮融资，并不缺钱，自己的赚钱能力也蒸蒸日上。最新消息显示，OpenAI近期的年化收入突破16亿美元，主要收入来源包括ChatGPT Plus会员服务、API模型访问以及与微软的合作，预计到2024年年底，OpenAI的年化收入可达50亿美元。

更何况，刚刚上线的GPT Store给OpenAI带来了新的商业变现的想象空间。

不过，小冰公司CEO李笛认为：“GPT Store的建立不仅是为了商业模式，还是为了收集究竟什么样的GPT落地应用是有效的，是为了从开发者处获得想法和灵感。”这样看起来，OpenAI建立GPT Store的目的并不单纯，而开发者会不会把好的创意拿出来，也还是未知数。澜码科技创始人、CEO周健认为，当前GPT-4的能力对于发展GPT Store还存在瓶颈，GPT Store开发者是否能够做出用户愿意买单的应用，需要等到OpenAI发布GPT-5后才知道。

从现在的情况来看，芯片并不是想买就能买到的，OpenAI更不希望沦为GPU和云计算公司的“打工人”。目前，OpenAI与微软的关系依然牢固，微软也拿到了OpenAI无投票资格的董事会成员资格，但《福布斯》杂志预测，2024年OpenAI与微软有可能走向分手，“随着OpenAI大到蚕食微软客户”。OpenAI还在寻找英伟达之外的其他途径来解决眼下算力不足的问题，包括最近与AMD洽谈合作的可能性，同时也在自主研发芯片，评估潜在收购目标。

基于种种不确定性，OpenAI希望将更多的“算力主动权”掌握在自己手里。最近，OpenAI与人工智能芯片初创公司RainAI在2019年签署的一份意向书被曝光，OpenAI将购买该公司总价值5100万美元的NPU芯片，而在此前，山姆·奥特曼自己也投资了这家公司。这一举动引起广泛关注，因为RainAI的NPU芯片采用神经拟态技术，模仿人脑的结构和功能，被认为具有低成本、高能效的潜力，有望为OpenAI提供所需的硬件支持。

与Rain AI的这一交易，被视为OpenAI为确保其AI项目的芯片和硬件供应而进行的关键举措之一。但事实上这家芯片创业公司给出的出货时间至少要到2024年年底。算力的差距，可能在2024年成为OpenAI与谷歌AI竞争的关键变量。

酝酿底层变革 谁先实现AGI？

或许，谷歌与OpenAI的真正较量并不是GPT-5和Gemini Ultra，而是谁先实现AGI（通用人工智能）。

从目前来看，无论是GPT还是Gemini都基于Transformer架构。谷歌团队在2017年发表的论文*Attention Is All You Need*提出了Transformer架构，如今的主流大模型产品，大多基于该架构。

有人分析称，谷歌拥有原创AI架构、算力、数据、技术、人才等显著优势，但谷歌推出的Gemini也就领先GPT-4一点点，这或许意味着Transformer架构存在“天花板”。

Google DeepMind资深工程师卢一峰在与美国工程院院士张宏江对话中坦言，“现在的Transformer架构已经比当年有了很大的优化和改进”，从2016年到现在，整个业界在软件、硬件和数据方面对Transformer架构进行了许多组合优化，“已经将其推进到了一个局部最优状态”。“我认为我们可以继续改进它，还有很大的空间，但要显著改变它则有一定难度。这个难度在于这几个维度已经彼此交织在一起。”卢一峰说道。

需要有新的架构来“接力”这场AI加速跑。

《福布斯》杂志在最近出炉的《2024年十大AI趋势预测》中指出：“尽管我们不认为Transformer架构在2024年将消失，但确信将出现新一代更先进的AI架构替代方案，而且新的替代架构将在2024年得到真正的应用。”

该预测提及，斯坦福大学的Chris Ré实验室正在构建一种新的模型架构，这种架构可随序列长度以次二次方的方式扩展（而不是像Transformer那样以四次方的方式扩展），将使得人工智能模型的计算密度更低，并能更好地处理长序列。“候选”的替代方案还包括麻省理工学院开发的液态神经网络以及由Transformer联合发明人之一Llion Jones所创办公司推出的Sakana架构。据悉，目前Transformer架构的五位共同发明者均已离开谷歌，开启了各自的AI创新创业之旅，这些亲手孕育了Transformer的人有可能就是Transformer的“掘墓人”。

此外，随着大模型能力的不断演进，需要带来“跨越式变革”的未来计算，比如量子计算机或许是硬件的替代方案。创世伙伴资本主管合伙人周炜表示：“量子计算擅长的就是处理排列组合、并发的问题，当量子计算与大模型结合在一起就能够解决很多问题。”

“首先，人工智能领域的算法大部分属于并行计算范畴。而量子计算机擅长进行并行计算，因为它可以同时计算和存储0、1两种状态，无须像电子计算机那样消耗额外的计算资源，譬如串联多个计算单元，或将计算任务在时间上并列。计算任务越复杂，量子计算就越具备优势。其次，运行ChatGPT所需的硬件条件，同样也十分适合导入当前体积庞大的量子计算机，二者都需要安装在高度集成的计算中心，由一支专业化技术团队进行管理支撑。”中国现代国际关系研究院科技与网络安全研究所人工智能项目负责人谭笑间表示。

2022年，来自谷歌、微软、加州理工学院等机构的研究者从原理上证明了“量子优势”在预测可观测变量、量子主成分分析以及量子机器学习中确实存在。量子计算与人工智能两大前沿技术合流的趋势正在变得越来越明朗。在量子计算、量子机器学习方面，谷歌是先行者，如果量子计算机能够成为未来AI硬件的替代方案，谷歌无疑拥有比OpenAI更多的优势。

谷歌会比OpenAI更快实现AGI吗？抑或，最先实现AGI的既不是谷歌也不是OpenAI，而是其他公司？一切皆有可能。