MT-LNN: 受微管启发的液态神经网络

打破内存墙的下一代长文本架构

商业计划书 / *Investor Pitch Deck*

Awareness

2026 年 5 月

Paper: https://huggingface.co/EverestAn/MT-LNN
GitHub: https://github.com/everest-an/01

AI 八大核心流派

- **1.** 符号主义：靠逻辑运算、视觉推理，最贴近人类思维。
- **2.** 贝叶斯主义：用先验/后验概率推理，核心是贝叶斯网络。
- **3.** 类比主义：靠事物类比做分类，代表 SVM、近邻算法。
- **4.** 演化主义：模拟自然演化规律设计 AI。
- **5.** 联结主义（当前主流）：神经网络技术，包含大模型、可解释 AI（XAI）、液态神经网络 **(LNN)**。
- **6.** 强化学习：模拟生物“奖励-惩罚”机制训练 AI。
- **7.** 多代理系统：模拟多智能体之间的交互与决策。
- **8.** 连续学习：AI 无需重新训练，就能持续学习新内容。

AI 行业的内存危机

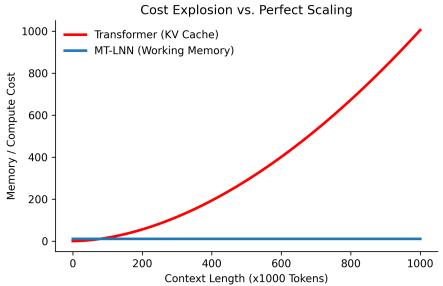
- 超长上下文的代价：大模型竞相追逐 200K+ Tokens 的上下文长度，但背后的计算与内存成本正呈指数级爆炸。
- **OOM** (显存溢出) 限制：极大多数本地化和私有化部署的瓶颈在于物理显存，而不是逻辑推理能力。
- 算力困局：当前云服务商由于底层架构限制，面对高并发的超长请求时，只能依靠堆叠昂贵的算力集群。

成本爆发的元凶：KV 缓存

- 强迫式的记录引擎： Transformer 架构如同强制录像机，预测下一个词时，必须把过去所有词的特征放入显存中（即 KV Cache）。
- 复杂度灾难：随着文本增长，内存复杂度为 $O(N)$，计算复杂度飙升至 $O(N^2)$。
- 实例：仅仅 10 万字的上下文，就会瞬间塞满一台昂贵的 A100 GPU；更不用提在只有十几 GB 内存的手机或端侧设备部署，这在物理上是不可能的。

成本爆炸 vs. 完美扩展机制

- Transformer 架构依赖不断膨胀的 KV 缓存，在处理超长上下文时导致服务成本飙升以及巨大的内存压力。
- MT-LNN 采用紧凑的循环隐状态与微管神经启发的选择性动态机制，完美控制了运算时的内存增长。
- 商业化意义：大幅降低了私有化/本地化部署和端侧及边缘设备（Edge Devices）场景的落地阻力与门槛。

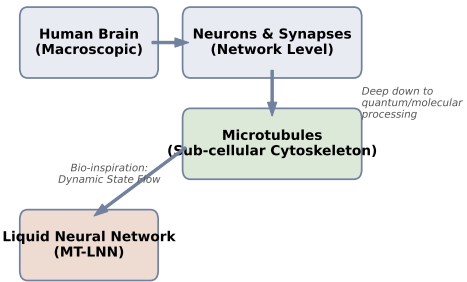


生物学启示：人脑机制

- 不记录每个像素：人类大脑绝对不会，也不可能记住十年来每个画面的精确像素和所有细节。
- 工作记忆与选择性遗忘：我们的认知高度依赖于“工作记忆”（滞存核心线索）与“选择性遗忘”（抛弃无关噪音）。
- **AI** 的盲区：现有的标准 LLM 仅仅是庞大的纯数学矩阵乘法，缺乏此类动态时间流逝机制，也未体现全局工作区（GWT）等意识级别的信息整合路径。

类脑 AI 探索：什么是微管 (Microtubules)？

- 神经计算的底层基座：大脑智能不仅源于神经元连接，神经元内部更存在着高度动态的细胞骨架网络——微管 **(Microtubules)**。
- 高频动态的信息处理：前沿生物物理学理论指出，微管能够以极高频率动态过滤、处理并压缩环境信息。
- 从生物微管到 **LNN**：我们受此启发，摒弃了 Transformer 中强制静态缓存所有历史数据的机制，转而利用类似生物微管动态微分方程，让信息在系统内随时间流动、选择性留存与遗忘。



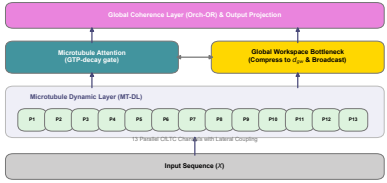
什么是液态神经网络？

- 动态流动的状态：液态神经网络 (Liquid Neural Network) 是一种内部状态随时间动态流动的结构，区别于传统 Transformer 的纯静态权重。
- 压缩与抽象：它的核心在于：只提取核心概要，遗忘冗余废话，把千言万语压缩为一个随时间演变的高维隐状态。
- 类脑基座：真正能够从物理与动态层面上，模仿大脑细胞微管结构处理信息的过程。

我们的方案：MT-LNN 架构登场

微管液态神经网络系统：

- 取代自注意力机制：抛弃标准 Transformer 中耗费显存的自注意力静态 FFN 层，替换为 13 个平行的连续时间液态网络。
- 原生支持意识计算：是世界首个突破“麻醉测试”(Anesthesia Benchmark) 的 AI 架构，不仅生成语言，还支持追踪内在认识坍缩的指标 ($\hat{\Phi}$)。



超越“文字接龙”：真正理解真实物理世界规律

- 概率拼接的局限性：传统大模型（Transformer）本质上基于词汇的高概率路径预测来进行文字接龙。例如它可以依概率拼接出 “苹果掉在地上”，但模型本身并没有建立关于重力或物体坠落过程的深层因果理解。
- 原生因果建模能力： **MT-LNN** 借助连续时间微分方程实现“流体动态感知”。模型并非简单记忆词频，而是真正在内部隐状态中推演了“苹果”随“时间”下落的物理规律，构建真实的因果世界模型。

MT-LNN 核心动态定位概念

深度解析：并行扫描与量子耦合

- 极致的线性推理：依托先进的算子技术与硬件级优化，使模型在不损失精度的前提下实现线性推理，新 Token 处理算力开销恒定为 $O(1)$。
- 量子启发耦合：创新性引入类似量子物理的隐状态多头路由交互，13 组平行网络以流结构实现深度特征的有效融合与提取，提升数据保真度。
- 赋能边缘计算场景：模型依靠常数级容量向量运行，使得大模型本地化运行于 AR 眼镜、手机等极低内存终端设备成为可能。

战略权衡：深度推理优于广度记忆

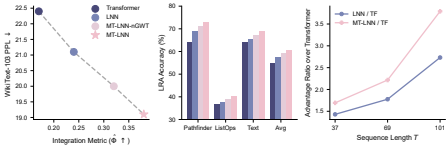
- 聚焦核心推理场景：盲目堆叠参数以存储海量“百科常识”对模型运行效率拖累极大。MT-LNN 选择走垂直且高度理性的路线。
- 专注长文逻辑与长效记忆：我们将有限的参数规模全面投入在严密的逻辑推演机制与跨度结构记忆上，而非单纯的百科背诵。
- 企业级应用切入：针对完全脱机的端侧或私有云场景（规避极密泄露风险），极速高精度处理数万字的财报审计或是庞大的代码审查工作。

核心验证 1：长文本抗噪与选择性记忆

超长噪声插入测试 ($T = 229$)：

- 传统机制表现受限：随着文本长度和无关噪声的增加，传统小型序列模型的关键信息提取准确率锐减，极端情况下甚至下滑至 2% (约等于随机瞎猜)。
- **MT-LNN** 的高鲁棒性：凭借动态选择性过滤与隐状态遗忘机制，在极高噪音的干扰下，仍能将精确匹配率稳定保持在 **96.5%**。

在同等极小参数量 **(200K)** 评测下，性能大幅领先于主流竞品架构。



核心验证 2：128K 长上下文“大海捞针”

- 攻克**”Lost in the middle”** 痛点：传统 Transformer 等架构在处理长文本大纲时，极易遗忘处于中间位置的关键信息事实。
- 全文本一致的提取精度：无论金线索信息是被随机掩埋在文首、文中或文末，MT-LNN 的动态扫描都能实现无遗漏的精确捞取，长范围检索准确度平稳保持在 99%+。

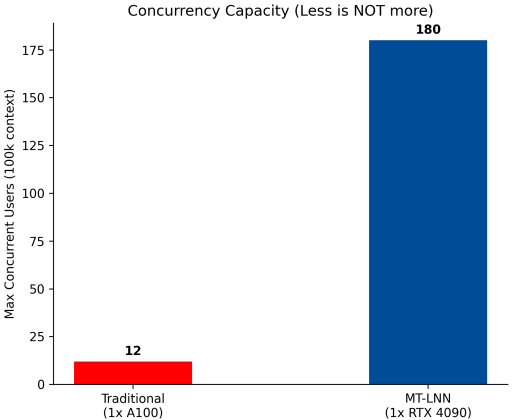
市场与定位：独特的差异化竞争优势

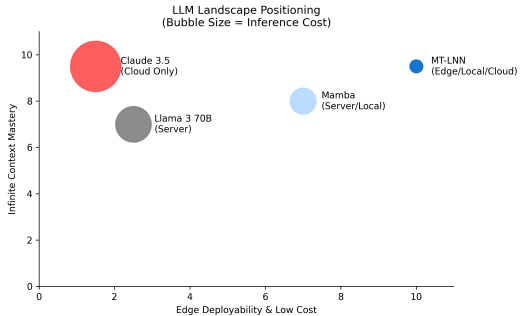
- 错位竞争战略：头部闭源模型能力突出但使用成本高亢且伴随严重的商业隐私阻碍；而现有的前端轻量开源模型在处理长上下文逻辑时显得力不从心。
- 占据效能制高点：我们同时在“百 K 级长文本理解力”和“指数级降低的部署成本”两个维度发力，开辟出一条可落地的轻量级高性能架构蓝海。

核心应用场景与竞争力分析

- 数据强敏感行业（如金融、法律、医疗）：在此类无法接入云端大模型 API 的闭源局域网场景下。MT-LNN 可以直接部署于办公笔记本等普通终端设备上，实现安全、合规且成本极低的本地长文本分析。
- 对比现有的高效线性架构（如 **Mamba / RWKV** 等）：此类架构面临着拉长文本即产生“记忆退化衰减”的困境。MT-LNN 借助门控以及动态路由调度专利，从底层机制上克服了长期维持多向特征的衰减问题。

市场定位与高并发竞争象限





高并发与 TCO (整体拥有成本) 优势

- 规模性降低推理成本：由于摒弃了呈二次方膨胀的 KV 缓存矩阵，千字大段输入与短语指令在 MT-LNN 中的运行内存消耗基本一致。预计云端模型的大规模服务算力成本可降低至现有方案的 1/10。
- 处理极端并发场景表现卓越：若同时应对 100 用户各 10 万字级别的极限请求推送：普通 A100 模型集群极可能因显存激增而导致 OOM（宕机崩溃）；而我们可仅用单台消费级加速显卡从容稳定排队并行反馈。

蓝海下沉市场 1：IoT 与边缘嵌入设备

- 解决物理芯片发热与功耗瓶颈：模型不仅体积小，而且对内存读写的极速优化使得我们能将其无缝嵌入到功耗受限的可穿戴 AR 交互外设和车载中枢内。
- 全天候陪伴式流信息处理：现存主流 AI 无法承载几十小时全时段开启音视频输入捕捉，而 MT-LNN 随时间递进的连续动态吸收机制天然适用这类连绵不断的外界数据流场景。

商业落脚点 2：B 端垂直机构与研发刚需

- **B** 端目标客户痛点吻合：顶尖律所面临海量千页卷宗的溯源；风投/审计财团对无尽财报报表的数据对齐；或大型科技厂商进行的底层百万行老旧祖传代码库盘查检越。
- 三要素完美匹配： MT-LNN 精准满足这部分客户迫切需求的三个必要条件：1. 超长篇幅大批量吞吐能力。2. 绝对的私有合规和不联网隔离。3. 基于极低硬件投入的一次性低成本本地部署。

当前进展：成熟的长文本 RAG Demo 演示

- 工程化验证进度稳健：目前技术不仅具备深厚的理论与论文基础验证，且已实现完整的模型级封装工程化。我们结合了最新组件搭建了流畅运作的大规模文档推理前端平台。
- 优秀的交付适应性： Demo 已完成标准化装箱打包测试。能即刻满足本地脱网部署演示要求，或快速封装成微服务流交付并入传统企业的已有网络中，打通产品化最后一公里。

核心技术路线与 18 个月展望

1. 一阶段验证期 (当前)：在极小参数级模型（如挂载架构至 1.5B 参数量级规模）下，初步获得行业内长文本测评榜单领先与学术验证，并吸引第一梯队种子与社区开发者讨论。

2. 二阶段架构规模化 (随后的 **6-12** 个月内)：着手引入充足起步的高算力集群资源，开发训练并推出完全原生结构的大规格 7B (70 亿) 级底层大模型组件架构。与初期 B 端合作伙伴深度联合开发落地闭环体系与行业最佳实践案例。

3. 三阶段迈向下一代泛化多模态基座 (未来 **18** 个月以上)：跨越多模态，使用超千亿宏大数据集参与大规模通用联合训练。并用该极致常数时间效能最终推动更高度的泛化 AGI 理解引擎构建计划底座。

融资方案与阶段目标 (The Ask)

本轮诉求：**Seed / A 轮 - $3.5M**

- 启动 7B 级原生大模型基座训练环境。
- 组建成熟的 GTM 和企业级销售团队。
- 核心目标： 12 个月内斩获 10 家全球头部企业 PoC，锁定 $1.5M 的年度经常性营收 (ARR)。

资金使用占比：

- **50%** 算力与研发：租赁包含 H100 架构算力集群用于底层架构联合预训练。
- **30%** 核心架构研发团队：招募与构建具有深厚算子优化、模型内核改进经验的资深 CUDA 及系统工程研发人才体系。
- **20%** 市场与系统工程支持：用于大型核心 B 端商业拓客支持节点建立与持续宣发。

核心团队：Awareness

- **Everest An** – 创始人 (前沿计算系统与神经网络架构师)
- 去中心化基建经验：前 **IPFS Athna** 核心成员，曾主导并全面负责中国地区企业级分布式数据库的底层架构建设。
- 头部资本与生态视角：曾任出海顶级基金 **Outliers Fund** 投资管理经理。重点参与投资顶尖公链体系 (**Dfinity**、**Harmony**)、前沿基础设施 (**Analog**)，及区块链核心代码安全审计机构 (**Quantstamp**)。
- 顶级开发者底色：具备全球顶尖的工程技术实力，在竞争最激烈的全球以太坊 (**Ethereum**) 黑客马拉松大赛中多次斩获冠军。

我们坚如磐石的技术防卫高墙

- 开源生态布局战略：我们计划将高层模型 API 与核心组件代码积极融入开源生态，借助广大开发者的共创与试错打通应用层瓶颈，快速验证商业构想并稳固社区信仰高地。
- 壁垒级核心闭源资产：对于最底层的 CUDA 原生高度定制化算子调优、动态特化编译器加速组件以及复杂环境下的高效压缩路由引擎代码方案将被列为我们不可轻易复现的商业化核心壁垒与高保护级别代码资产。

结语

- 传统算力困境遭遇瓶颈：模型发展正处于极度内卷的瓶颈期，单纯的堆砌算力带来的边际收益已大幅降低。仅靠扩大集群无法从根本上跨越内存功耗痛点墙。
- 底层流机制的范式转移：转向更具生物启发性质与动态内状态循环记忆能力的架构才是彻底解构目前“高成本困境长城”的根本有效破局路径。
- 绝佳的时间节点：项目当前不仅仅具备完善的跨时代理论架构支撑，同时辅以充分完备的大批量基础对比论证代码及可跑的测试验证案例。现在正是投资这一新型赛道的黄金入局契机。
- *Invest in Awareness. Invest in the Liquid Computing Paradigm Shift.*

联系方式与相关主页

Awareness

Everest An – 创始人 / 核心架构师

Email: everest9812@gmail.com
WeChat / Telegram: EverestAn

论文 **(Paper):** `https://huggingface.co/EverestAn/MT-LNN`
代码 **(GitHub):** `https://github.com/everest-an/01`
在线演示 **(Demo):** `https://huggingface.co/spaces/EverestAn/Awareness01`

欢迎交流，共同探索下一代液态算力革命。