谷歌Gemini让大模型走向多模态 算力需求将进一步暴涨

本报记者 李玉洋 上海报道

12月7日，谷歌发布新模型Gemini 1.0系列，该系列有三个尺寸版本（超大杯Gemini Ultra、大杯Gemini Pro和中杯Gemini Nano），被称作可以真正叫板GPT-4的模型。“这是我们迄今为止功能最强大、最通用的模型，在许多基准测试中都领先。”谷歌CEO桑达尔·皮查伊（Sundar Pichai）表示。

为显示Gemini强大的多模态能力，谷歌一段仅靠视觉和声音来驱动Gemini的视频演示，刷屏了全网并惊呆了很多人，一些评测指标上追平甚至超过OpenAI的多模态模型GPT-4V。然而，这段演示被指存在造假嫌疑，而谷歌方面的回应是“所有用户提示和输出都是真实的，只是为了简洁起见进行了缩短”。

尽管如此，研究机构Omdia人工智能首席分析师苏廉节对《中国经营报》记者表示：“谷歌是第一个把大模型和应用完美结合展示出来的公司，具有划时代的意义。像百度、谷歌、腾讯、Meta这种有大量消费者业务的企业，首要的目标应该是考虑怎么利用多模态将人机交互丰富化。”

“随着谷歌Gemini模型的发布，AI进入多模态时代。”这是外界对于谷歌新近发布大模型Gemini（中文名“双子座”）一个观察。苏廉节对这一看法表示认同。

多模态大模型已经是行业内公认的发展趋势之一。“这是很自然的趋势，文本处理完，就需要处理其他模态的能力，比如图像、声音。”AI算法专家、连续创业者黄颂表示，谷歌Gemini的推出对于多模态大模型的发展具有促进意义。中信证券研报指出，短期来看，Gemini将进一步激发市场对多模态模型的期待，对产业而言，多模态也将带动算力需求的提升。

原生多模态有多强？

根据谷歌的说法，Gemini Ultra在30项大模型能力测试中超过此前最强的大模型GPT-4。

目前业界比较公认的是，谷歌Gemini是真正能与GPT-4正面硬刚的模型，就算存在自吹自擂的成分。

根据谷歌的说法，Gemini Ultra在30项大模型能力测试中超过此前最强的大模型GPT-4，在检验大模型数学、历史、物理、法律等57个学科知识水平的MMLU（大规模多任务语言理解）测试中得分率达到90%，是第一个超过人类专家的模型。此外，在推理、数学和编码等几个评判大模型真正能力的测试中，Gemini Ultra几乎全面领先GPT-4。

需要指出的是，Gemini Ultra要到2024年才会向公众开放，它的真实效果还有待验证。皮查伊解释称，花更多时间是为了进行严格的安全测试，并挖掘它真正的功能。事实上，OpenAI在训练完GPT-4后，也花了半年时间做类似的事情。

Gemini Pro则会成为谷歌聊天机器人Bard背后的模型，替换原来的PaLM 2模型，一些开发者测试后发现，效果要比原来的好，但与GPT-4仍有不小的差距，大致相当于GPT-3.5的水平。

而Gemini Nano将搭载于谷歌手机Pixel 8 Pro，是一个定位在端侧的模型。据悉，Android开发者已能在Pixel 8 Pro上使用Gemini Nano开发应用，用户也可以用它总结录音纪要等。

撇开谷歌的自我宣传，科技圈大佬也对谷歌Gemini模型做出了较高评价。比如Meta的AI框架PyTorch联合创始人Soumich Chintala表示：“（Gemini）似乎在基准测试上可以硬刚GPT-4。谷歌拥有客户基础，无须担心模型采纳问题。而且谷歌将使用TPU进行推理，因此不必像OpenAI和微软那样支付给NVIDIA 70%的利润（直到它们的芯片准备好并投入生产）。”

在谈到“谷歌Gemini和GPT-4谁更强”时，360集团创始人、董事长周鸿祎表示：“谷歌的商业模式靠搜索和广告，做大模型等于左手打右手，所以没有全力做，这才给了OpenAI表现的机会。现在谷歌想明白了，与其被人打死不如主动转变。”

“从长期看，谷歌赶上GPT-4绰绰有余，毕竟是做搜索出身，有数据优势，有大量的知识积累和沉淀。搜索和大模型融合，能让大模型变得更实时，知识更全面更准确，搜索本身也会变得更智能。”周鸿祎称。

原生多模态，是谷歌Gemini的主要特色。与之形成对比的是，OpenAI的文字、图像和语音的模型分别是GPT-3.5/4、DALL-E和Whisper，直到三个月前低调发布的GPT-4V才能做多模态任务。

“谷歌Gemini模型的核心优势，在于其原生多模态的特性。”黄颂指出，多模态大模型已是大模型发展的明确趋势之一，Gemini的到来会刺激国内公司加速研发。

业内人士普遍认为，多模态是生成式AI下一步的重点方向，百花齐放的应用场景有待继续探索。苏廉节也表示：“目前的主流人工智能应用都是以文本和语音为主，包括现在最火的类ChatGPT应用也是用语言来交互，像百度、谷歌、腾讯、Meta这种有大量面向消费者业务的企业，首要的目标应该是要考虑怎么利用多模态将人机交互丰富化。”

东方证券研报认为，现阶段大语言模型的竞争已经非常激烈，从技术突破的角度来看，下一阶段的重点攻克方向必然是多模态技术。能真正处理和应用好多模态AI能力，才能真正打通物理世界和数字世界的障壁，用最基础的感知世界能力直接生成操作，实现与物理世界最自然的交互。

算力需求将进一步增长

相比于大语言模型，多模态大模型对算力的消耗呈指数级增长趋势。

“这是属于非常前沿的科技，就算谷歌Gemini展示出的能力也是经过精心调教的，没有那么顺其自然。”苏廉节指出，多模态的意义就在于，为AI应用带来了更多可能性，是通用人工智能（AGI）发展的关键。

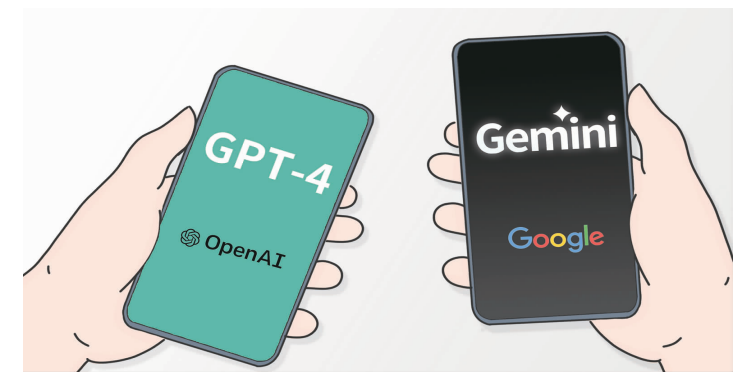
在Gemini技术文档和公开发言中，谷歌高管多次强调目前只是1.0版本，明年还会发布更先进的大模型。这显示出“没抢在ChatGPT前发布Bard”的谷歌，正在挽回自己在新一轮AI浪潮中的落后局面。

今年3月GPT-4发布之后，谷歌把两个人工智能部门DeepMind和Google Brain合并，并让DeepMind的创始人Demis Hassabis来领导新部门，主要使命就是牵头研发多模态大模型Gemini。

相比于大语言模型，多模态大模型对算力的消耗呈指数级增长趋势。国泰君安研报认为，当前多模态大模型仍在不断迭代，训练阶段的算力将保持增长。根据测算，GPT-4对于算力的需求在同等训练时长下相比GPT-3增长445倍。而根据谷歌内部消息，Gemini有万亿参数，训练所用的算力达到GPT-4的5倍。

只不过，有别于其他大模型对英伟达硬件及生态的依赖，Gemini训练所需的算力基于谷歌自研的TPU V4和V5e等硬件。在推出新模型的同时，谷歌宣布推出迄今为止功能最强大、最高效、可扩展性最强的TPU系统Cloud TPU V5p，将用于开发更高层次的AI大模型。

“对于人工智能基础设施而言，系统能力比微架构更加重要。”芯片研究机构Semianalys的首席分析师迪伦·帕特尔（Dylan Patel）表示，谷歌擅长把上千块AI芯片连接在一起，组成一个强大的算力平台。



谷歌Gemini和GPT-4谁能笑到最后？ 视觉中国/图

此外，谷歌还拥有围绕TPU的软硬件集成能力、研发出大模型基础技术的Transformer等基础实力，通过软硬件高度集成能力，做出一个能替代英伟达的方案，不是难事。谷歌云高管阿明·瓦赫达（Amin Vahdat）表示：“借助TPU V5p，可以让他们更划算地利用人工智能。”据悉，Salesforce、Lightrick等客户已经在使用谷歌云的TPU V5p超级计算机来训练大模型。

事实上，英伟达GPU作为2023年最紧俏的AI硬件，已经被各大科技巨头瓜分殆尽。根据Omdia近期发布的一份半导体研究报告，微软和Meta位居榜首，双双从英伟达购买了15万块H100 GPU。

从第三名开始，购买数量开始断崖式下跌。谷歌、亚马逊和甲骨文等公司各抢到了5万块GPU。其中，谷歌通过自研的张量处理单元弥补了一些芯片需求。国内科技巨头也是英伟达GPU的大客户，比如腾讯购买了5万块H800，百度和阿里巴巴分别购买了3万和2.5万块GPU。

Omdia的报告还显示，今年向英伟达采购H100（或H800）最多的12家客户里，有4家公司来自中国（分别是腾讯、百度、阿里和字节跳动）。

另一方面，英伟达也是动作频频。据Omdia统计，被称为“算力黄牛”的公司CoreWeave获得了4万块GPU，仅比谷歌少了1万块。而据华尔街见闻的报道，英伟达瞄准云服务领域后看上了CoreWeave，联手谷歌来扶持这家公司。

即使在H100紧缺的情况下，英伟达还是把大量的新卡分配给了CoreWeave，并直接参与投资。今年4月，在CoreWeave 4.21亿美元B轮融资中，英伟达成为了主要参与者，让CoreWeave估值升至20亿美元。

近期，英伟达CEO黄仁勋先后走访日本、新加坡、马来西亚和越南，跟当地政府和大企业谈合作、建AI基地。此前，黄仁勋还在今年9月和10月去了印度和中国台湾，合作对象分别是信实工业、塔塔和富士康。

国泰君安研报认为，训练成本持续高企，算力租赁商业模式具备可行性，短期持续看好算力以及算力租赁赛道。“AI算力有望超越普通算力服务市场，2023年格局分散。AI算力租赁市场规模有望超过普通算力服务市场，规模达到数千亿元。”东吴证券也指出。

市场广阔的算力租赁市场，吸引了众多玩家涌入。数据显示，近十年，我国算力基础设施相关企业呈逐年高速增长态势。2020年、2021年、2022年分别新增47.85万家、76.97万家、80.14万家，同比增长52.62%、60.85%、4.11%。天眼查数据显示，截至目前，今年我国算力基础设施相关企业注册量达86.66万家，已超去年全年注册量，其中今年前十个月新增80.27万家，同比增长20.48%。

中尺寸OLED市场竞争升温

本报记者 陈佳岚 广州报道

备受业内关注的TCL华星印刷OLED面板，计划于2024年下半年实现量产。

近日，TCL华星在其2023全球显示生态大会（DTC2023）上宣布，全球首款65英寸8K印刷OLED曲面显示屏、全球首款14英寸2.8K印刷Hybrid OLED笔记本电脑显示屏等多款印刷OLED新品发布。而据TCL科技CTO、TCL华星CTO闫晓林透露，公司计划于2024年下半年开始量产印刷OLED相关产品，并率先应用于IT和医疗显示领域。

麦吉洛咨询研究总监司马秋向《中国经营报》记者表示，TCL华星印刷OLED计划量产时间明确，意味着印刷OLED商业化进程正在提速。

值得留意的是，除TCL华星外，近期京东方、三星、维信诺都有对OLED中尺寸的布局动作，似乎也意味着平板、笔电等中尺寸已经成为诸多面板厂商共同发力的战场。而相较于韩国面板厂商以往在小尺寸、大尺寸OLED领域的相对先发优势，OLED中尺寸之战或许会成为新的战场。

印刷OLED量产在即

近年来，三星、LG等都在主攻OLED，退出LCD市场。随着各方的竞相入局，OLED领域的竞争也将愈演愈烈。

据了解，在OLED显示领域，蒸镀和喷墨打印是两大工艺路径。目前OLED面板量产的主流方法是真空蒸镀工艺，生产技术主要是三星、LGD等韩国面板厂商使用的“蒸镀式”：即在真空状态下，将红、绿、蓝等f发光材料汽化附着于基板上。当前，蒸镀+FMM（精细金属掩模版）工艺在中小尺寸OLED市场占据绝对优势地位，一些中国厂商跟随三星显示的技术路线，如今取得了一定的市场成绩。

印刷OLED工艺，指像喷墨打印一样，可定点喷印发光材料，发光材料使用率相较蒸镀工艺可提升两倍。印刷OLED技术可在低环境要求下进行，有效降低工厂能耗。显示效果上，印刷OLED技术在广色域、低功耗、高分辨率、透明显示、柔性显示方面更具优势。不过，印刷OLED的大规模商业化一直没有到来。OLEDindustry的一篇文章中提到，JOLED喷墨打印技术的缺点是，与真空设备中沉积的材料相比，屏幕看起来有污渍，并且使用寿命较短，提高产量并不容易。

而在印刷OLED工艺布局上，TCL华星动作较多。早在2020年6月，TCL华星投资300亿日元，获得了日本面板厂商JOLED约10%的股份，开始与其在印刷OLED技术上展开深度合作，随后几年TCL华星一直在进行印刷OLED领域的探索。

值得留意的是，JOLED因追求自己的生产方式，难以与韩国和中国的同行对手竞争。今年3月，JOLED申请破产保护，负债337亿日元。停止了生产和销售，关闭了Nomi工厂以及千叶县的另一家工厂。

12月7日，TCL华星不仅发布多款印刷OLED新品，还宣布预计明年下半年开始量产印刷OLED。据悉，12月初，全球首条印刷OLED显示屏生产线已在TCL华星武汉工厂开始搬入设备，后续将进行设备调试。

记者亦了解到，TCL华星在武汉的印刷OLED实验线正是来自其与JOLED合作中的产线。

TCL华星CEO赵军对记者表示，“TCL华星已于近日搬入上述印刷OLED产线设备，后续将依托目前引入的产线进行技术和产品开发，实现小规模量产。在确保技术和产品成熟度能够达到市场和客户的要求后，TCL华星会向IT、医疗专显等领域客户进行技术和产品推广，当客户认可和市场需求得到充分的论证后，我们后续才会考虑新的产线。”

“印刷OLED是TCL华星引领全球的下一个显示技术。”针对印刷OLED的量产进度，赵军对记者表示，“我们已进行印刷OLED设备搬入，这就意味着印刷OLED技术产品产业化将会以此为起点加快推进。”

OLED的工艺路线之争

此前在等离子电视和液晶电视的路线选择上，有些厂商因为赌对了液晶电视的方向赢得了市场。对于科技公司而言，押注技术和工艺路线机遇和风险并存。

近年来，AMOLED领域的蒸镀工艺和喷墨打印之争也浮出水面。

司马秋向记者表示：“等离子电视和液晶电视是技术路线的差异，蒸镀OLED和印刷OLED是工艺路线的差异。”

对于厂商而言，印刷OLED是否又是一个重要的工艺路线路口呢？

目前OLED主流工艺还是以蒸镀OLED为主，TCL华星也是国内主要的蒸镀OLED厂商之一。但TCL华星亦对印刷OLED寄予厚望。

闫晓林在DTC2023上表示：“TCL华星蒸镀OLED的目标是实现国内领域领先，而印刷OLED的加入将给TCL华星未来在OLED领域实现全球领先的目标奠定良好的基础。”

事实上，三星、LG、京东方也有研发印刷OLED的动作，京东方这几年也陆续发布过一些印刷OLED样机。但是，相对于其他厂商的多点布局，“TCL华星在印刷OLED技术方面取得的突破较为靠前，并投资了日本JOLED公司，双方在IJP-OLED（喷墨印刷OLED）领域开展深度技术合作。”长期关注显示产业的资深观察人士马聪表示。

司马秋对记者分析称，印刷OLED工艺材料利用率更高，理论上可以大幅降低成本，有希望适应电视市场的需求，所以TCL华星和京东方都在探索印刷OLED工艺路线，只是TCL华星在印刷OLED工艺上比京东方的做法更加激进。如果TCL华星能打通印刷OLED工艺路线，就有机会在电视市场和中尺寸市场与韩国厂商进行差异化竞争，形成优势。

“印刷OLED相比于蒸镀OLED更为先进。”在马聪看来，印刷OLED与蒸镀OLED不是选择的问题，印刷OLED一定是未来的趋势，只是当前印刷工艺成本高，制程上尚有技术难点需要攻克。但印刷OLED代表更先进的工艺，有能力的话，大家都会上马印刷OLED。

中尺寸OLED竞争加剧

当前OLED面板已逐渐成为智能手机面板的“标配”，并逐渐加速在平板电脑、笔记本电脑等中尺寸应用上的渗透率，其应用场景和市场边界正在不断拓展。

可以看到，面板厂商们已经将更多资源投入中尺寸OLED的产业化进程中。

赵军对记者介绍道：“TCL华星印刷OLED技术主要应用在中尺寸产品，包括显示器、笔记本电脑、平板、专业医疗等场景。我们会以中尺寸作为突破口，加快印刷OLED的产业化进程。”

京东方A（000725.SZ）近日亦公告，拟在四川省成都市高新西区投资建设京东方第8.6代AMOLED（OLED）生产线项目，预计项目总投资630亿元，主要生产笔记本电脑、平板电脑等高端触控显示屏，以顺应OLED屏从手机、穿戴领域向中尺寸显示领域渗透的市场需求，提升半导体显示的竞争力。

而今年三星旗下面板公司三星显示已经规划了高世代的OLED产线。据报道，三星显示宣布建设8.6代AMOLED产线，产品将用于平板电脑、笔记本电脑等，同样瞄准了IT面板应用。该产线投资4.1万亿韩元（约合人民币215亿元），计划于2026年量产。

今年5月，维信诺发布无金属掩膜版RGB自对位像素化技术——ViP技术（维信诺智能像素化技术），亦是一项适用于中大尺寸产品的OLED技术。

有面板行业人士向记者分析：“大家都是为了接下来OLED可能在中尺寸领域的渗透做准备。”

司马秋亦对记者分析，目前在苹果需求的牵引下，三星显示和京东方都宣布投资8.6代OLED产线，接下来LG显示也有可能公布8.6代OLED产线的投资计划，待这些产能释放，中尺寸OLED将在IT市场快速渗透，挤压LCD市场空间。

群智咨询（Sigmaintell）数据亦预测，2023年全球OLED平板面板渗透率约为1.4%。随着2024年苹果iPad Pro采用OLED面板，以及三星、华为、荣耀等品牌高端产品线布局OLED技术，将引发自上而下的技术迭代潮流，预计2024年全球OLED平板面板渗透率将迅速提升至5.7%。

目前，全球的OLED 6代线均采用蒸镀工艺，可以看到三星显示和京东方计划上马的8代OLED产线都是蒸镀工艺，但在部分厂商看来，接下来8代线如何走仍有待观察。

值得留意的是，目前TCL华星中尺寸印刷OLED实验线为5.5代，TCL华星并没有OLED 8代线面板的相关布局。“8代线即使继续采用蒸镀的话，也有很多技术问题需要解决，目前没有定论。”记者以投资者身份致电TCL科技投资者关系部，其相关人士表示，“OLED在中尺寸领域的确有渗透的趋势，不过，量产能力、未来的市场空间并没有非常明确的预期，公司对此也保持着相对谨慎的态度。”