生成式AI进入终端侧 产业链影响几何？

本报记者 陈佳岚 广州报道

生成式AI正跑步进入移动时代。

《中国经营报》记者注意到，自2023年7月以来，包括荣耀、小米、华为、OPPO、vivo在内的国内主要手机厂商，都已经将生成式AI带入手机终端；而且在实践层面上，国产手机厂商布局大模型已经产生了两种不同的路径：一种是荣耀、小米引入端侧AI大模型，另一种是华为、OPPO、vivo采用的“端云协同”部署方案。

端侧大模型的实现，离不开手机芯片厂商的支持。记者注意到，手机芯片厂商联发科（MediaTek）11月21日发布天玑8300 5G生成式AI移动芯片，采用台积电第二代4nm制程，在11月初，联发科还发布了天玑9300旗舰5G生成式AI移动芯片；而高通亦于10月底率先推出骁龙8Gen3，将生成式人工智能功能直接引入SoC系统级芯片组中。

随着SoC芯片巨头为在手机上跑大模型奠定硬件基础、手机厂商们纷纷支持大模型的应用，这也就意味着端侧AI应用或迎来爆发。

不过，多位行业人士提醒记者，端侧大模型对手机硬件来说仍是不小的负担，可能导致手机发烫、电池寿命缩短、续航能力下降等问题。

手机纷纷落地大模型

AI手机和AI PC正成为一个新趋势，下游端侧AI设备需求量有望增长。

在手机上使用大模型已经不是什么新鲜事，ChatGPT、文心一言、讯飞星火大模型等都推出了APP，甚至文心一言、讯飞星火大模型可以接入学习平板、学习机中，但这些应用都是依赖云端算力。而现在的趋势是，手机厂商正在努力使大模型直接在手机终端运行，不依赖于网络。端侧的模型应用成为不容忽视的重要场景。

终端侧AI，是指将AI的处理和数据存储、计算等任务放在终端设备（如手机、电脑等）上执行。终端侧AI的优势在于低延迟、低功耗、高隐私保护和低成本等方面。而端侧大模型，也意味着即使不联网，也可以在本地完成计算。

7月，荣耀发布了Magic V2折叠屏手机，声称是“全球首款原生集成AI大模型的国产手机”。10月，荣耀CEO赵明宣布，荣耀Magic6系列将搭载第三代骁龙8移动平台，支持70亿参数的端侧AI大模型。

8月，小米创始人、董事长兼CEO雷军透露，小米做大模型的思路是轻量化和本地部署，并表示目前手机端侧的大模型已经初步跑通。

在“端云协同”部署方案中，例如，华为的智慧助手“小艺”背后的大模型有端侧和云侧两种形态，可以根据不同设备和场景的需求进行处理。这种处理方式最大化地发挥了“端侧快”和“云侧强”的优势。

再比如，OPPO AndesGPT以“端云协同”为基础架构设计思路，推出从十亿至千亿以上多种不同参数规模的模型规格，能够基于“端云分工、端云互补、端云协作”等方式，灵活支撑多元化的应用场景。AndesGPT着重强调了“对话增强、个性专属、端云协同”三个层面的技术特性。

vivo大模型亦采用了云侧和端侧两种方式进行部署。在vivo X100系列手机上，vivo与联发科合作，实现全球首个跑通了130亿参数的大模型，端侧支持70亿参数的大模型手机。此外，vivo OriginOS 4在端侧部署了具有10亿参数的大模型，同时在云侧部署了660亿和1300亿参数的大模型。vivo的五个大模型通过云侧和端侧的融合，构成了一个大模型矩阵，可以快速响应各种问题和需求。

而在PC产品方面，联想在其Tech World创新科技大会上展示了端侧大模型方面的能力以及首款AI PC。

随着主流手机、PC厂商纷纷将落地端侧大模型，分析人士称，AI手机和AI PC正成为一个新趋势，下游端侧AI设备需求量有望增长。

端侧部署逻辑何在

未来移动端布局大模型更可能是端侧大模型和网络侧的大模型相结合。

在vivo X100系列手机上，vivo与联发科合作，实现了全球首个跑通了130亿大模型，端侧支持70亿大模型的手机。 视觉中国/图

终端厂商们纷纷都有端侧落地的方案，背后都是希望智能手机端侧AI模型支持本地问答、语音互动等功能。

据了解，与云端的通用大模型相比，端侧模型不需要将用户的数据上传到云端进行处理，因此可以更好地保护用户的隐私。端侧模型可以在本地进行计算，避免了网络延迟和带宽限制，提高了计算效率。

此外，成本也是手机厂商在布局大模型时候会选择端侧的原因。中信研报提到，ChatGPT测算生成一条信息的成本在1.3美分（约合人民币0.0929元）左右，是目前传统搜索引擎的3到4倍，单次搜索成本过于高昂。

vivo副总裁、OS产品副总裁周围在接受记者采访时就提到，如果需要调用云端大模型的话，目前最少也要0.012元，以vivo3亿用户来看，每天用十次，一年算下来也要100亿元左右的支出，成本非常高，但如果在手机端侧运行大模型，则不会产生推理成本。

可靠性方面，与云端互联的网络可能不稳定、甚至断线。决策在本地大幅降低了数据经过更长的通路产生错误的概率。终端侧AI处理能够在云服务器和网络连接拥堵时，提供媲美云端甚至更佳的性能。根据浙商证券研报中引用的“Cloud vs On-device AI? Maybe something in between!”的测算，如果所有的推理案例都在云服务器上进行，准确率是79.31%；如果49.88%的推理案例在移动端进行，其余在云端进行，仍可达到79.31%的云级准确率。

而芯片厂商们也敏锐识别到了手机厂商的诉求。比如，高通发布的骁龙8 gen3支持包括Meta Llama2、Chat GPT等在内的多模型生成式AI，其可处理的大模型参数超过100亿，推理速度达到了每秒20个token。

再比如联发科发布生成式AI移动芯片天玑9300，集成第七代AI处理器APU 790，结合内存硬件压缩技术NeuroPilot Compression来减少AI大模型对终端内存的占用，天玑9300支持在终端运行10亿、70亿、130亿参数的AI大模型，联发科的AI开发平台NeuroPilot支持Android、Meta Llama 2、百度文心一言大模型、百川智能百川大模型等主流AI大模型。

在PC端，英特尔预计于2023年12月14日发布面向下一代AI PC的酷睿Ultra处理器，且公布了“AI PC 加速计划”，宣布将为软件合作伙伴提供工程软件和资源，以期在2025年前为超过1亿台PC实现人工智能特性。

中银证券发布研报指出，随着高通和联发科相继发布支持边缘AI功能的旗舰SoC，未来手机有望部署本地大模型。预计随着生成式大模型进入移动时代，对应的边缘AI需求也有望迎来较快增长。

IDC则预测，到2026年中国市场近50%的终端设备处理器将带有AI引擎，AI落地空间广阔。AI手机是智能手机行业大势所趋。光大证券认为，AI手机等智能终端可能成为未来消费电子行业变革的方向。

不过，当端侧大模型渐成趋势之后，相比通用大模型，端侧大模型仍存在一些不足和缺点。

IDC中国高级分析师郭天翔对记者提醒道，目前端侧的算力要求比较高，功耗较大，更为重要的是端侧大模型的参数量级无法与通用大模型相比。

而未来移动端布局大模型更有可能是端侧大模型和网络侧的大模型相结合。

荣耀公司CEO赵明表示：“真正帮助我们更好管理自己的事情的时候还是要靠端侧AI能力，而未来一定是端侧大模型和网络侧的大模型相结合。混合式AI未来能够真正解决我们的问题。目前荣耀已经和很多云侧大模型供应商进行沟通，酝酿端云大模型互补的落地。”而郭天翔亦认为，未来移动端布局大模型的趋势也是端侧大模型和网络侧的大模型相结合，这样既可以保证有体量较大的参数量级，可以带来更好的使用体验，又能结合本地算力，在无法联网的情况下使用。

产业链准备好了？

由于终端设备的算力有限，一些芯片、手机厂商对落地在手机中的大模型都进行了压缩。

而随着生成式AI加速落地终端，也将对上游产业链带来影响。郭天翔对记者分析，这种情况将对高算力的芯片、大存储组合、大密度电池等提出更高要求。

麦格理（Macquarie）最新发布的报告就提到，支持终端侧AI大模型功能的智能手机将需要比以前更大容量的内存。目前主流的智能手机多是配备8GB RAM，具有AI自动生成图像功能的设备至少要配备12GB RAM，具有数字AI助手功能的设备需要大约20GB RAM。

实际上，记者留意到，由于终端设备的算力有限，一些芯片、手机厂商对落地在手机中的大模型都进行了压缩。比如，在手机端，高通已经将FP32模型量化压缩到INT4模型，实现64GB内存和计算能效提升。vivo通过模型压缩、异构计算等手段，让模型可以在手机CPU、GPU等硬件上高效执行。

TrendForce集邦咨询分析师黄郁璇对记者分析，目前AI功能主要是搭载在旗舰手机上，而安卓旗舰机一般都配置有12GB、16GB的容量，足以应付初阶的AI功能所需，不过的确会带动旗舰机向更高内存转进，之后随着AI在终端的应用更趋成熟，内存容量会进一步向上推升。

郭天翔亦对记者分析，未来肯定会推动手机存储向更大容量发展，可能目前落地使用最大24GB的RAM在未来就只是起步内存了。

联发科无线通信事业部AI技术高级经理庄世荣就曾表示，一款130亿参数的AI大模型，大概需要13GB 内存（INT8）才能运行，而智能手机本身运行安卓操作系统通常就要占用4GB的内存，如果还要相对流畅地运行其他常规APP任务并保活，还需要6GB的内存，即总的手机内存容量需求将达到23GB。

不过，生成式AI的发展，对推动存储芯片需求回暖、存储价格回升仍有限。黄郁璇表示，对于今明两年而言，存储需求受到AI应用鼓舞的成分还很低，目前主要仍是透过全面的终端需求好转，或是原厂控制产出的影响为主。

罢工落幕赶工开启 春节档前好莱坞大片恐仍“难有起色”

本报记者 张靖超 北京报道

近日，《中国经营报》记者注意到，美国演员工会—美国电视和广播艺人联合会（SAG—AFTRA）发布公告称，已与美国影视制片人联盟（AMPTP）就一份为期三年的新合约达成初步协议，并宣布于11月9日凌晨12时01分正式解除罢工令，这意味自今年5月份开始的、历时118天的好莱坞演员、编剧大罢工终于按下了终止键。

这场罢工中，“争取流媒体播放与重播分成”是谈判核心问题之一。美国演员工会公告显示，此次新合约将带来价值超10亿美元的工资与福利计划基金。工会主席弗兰·德莱斯切尔（Fran Drescher）表示这是一项“革命性的协议”，可有效增加最低薪酬与流媒体分成奖金，提高医疗和养老基金上限，并首次履行对AI技术使用的监管措施，为基层演员提供权益保护。

据《好莱坞报道》消息，有分析师估计好莱坞所处的加利福尼亚州在停工数月影响下，至少损失了60亿美元。但记者翻阅迪士尼、华纳兄弟探索、派拉蒙环球、奈飞等公司的财报发现，在今年第三季度，这四家公司的营收、净利润、自由现金流净额均有不同幅度的增加。同时，这四家公司对第四季度的业绩指引，也都呈现出环比增长的态势。

复工赶工

本次好莱坞演员大罢工始于编剧罢工期间的7月中旬，是美国演员工会史上持续时间最久的一场劳资拉锯战。工会要求提高最低薪资，共享流媒体收入分红，并保护群众演员不被生成式AI技术创造的“数字复制品”替代，但一度未能与制片人联盟缩减成本的初衷实现一致。此前，美国编剧工会大罢工已经结束，并在薪酬待遇、成员规模、保险缴纳、AI使用限制等方面与制片人联盟达成新协议。相较之下，演员工会的罢工协谈进展稍显缓慢，外媒消息显示，多位好莱坞一线演员曾出面协调谈判双方，并主动提出多缴会费以缩减双方诉求差距，期望尽快结束罢工对整个影视行业带来的经济影响。

据Variety消息称，多数人的最低工资标准将较以往提升7%，比WGA和美国导演工会的工资涨幅还高出2%。

如今，这份为期三年的演员工会新合约让两大工会接连罢工事件终于告一段落。影视制片人联盟发言人在声明中表示：“我们很感激整个行业都在热情地复工。”罢工结束后，演员们可以再次出席新片发布会和红毯首映式。值得关注的是，罢工期间，艾美奖颁奖典礼被推迟至2024年1月举行，这也意味着明年将迎来一个艾美奖、格莱美奖、演员工会奖和奥斯卡奖云集的繁忙颁奖季。

此外，由于此前两大工会接连罢工，各大影视公司纷纷推迟重要电影的上映时间，叫停影视节目的制作，许多电影、剧集、综艺节目都宣布将推迟至2024年重启制作。罢工结束后，由于2024年作品囤积量过多、制作进度紧凑且档期有限，一些原定于今年第四季度首播的新剧本剧集也都将改为在2024年至2025年播出，以获得足够的宣传预热时间。

特别是在电影制作部分，罢工产生的影响则引发了滚雪球般的连锁反应。据Deadline最新报道，漫威旗下的电影《死侍3》将从2024年5月推迟至7月上映，而原定于7月上映的《美国队长4：勇敢新世界》则被推迟至2025年2月14日上映。为了腾出上映档期，《刀锋战士》的上映日期从2025年2月移至当年11月7日。此外，罢工影响下漫威未能开启《雷霆特工队》的制作，因此该片的上映日期也将由2024年12月20日改为2025年7月25日。

迪士尼将《狮子王前传：木法沙》的续集定为2024年7月5日至12月20日上映，并暂时取消了原定于2025年7月25日和2025年11月7日上映的两部未命名电影。日前，迪士尼还宣布将真人版《白雪公主》电影的上映时间推迟一年，定为2025年3月21日上映，皮克斯动画工作室的动画电影《艾里奥》也将推迟至2025年6月13日与观众见面。此外，派拉蒙的《碟中谍8》档期推迟已近一年，预计于2025年5月23日上映。《寂静之地》、暂未定名的《海绵宝宝》电影则分别推迟至2024年6月28日和12月19日上映，但瑞安·雷诺兹主演的影片《如果》则将提前一周，于2024年5月17日在北美上映。

索尼将《毒液3》的档期从2024年7月14日推迟至11月8日，并开启了动画三部曲《蜘蛛侠：超越宇宙》的配音工作，该片原定于2024年3月上映，但目前还未确定该片推迟后的具体上映日期。不过，华纳表现出较强信心，表示由蒂姆·波顿执导的奇幻恐怖片《阴间大法师2》仍将在原定的2024年9月6日上映。

损失几何？

记者翻阅了迪士尼、奈飞、华纳兄弟探索、派拉蒙环球四家上市公司的最新财报，发现除派拉蒙环球外，奈飞、华纳兄弟探索的销售费用均有不同程度的下降，其中，华纳兄弟探索的营业总成本同比减少了5.65%，环比减少了约20%。迪士尼则由于业务架构的调整，费用方面暂时不具备可比性。

几大好莱坞和流媒体巨头在罢工期间降低费用支出和营业成本，虽然让他们的财务业绩表现亮眼，却对上游公司造成了损失。

据《好莱坞报道》消息，有分析师估计好莱坞所处的加利福尼亚州在停工数月影响下至少损失了60亿美元。受此影响，洛杉矶地区的总制片量缩减超四成，FilmLA报告也显示，今年三季度该地区“有剧本电视节目”的制作量较去年同期下降近99%，电影制作产量跌幅也达55%。经济的低迷影响了许多依赖娱乐业发展的企业和家庭，道具师、服装设计师等剧组相关职业成员也由于工作量骤减而逐渐陷入财务危机。

除美国本土外，好莱坞大罢工的余震也波及了其他国家的影视行业。《好莱坞报道》称，英国影视行业因大罢工一度陷入瘫痪，工作人员接连失业。

在中国内地，从业者最直观的感受就是好莱坞大片最几个月在中国内地的溃败。

国家电影专资办初步数据统计显示，2023年11月13日，中国电影年度票房突破500亿元，但在这500亿元中，进口影片票房约83亿元，占比仅16.6%，为最近十年来最低。猫眼专业版的数据显示，截止到11月23日，2023年度中国内地票房榜前二十的影片中，只有《速度与激情10》和2022年年底上映的《阿凡达：水之道》两部好莱坞影片，但票房均未超过10亿元。

“因为两大工会接连罢工长达数月，所以很多电影、剧集的拍摄制作都停滞了，所以上映、播出时间会推迟。目前还能在院线、流媒体上看到的，大部分是之前已经完成的、或者是整体流程进入尾声的作品。但这种存货还能支撑多久，会是一个疑问。尤其是在2024年春节档前，好莱坞大片恐怕难有起色。”从事电影制片的金妍（化名）说。

AMC院线首席执行官亚当·阿伦（Adam Aron）在发布今年第三季度财报后也表达了好莱坞双罢工对电影院线造成冲击的担忧，“编剧和演员的罢工，很可能会在明年给电影院线带来更大挑战。”好莱坞编剧和演员工会的罢工虽然已经结束，但对明年电影排映的影响已经形成——无论影院，还是流媒体平台，很多影片的上映日期都因为罢工被推迟。电影公司需要重新规划内容制作、后期、宣传、发行等一系列流程。

国内一家院线公司的人士则认为，在罢工期间，演员需履行工会要求，避免出席和参加所有影片的拍摄与宣传活动，例如《碟中谍7》《芭比》等一些质量较好、关注度较高的影片在内地的宣发声量都很低，而中国内地市场已是好莱坞在全球的第二大票仓，因此在疫情前，通常都会有主创人员来华宣传，或者在华举办首映礼。罢工结束后，这样的现象应该会减少，好莱坞明星来华路演、宣传，或许会在未来一段时间带动个别影片的票房。

如今，罢工结束虽令行业振奋，但数月的停工对影视行业经济发展带来的影响仍将持续，娱乐社区基金总负责人在一份声明中指出，大罢工将在未来数月继续对行业经济产生直接或间接影响，失业者还需寻找新工作或等待岗位复原，好莱坞离“行业的完全复苏”还有很长一段路要走。