苹果为何迟迟未推出大模型？

本报记者 张琪玮

近期，苹果公司悄悄放出一条不起眼的消息：其研发部门发布了一篇题为《闪存中的大型语言模型：在有限内存下高效的大型语言模型推理》的论文。在冗长的标题下，掩藏着大模型落地端侧的技术亮点。业内人士纷纷猜测：在“AIGC元年”里始终保持着沉默的苹果，是否终于要在大模型领域出手了？

浮出水面？

近年来，在新技术方面的“后发制人”已经成为苹果身上的独特标签：不是行业首发，却能凭借更加优越的技术能力、设计理念与生态体系重新定义产品甚至整个产业生态。这一特质早在iPhone、Macbook等系列产品上就有所显现，2023年6月发布的MR头显初次进入人们的视野中时，更被认为是苹果“后发制人”的高光时刻。

基于此，虽然苹果始终对百舸争流的大模型保持缄默，业内却认为苹果在暗自“憋大招”。多位业内专家对《中国电子报》记者表示，在大模型领域，苹果手握“门票”却迟迟没有“入场”，或许也是其“后发制人”风格的延续。

这也解释了为何苹果在大模型方面的任何风吹草动，都格外引人注目。2023年7月，首次传出苹果暗中测试大模型工具“Ajax”，并推出代号为“Apple GPT”的内部聊天机器人的消息。人们纷纷猜测，“Apple GPT”将为苹果的人工智能助手Siri带来革命性的升级。然而，这一猜测迄今尚未成为现实，“Apple GPT”最广为大众接受的解释落定为“工作人员在开发层面开始使用能够适配苹果系统的人工智能工具”。

2023年11月，在“AIGC元年”的尾声，苹果悄然推出一款名为Ferret的开源多模态大语言模型。据了解，这款大模型拥有70亿和130亿两个参数版本，从测试结果上看，其图像处理技术走在行业前列。然而，这款大模型由于仅面向研究机构开放，最终并没有走入大众视野。

2023年12月，随着上述论文的发布，“苹果大模型”再次被推上风口浪尖。略过晦涩的文字表述和复杂的技术原理，论文的结论是：大参数模型，终于有望在“内存有限”的端侧设备中落地了。

另辟蹊径

大语言模型要实现普及，落地智能手机是一条必经之路。当下，荣耀、vivo、OPPO、小米、华为等智能手机厂商纷纷推出“大模型手机”，大模型在手机终端的适配与落地已是大势所趋。

然而，训练参数大、体量庞大、难以部署在端侧离线使用，是大模型“走入用户掌心”的一大难题。记者了解到，在当下主流智能手机市场，16GB运存是较为广泛的终端配置，这样的运存处理手机日常运行绰绰有余，但要实现模型的加载与数据分析，就显得捉襟见肘。

为此，谷歌、Meta、微软等头部企业均选择了“让模型适应终端”的路线，纷纷推出了训练参数更少、体量更小的“小模型”。以微软为例，2023年12月，微软正式发布了参数规模仅有27亿的“小模型”Phi-2，并宣称该模型性能能够“吊打”体量在其25倍以上的大模型。

而面对这条“卷起来”的“小模型”之路，苹果却另辟蹊径，首次提出利用闪存技术创新来突破大模型端侧部署难点的概念。苹果发布的论文指出，利用其创新的闪存技术，可以让模型的运行规模达到iPhone可用内存的2倍。在该技术的加持之下，大模型的推理速度在Apple M1 Max CPU上提高了4~5倍，在GPU上提高了20~25倍。“这一突破对于在资源有限的环境中部署先进的大语言模型至关重要，极大地扩展了它们的适用性和可访问性。”研究人员写道。

具体而言，论文中提到了两种关键技术：一是“窗口化”技术，允许模型重复使用部分已处理的数据，从而减少频繁读取内存的需要，提高大模型运行效率；二是“行-列捆绑”技术，通过对数据进行更有效的分组，令大模型能够更快地从闪存中读取数据，从而加速AI理解和生成语言的能力。

从论文内容看，大模型在端侧的部署难题似乎可以迎刃而解。但也有业内专家指出，闪存技术仍有“漏洞”，离实际应用尚有距离。专家表示，闪存技术可用的核心假设是大模型所处理的相邻数据前后具有相似性，但论文中苹果并未对这一必要条件展开论证。闪存技术能否成为大模型端侧部署的“转折点”，还有待验证。

谨慎前行

苹果公司CEO库克曾说过：“苹果有计划在更多产品中加入AI，但要‘深思熟虑’”。

过去一年，苹果虽未对“AI”大书特书，但却处处可见AI的影子。在2023苹果全球开发者大会上，库克始终强调ML（机器学习）概念，称无论是硬件领域还是软件领域，苹果都早已为ML做足了准备。一方面，苹果最新推出的M2 Ultra芯片可以负担大规模ML的性能需求，在某些场景和需求下甚至可以部分替代独立图形处理器；另一方面，从系统到软件，苹果将ML的应用重点放在提升用户体验上。比如iOS17输入法方面的更新，其本质就是大语言模型的应用。

在技术层面的准备之外，苹果在内容方面也逐渐开始了动作。2023年12月，有消息传出，苹果正在就“价值至少5000万美元的多年期合作协议”展开讨论，并与康泰纳仕、NBC新闻和IAC等媒体接洽，获取他们过往新闻文章的使用授权，以作为大模型训练之用。相比同期微软、OpenAI被《纽约时报》因版权原因起诉的尴尬，苹果的这笔“版权投资”更显示了其在内容生成领域的计划性。

在硬件准备方面，香港海通国际证券分析师Jeff Pu发布报告称，2023年苹果可能已经建造了几百台AI服务器，而这个数量在2024年还将显著增加。他还指出，苹果计划最早于2024年年底在iPhone和iPad上采用生成式AI技术。这意味着，如果该计划得以实现，用户有望在2024年年第发售的下一代iPhone和iPad上亲身体验苹果大模型。

OpenAI GPT商店正式上线

本报讯 记者**张琪玮**报道：美国时间1月10日，OpenAI官宣GPT Store（GPT商店）正式上线，OpenAI联合创始人山姆·奥特曼将其称为“人工智能领域的苹果应用商店”。

OpenAI表示，到目前为止，社区成员已经构建了300万个GPTs（自定义GPT），并已批准其中一系列GPTs在GPT商店中下载。山姆·奥特曼宣称，GPT商店不需要用户有任何编程经验，只需用简单的自然语言输入希望GPT实现的功能，OpenAI的GPT构建工具GPT Builder就会根据要求定制出一个AI聊天机器人。在此基础上，用户可以通过多轮指令微调，轻松自定义AI机器人的属性和风格，并与他人分享。

数据显示，目前已有9.11万个GPTs涌现于公共网络，覆盖领域广泛，涵盖开发工具、生产力、绘图、语言学习、财经、客户支持、市场分析和医疗健康等多个领域。据奥特曼介绍，在GPT商店，GPTs有不同的分类以及排行榜，里面会依次排出安装量最高、评分最高的GPTs，相当于人工智能领域的苹果应用商店。

在营收模式方面，GPT商店也试图借鉴苹果应用商店建立起的成熟商业模式。奥特曼透露，未来，OpenAI将从自身营收中提取一部分用于激励最常用、最有用的GPT的开发者。此外，他还在采访中表示，GPT商店首先会根据使用量进行直接的收入分成，此后，还可能考虑推出订阅个人GPT的服务。

此外，记者了解到，OpenAI计划于2024年推出GPT-5。相较于GPT-4，GPT-5将具备更强大的语音、视频和推理能力。数据显示，2023年，OpenAI营收达到16亿美元，同比增长了56倍，公司估值高达1000亿美元。专家认为，随着GPT商店和GPT-5的推出，OpenAI将迎来收益与产业生态的新突破。

中国电信星辰语义大模型宣布开源

本报讯 记者**张琪玮**报道：近日，中国电信星辰语义大模型TeleChat-7B版本宣布开源，并开放1T高质量清洗数据集。此外，中国电信还透露，将在1月20日开源12B版本模型，拥抱更多开发者共建开源大模型生态。

星辰语义大模型是由中电信人工智能科技有限公司研发训练的大语言模型，采用1.5万亿Tokens中英文高质量语料进行训练。星辰语义大模型在业界首次提出缓解多轮幻觉的解决方案，通过关键信息注意力增强、知识图谱强化、多轮知识强化、知识溯源能力四大技术，将AI大模型的幻觉率降低了40%。

在模型开发方面，星辰语义大模型已与昇腾AI基础软硬件完成适配，并同步开源了适配后的代码。目前，该模型支持Atlas 300I pro推理卡，具备int8量化能力，精度与性能表现均与业界第一梯队持平。此外，它还支持Atlas训练服务器，用户可使用昇思MindSpore和PyTorch框架进行模型训练和推理，在两个框架下，该模型的精度与性能均有不俗的表现。