首页

推荐

亚运会

关注

朋友

我的

直播

放映厅

知识

热点

游戏

娱乐

二次元

音乐

美食

体育

时尚

业务合作

京ICP备16016397号-3

京公网安备 11000002002046号

广播电视节目制作经营许可证

京B2-20170846

网络文化许可证-京网文-（2022）0938-030号

互联网宗教信息服务许可证 京（2022）0000057

药品医疗器械网络信息服务备案 （京）网药械信息备（2023）第00318号

网络谣言曝光台

网上有害信息举报

违法和不良信息举报 400-140-2108

青少年守护专线 400-9922-556

算法推荐专项举报 sfjubao@bytedance.com

网络内容从业人员违法违规行为举报 feedback@douyin.com

广告投放

用户服务协议

隐私政策

账号找回

联系我们

加入我们

营业执照

友情链接

站点地图

下载抖音

搜索

投稿

- 发布视频
- 视频管理
- 作品数据
- 直播数据
- 创作者学习中心
- 创作者服务平台

登录

登录后即可观看喜欢、收藏的视频

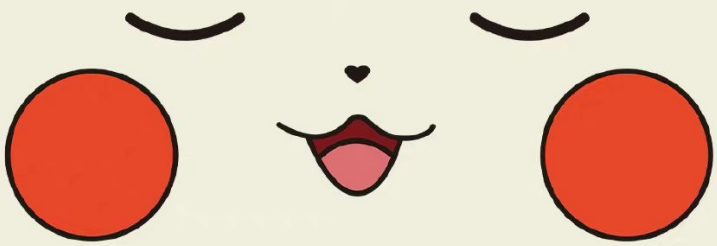
- 我的作品

- 我的喜欢

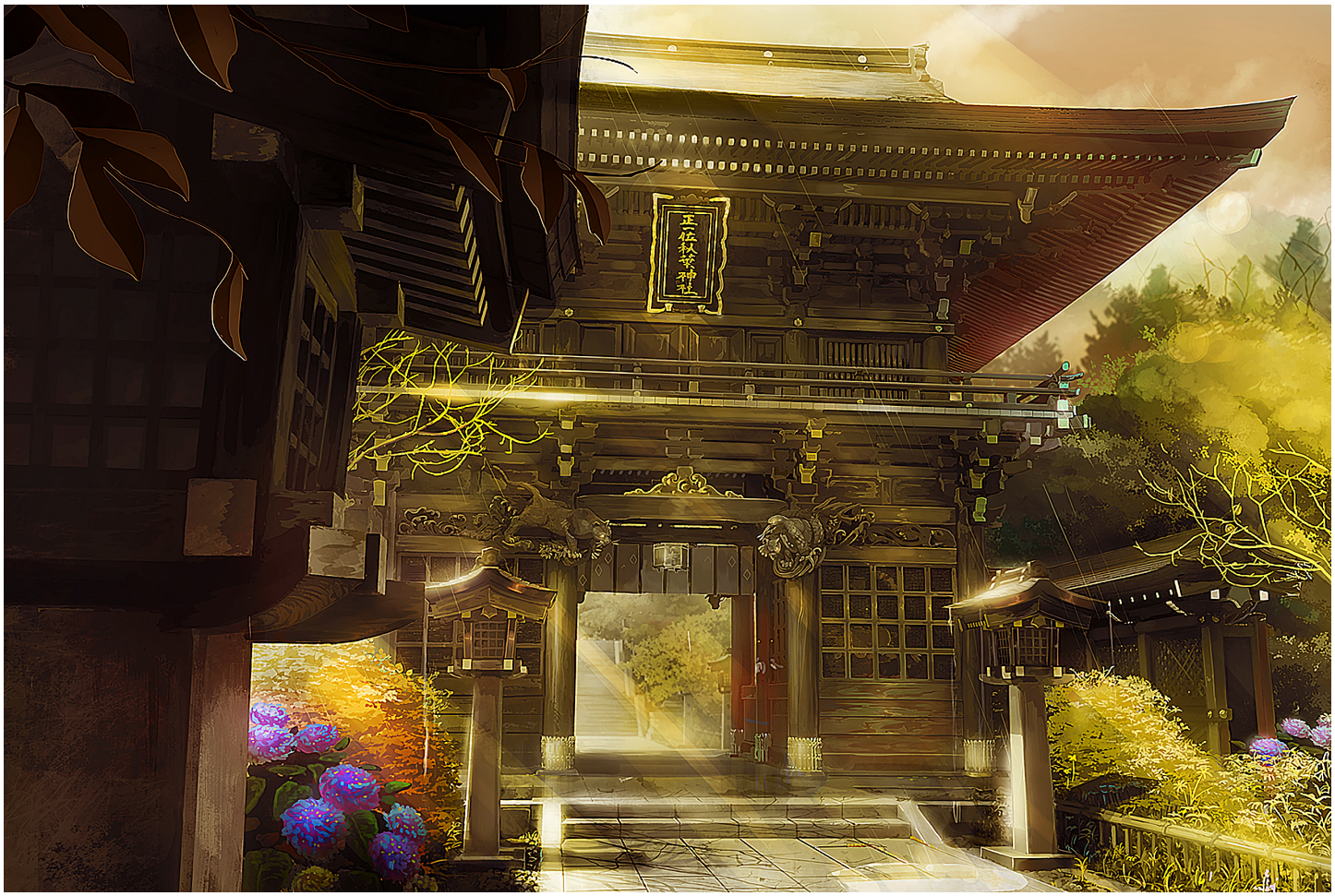
- 我的收藏

- 观看历史


正一位枚葉神社




Copyright @ 2018 IMIS



0

0

0

分享

音乐

愿你我皆安好（剪辑版）

贾晓龙

举报
发布时间： 20260403 18:53:08

全部评论

请先 登录　　后发表评论

暂无评论

[桦恺佑信息咨询](#)

[桦恺佑信息咨询工作室官方账号](#)

粉丝 57 获赞 1

关注

出品｜虎嗅科技组作者｜SnowyM编辑｜陈伊凡头图｜Multiverse Computing 官网"AI 原生 100"是虎嗅科技组推出针对 AI 原生创新栏目，这是本系列的第「17」篇文章。端侧模型和小模型这件事，在人工智能行业如今并不新鲜。去年，Meta、微软、苹果等就集中发布了一系列小模型，Llama-3、Phi-3、OpenELM 等。2019 年成立的 Multiverse Computing，试图用所谓 " 量子物理 " 方式给模型瘦身：它的核心技术 CompactifAI 能将大模型体积压缩 95%，却几乎不损失性能，让原本只能在数据中心运行的 AI，装进手机、汽车里。这也让这家公司获得了资本的青睐，截至今日，Multiverse Computing 已经完成了 5 轮融资。2024 年 3 月，这家公司完成了 2500 万欧元的 A 轮融资，一年多后 B 轮融资直接冲到 1.89 亿欧元，估值从 2024 年的 1.08 亿美元，涨到 5 亿美元，一跃成为西班牙最大的 AI 初创公司之一。两周多前，这家公司发布了两款 " 世界最小的模型 " ——鸡脑（chicken ’ s brain）和苍蝇脑（a fly ’ s brain）。" 苍蝇脑 " 是 Hugging Face 开源模型 SmolLM2-135 的压缩版本，原始参数是 1.35 亿，压缩之后只有 9400 万参数。" 鸡脑 " 则是 Llama3.18B 模型的压缩版本，可以直接在苹果电脑上运行，无需联网。这背后藏着太多值得拆解的问题：" 量子瘦身 " 技术究竟是噱头还是真功夫？当模型被压缩时，是否也会影响其性能？团队推出的 " 苍蝇脑 "" 小鸡脑 " 超小模型，又是如何突破硬件限制，甚至吸引苹果、三星等巨头洽谈合作？在 Meta、谷歌、微软纷纷下场做小模型，众多初创公司争抢 AI 效率赛道的当下，Multiverse 凭什么建立技术壁垒，成为西班牙估值最高的 AI 初创企业之一？虎嗅与量子计算领域的业内人士交流，试图理清这些问题。4 年 5 轮融资，估值一年涨 5 倍Multiverse Computing 并非一开始就进入模型赛道。2019 年团队成立之初，其聚焦量子计算软件，试图用量子技术解决金融领域的投资组合优化、风险管理等难题，这些在传统 IT 技术上难以被功克。凭借技术积累，Multiverse 很快被第三方数据分析与咨询机构 Gartner 评为量子计算领域的 "Cool Vender"。Gartner 的这份 Cool Vendor 的报告，主要涵盖科技创新领域，每个领域只有 4 家 -5 家公司能上榜，金融人士更是将这份榜单视为 " 投资宝典 "。借此，Multiverse 还获得了欧盟加速器 EIC 1250 万欧元的资金支持，成了欧洲资本最充足的量子初创公司之一。Multiverse 的团队中，40% 成员拥有博士学位，核心成员更是横跨金融、量子物理与科技创业三大领域 —— CEO 恩里克身兼数学、计算机、医学博士与 MBA，有 20 年银行业经验，曾任西班牙 Unnim 银行副 CEO；联合创始人罗曼是欧洲顶尖量子物理学家，专攻张量网络，拿过欧洲物理学会青年研究奖；CTO 塞缪尔则是量子计算与机器学习双料专家，熟悉创业与投资逻辑。转折点出现在 2023 年。生成式 AI 爆发后，大模型参数规模暴涨，算力成本飙升成了行业普遍痛点 —— OpenAI 每周在 ChatGPT 推理上的支出甚至超过训练成本。恩里克和团队敏锐发现，他们深耕多年的量子张量网络技术，恰好能破解这一困局：量子多体系统中的数学技巧，可用于大模型参数的高效压缩，且能最大程度保留性能。基于这一判断，团队火速组建 AI 压缩专项组，年底就推出了核心技术 CompactifAI，正式从 " 量子 + 金融

"转向"量子 + AI"。这次转向不仅让 Multiverse 踩中了"小模型"风口，更让它在 2024-2025 年迎来爆发，成为西班牙最大的 AI 初创企业之一。"量子瘦身"靠谱吗？Multiverse 的故事核心，是一套叫做 Compactif AI 的压缩技术。它不像行业常用的量化、蒸馏技术那样简单削减参数，按照 Multiverse 自己的介绍，这套技术是用量子物理张量网络方法，融合张量分解、矩阵低秩近似等复杂数学技巧，从模型底层重构参数逻辑。正如联合创始人奥鲁斯所说："我们的压缩技术并非计算机科学领域常见的套路，而是源自我们对量子物理的理解，更加微妙而精炼。"不过，虎嗅询问了量子计算领域的业内人士，Multiverse 所使用的这套数学方法虽然是量子中常用的，但其实只是一类数学方法，严格意义上和量子物理无关，因为张量网络问题最初就是物理学家从数学研究中借鉴到量子物理中的。所谓的张量网络方法，通俗比喻就是，你要拼一个一万平方米的拼图，拼完后为了存放它，需要找一个很大的房子。但如果你把拼图重新打碎，装到罐子中，把维度升高，从二维升高到 3 维，维度越多越方便压缩，再去掉重复的碎片，就可以装到一个小盒子里，并且保留几乎所有信息，需要的时候可以重新还原成大拼图。这种方法对大部分模型都适用，因为现在的模型，大多都是神经网络的变体，差别不大，Multiverse 的方法有很强的泛化性。这件事情的难点在于，要把现有的大语言模型基础算子 / 结构抽象出来，形成一套通用的压缩工作流，这样无论什么模型都可以复用。Compactif AI 通常能将型体积缩小 80-95% 而准确率只下降 2-3 个百分点。例如，原本需要数十亿参数的模型压缩后可能只有几亿参数，却在绝大多数基准测试中与原模型表现相当。目前 Multiverse 已发布多个压缩模型版本，例如 Llama 4 70B 模型的精简版 "Llama 4 Scout Slim"，以及 Llama 3 系列和 Mistral 小模型的精简版等。2025 年 8 月，公司发布了两款号称"史上最小且高性能"的模型，并以动物大脑体积命名——SuperFly（苍蝇脑）和 ChickBrain（小鸡脑）。SuperFly 基于 135M 参数的开源 SmolLM 模型压缩而成，仅含 9400 万参数，相当于一只苍蝇的大脑大小；ChickBrain 则由 Meta 的 Llama 3.1 系列 8B 模型压缩成 3.2B 参数（压缩率 60%），大小如小鸡大脑，却具备一定推理能力。ChickBrain（3B）的基准测试结果这件事的商业价值也很明显，CompactifAI 带来的直接好处是成本与效率优化。根据 Multiverse 公布的数据，其瘦身版模型推理速度是未压缩模型的 4-12 倍，对应推理成本降低 50-80%。在 AWS 云服务上，使用 CompactifAI 压缩后的模型可大大节省费用。例如，压缩过的 Llama 4 Scout Slim 在 AWS 上的调用费用约为每百万 tokens 0.10 美元，而原版约为 0.14 美元，也就是说，每处理百万 tokens 可以节省约 30% 费用。另外，CompactifAI 让此前只能在昂贵服务器上运行的 AI 模型进入了"平民设备"时代。Multiverse 声称其部分精简模型"小到可以在 PC、手机、汽车上运行"。目前，Multiverse 提供了 3 种商业服务模式：（1）通过 AWS API，将压缩后的模型与原始模型均可通过 API 访问；（2）购买私有部署许可，提供企业级授权，允许用户在自己的本地基础设施或云环境中部署 CompactifAI；（3）通过服务提供商交付，让 Multiver 为用户压缩模型，并交付指定的推理服务提供商。CompactifAI 的用户主要是广泛使用大模型的企业和开发者。大型互联网和软件企业的 AI 团队是首要客户，他们往往部署开源 LLM 在自己的应用中，如客服聊天机器人、代码自动补全、文本分析等，但也必然面临高昂的推理开销和延迟问题。CompactifAI 可以帮助他们大幅削减算力成本，甚至支持离线部署。CompactifAI 在降本增效和边缘部署方面功能突出。它可以将一个部署在 8 张 A100 GPU 上的 LLM 压缩到 1-2 张 GPU 即可运行，甚至压缩到能够在 CPU 上实时推理。这为客户节省的不仅是每小时数百美元的云 GPU 租用费，还有巨大的能耗开销。小模型和端侧模型——巨头云集的赛道Multiverse 的技术，很快吸引了全球硬件巨头的关注。据其透露，目前已与苹果、三星、Sony、HP 等洽谈合作，核心是将"苍蝇脑""小鸡脑"这类超小模型嵌入下一代终端设备——这恰好契合苹果的战略：2024 年 WWDC 大会上，苹果推出 "Apple Intelligence" 框架，明确表示不追通用巨无霸模型，优先做适配 iOS/macOS 的轻量化本地模型。不过，赛道竞争也在加剧。2024 年起，科技巨头纷纷下场小模型：Meta 发布 13 亿参数 LLaMA 微型模型，Google DeepMind 推出 2 亿 - 7 亿参数的 Gemma，微软 Phi 系列用 14 亿参数模型在数学、编码任务上超越 50 倍体积的大模型；初创公司中，Neural Magic、Deci 等也在争抢 AI 效率赛道，聚焦模型加速、自动选型等方向。AI 推理优化已经成为创投圈新的竞技场。初创公司阵营也不甘示弱。除了 Multiverse 外，Neural Magic、Deci、OctoML 都在下场大模型效率赛道；还有初创公司专注于模型路由、自动选型等，将不同

模型按成本和效果自动分配。这些公司切入点各异，但都瞄准了"提高 AI 性能 / 成本比"这个共同目标。虎嗅与量子计算领域人士交流，鉴于如今的大语言模型基本架构类似，Multiverse 的壁垒并不算太高，端侧模型和小模型不同，虽然都需要模型轻量化，但端侧模型除了需要轻量化，还需要配合不同设备的计算资源（内存、算力），以及能耗、发热等调节小模型，需要有特别设计，是一个工程化的问题。Multiverse
如果能够绑定一家硬件厂商，或许能够在端侧模型上建立自己的生态壁垒。另一方面，Multiverse 如今大部分还是围绕已有模型压缩，而不是自己训一个小模型，在效果上，可能不会达到惊艳的效果，而且极度依赖原有的模型能力。目前已经有一些专注小模型的初创公司除了压缩模型，还自己训练小模型，达到了不错的效果。Multiverse 可能在模型压缩上，通过自身团队积累的技术，能够实现较小的压缩损耗，但后续在端侧模型布局上的工程化问题，以及模型能力本身的技术壁垒，仍然有待观察。

WhatsApp网页版全新上线，极速扫码，随地随时在线畅享沟通

。

随着互联网技术的飞速发展，移动通讯工具在我们的日常生活中扮演着越来越重要的角色。WhatsApp作为全球最受欢迎的即时通讯软件之一，其便捷的聊天体验和强大的功能深受广大用户的喜爱。为了满足用户随时随地在线沟通的需求，WhatsApp网页版全新上线，极速扫码，让您随时随地畅享在线沟通的乐趣。 WhatsApp网页版是一款基于浏览器运行的即时通讯工具，用户只需在电脑端打开网页，输入账号信息，即可实现与手机端无缝连接。相较于手机端，WhatsApp网页版具有以下优势： 1. 极速扫码，快速登录 WhatsApp网页版采用了最新的扫码登录技术，用户只需在电脑端扫描手机端生成的二维码，即可快速登录。整个过程简单快捷，无需繁琐的账号密码输入，大大提高了用户体验。 2. 随地随时在线 无论是在办公室、咖啡厅还是家中，只要有电脑和互联网，用户即可随时随地登录WhatsApp网页版，与亲朋好友保持实时沟通。这对于经常需要使用电脑办公的用户来说，无疑是一个极大的便利。 3. 功能丰富，体验更佳 WhatsApp网页版继承了手机端的所有功能，包括发送文字、语音、视频、图片等，同时还支持文件传输、位置共享等实用功能。在电脑端使用WhatsApp网页版，用户可以更方便地管理聊天记录，查看群聊动态，体验更加丰富的沟通方式。 4. 安全可靠，隐私保护 WhatsApp一直致力于保护用户的隐私和安全。网页版同样采用了端到端加密技术，确保用户通讯内容的安全。同时，WhatsApp网页版还提供了多种隐私设置，用户可以根据自己的需求调整，保护个人隐私。 5. 跨平台使用，无缝切换 WhatsApp网页版支持Windows、Mac、Linux

等多种操作系统，用户可以根据自己的电脑环境选择合适的版本。此外，WhatsApp网页版还支持与手机端无缝切换，用户可以在电脑端和手机端之间自由切换，实现无缝沟通。 总之，WhatsApp网页版的上线，为广大用户提供了更加便捷、高效的在线沟通方式。极速扫码、随地随时在线，让沟通变得更加简单、轻松。无论是商务洽谈还是日常交流，WhatsApp网页版都能满足您的需求。 在未来的发展中，WhatsApp网页版将继续优化功能，提升用户体验。相信在不久的将来，WhatsApp网页版将成为广大用户不可或缺的在线沟通工具。让我们一起期待WhatsApp网页版带来的更多惊喜吧！

TA的作品

更多作品