

全球科技行业

本文内容由 Alex Wu (吴俊豪) 提供

中信证券财富管理(香港)

产品及投资方案部

电话: (852) 2237 9250 / 电邮: wminvestmentsolutions@citics.com.hk

小即是美

摘要

中信证券财富管理与中信里昂研究观点一致。根据中信里昂研究在 2026 年 8 月 21 日发布的题为《Small is beautiful》的报告，小语言模型获得的关注少于其前沿对应模型，但这种情况可能很快会改变。学术进展使其能力更强、具备多模态能力且硬件效率更高。Microsoft 的 140 亿参数 Phi-4 在部分基准测试上可胜过 GPT-4o，而阿里巴巴的 Qwen3-VL 8B 为该模型配备了图像/视频处理能力。更好的软件也可以释放智能手机内部已配备的 NPU 潜力。AI 需求已准备好从云服务器扩展到消费设备、工业设备、车辆和机器人。

更好的数据驱动更好的、更小的模型：Microsoft Phi-4

- 小语言模型历来遵循规模定律：更大的模型通常表现更好。蒸馏（Distillation）产生了更便宜的“学生”模型，但它们通常仍明显落后于其更大的“老师”模型。
- Microsoft 的 Phi-4 对这一上限提出了挑战。Microsoft 并非主要依赖原始网络数据，而是使用经过精心筛选和专门构建的合成数据来训练这个 140 亿参数模型。更好的数据带来了一个更好的、尽管更小的模型。
- Phi-4 在部分数学和科学基准测试中优于 GPT-4o，证明模型规模不再是获得更强智能的唯一途径。

更小的模型也能处理视频和图像：Qwen3-VL

- 早期小语言模型面临两大限制。它们的上下文窗口较短，而且大多数主要是为文本而非图像或视频而设计。
- Qwen3-VL 解决了这两个问题。其较小的 20 亿和 40 亿参数版本可在 256,000 token 的上下文窗口内原生支持文本、图像和视频。
- 阿里巴巴对该模型进行了渐进式训练。该模型先掌握较短的材料，然后才被要求处理更长得多的输入。
- Qwen3-VL 还改变了其在文本、图像和视频之间跟踪时间和空间位置的方式，从而在利用各种信号时提供了更高的一致性。
- Qwen3-VL 8B 在多模态性能方面于某些基准测试上几乎可媲美 Gemini 2.5 Pro 和 GPT-5。

最大化利用智能手机中的神经处理单元：llm.npu

- 智能手机已经包含 CPU、GPU 和神经处理单元。NPU 是为大规模、重复性的 AI 计算而设计的。问题在于兼容性。
- 移动端 NPU 偏好固定大小的输入和整数计算，而语言模型处理长度不一的提示并需要非整数运算。
- llm.npu 系统将提示拆分成若干部分，将适合的整数计算发送至 NPU，并将不规则或对精度敏感的工作分配给 CPU 和 GPU。
- llm.npu 将预填充速度提升了 22.4 倍，并将能耗降低了 30.7 倍。

智能体需要专家，而不是每项任务都需要一个天才：NVIDIA 对小语言模型的论证

- NVIDIA Research 认为，小语言模型特别适合智能体 AI。
- 对用户而言，一个智能体可能看起来是通用型的，但其底层工作流通常由许多狭窄且重复的步骤组成。针对特定任务训练的较小模型，可能具有足够准确性、显著更快且运营成本更低得多。
- 小语言模型更容易通过微调来专门化以适应某项任务，这在智能体环境中很有用。许多工作流要求对输出进行严格约束，而一个经训练可执行单一工作的专用小语言模型可能提高系统的可靠性。
- 部署一个 70 亿参数的小语言模型成本可能比使用一个大得多的模型低 10-30 倍。当一个智能体重复调用模型时，这种节省会进一步累积。
- 小语言模型不会取代前沿模型。然而，它们可以处理底层的大部分工作，使智能体更便宜、更快速且更易于部署。

免责声明

本文由中信证券财富管理（香港）（“**中信证券财富管理（香港）**”）撰写。中信证券财富管理（香港）是中信证券经纪（香港）有限公司（“**中信证券经纪（香港）**”）的财富管理品牌。本文为一般性资讯，仅供参考之用。本文具有教学性质，但不被视为对任何司法管辖区的任何特定投资产品、策略、计划特征或其他目的的建议或推荐，中信证券财富管理（香港）或其任何子公司也不承诺会参与任何此处提及的交易。任何示例都是通用的、假设的并且仅用于说明目的。本文尚未经香港证券及期货事务监察委员会审核。

本文是由投资／产品专员而非分析师撰写的文章汇编。它不构成研究报告，也不应被解释为研究报告，也不旨在提供专业、投资或任何其他类型的建议或推荐。本文件所载观点为投资／产品专员个人观点，或投资／产品专员对公司观点的理解，可能与中信证券财富管理（香港）的官方观点和利益（包括投资研究部门的观点）存在差异，也可能不存在差异。

适合性、风险及适当性

本文不含足够资讯以支持投资决策，并且不得依赖本文来评估投资任何证券或产品的优劣。本文仅提供一般资讯，未针对任何接收者量身定制，亦未考虑任何接收者的投资目标、投资组合持仓、策略、财务状况或需求。本文具有教育性质，并非中信证券经纪（香港）或其附属公司参与任何交易的建议、推荐或承诺。任何范例均为通用、假设及仅供参考之用。

读者应自行独立评估法律、监管、税务、信用及会计方面的影响，并与其财务专业人员共同决定本文提及的任何投资是否适合其个人目标。投资者应确保在进行任何投资前获取所有相关资讯。本文中所涉及的任何预测、数据、观点、技术或策略仅供参考，基于某些假设及当前市场状况，可能会在未通知的情况下发生变更。

信息来源被认为是可靠的，但不保证其准确性，也不对任何错误或遗漏承担责任。请注意，投资涉及风险，投资的价值和收益可能会根据市场情况和税收协议而波动，投资者可能无法收回投资的全部金额。过去的表现和收益率都不是当前和未来结果的可靠指标。

本文不应被解释为在任何司法管辖区向任何不合法接受此类邀请或招揽的人发出购买或出售任何证券的邀请或招揽。此处包含的信息均未在任何司法管辖区获得批准。

责任限制

中信证券财富管理（香港）对于您或任何第三方在使用此材料或其内容时遭受的任何损失或损害（包括但不限于直接损失、间接损失、或由此导致的间接损失、利润损失或机会损失）不承担任何责任。此材料是根据您的请求提供给您的，并适用于熟悉相关市场且愿意并能够承担投资相关风险的个人。在您充分了解所有涉及风险并独立确定该投资适合您之前，请勿进行任何交易。请注意，我们并非您指定的投资顾问。

法律主体声明

香港：本资料由中信证券经纪（香港）有限公司（中央编号 AAE879）在香港发布及分发，该公司受香港证券及期货事务监察委员会监管。

附属披露

本文可能包含先前发布的信息，包括由 CLSA Limited（中央编号 AAB893）发布的信息，该公司受香港证券及期货事务监察委员会监管。中信证券经纪（香港）与 CLSA Limited 是关联实体，均属于中信证券国际有限公司（“**CSI**”）的共同所有或控制之下。

中信证券经纪（香港）与 CSI 及其子公司（“**CSI 集团**”）内的其他实体之间可能存在潜在的利益冲突，这些冲突可能涉及多个领域，包括投资产品、企业咨询服务、证券交易以及提供给本文提及公司（们）的其他金融服务。此外，CSI 集团可能在本文讨论的公司（们）中持有利益或头寸。

中信证券经纪（香港）遵循一项整个集团的政策（可能会不时更新），该政策阐明了 CSI 集团实体如何管理或消除任何实际或潜在的利益冲突。此政策旨在确保 CSI 集团发布的任何研究报告或评论保持公正。

针对不同司法管辖区的声明

香港：根据任何时间下所适用的法律和法规，中信证券财富管理（香港）、其各自的关联公司或与中信证券财富管理（香港）有关联的公司或个人可能在发布本资料前使用了此处包含的信息，并且可能不时购买、出售或在所提及的任何证券或相关证券中拥有重大利益，或在目前或将来可能产生或曾经有业务或财务关系，或可能或已经向本文提及的实体、其顾问和／或任何其他关联方提供投资银行、资本市场和／或其他服务。因此，投资者应注意中信证券财富管理（香港）和／或其各自的关联公司或公司、或此等个人可能存在一项或多项利益冲突。

未经中信证券经纪（香港）有限公司事先书面授权，任何人不得以任何目的复制、分发或销售本文。

版权所有 2026。中信证券经纪（香港）有限公司。保留所有权利。