2025 年末 AI 总结：国民应用和 Cursor for X

周枫

2025年是AI从早期浪潮步入中程阶段的一年。技术在继续演进，故事开始回归现实。借此年末，我做一个简短总结，谈谈过去一年看到的行业变化、创业机会与个人感受。

行业与模型

中美 AI 国民应用都已出现。不管是中国还是美国，AI十年与移动互联网十年呈现出高度相似性：都会诞生跨年龄的大众级超级应用。

OpenAI 的 ChatGPT 从 2022 年末上线，通过免费、手机 App、语音模式、GPT Store 等一系列升级，2025 年已是全球“国民应用”。与之对照，豆包在 2025 年底日活突破一亿，成为国内真正意义上的国民级 AI 助手。不同在于：ChatGPT 已深度商业化，而豆包、DeepSeek 在 C 端仍完全免费，路径尚未与 ChatGPT 完全重合。

模型思考能力和智能体能力取得巨大进展，多模态、长记忆等能力相对停滞。2024 年中的 GPT-o1 是第一个具有思考能力的模型，2025 年初的 DeepSeek R1 开源给行业带来巨大震撼，到 2025 年底，主流模型基本都具有了思考能力和多轮工具调用或者说智能体（Agent）能力。国内模型的多模态能力明显落后于美国同行，而前沿模型本身在 2025 年多模态方面的进展也不大，而长记忆等更高级的能力在 2025 年也没有明显的进展。

前沿模型和开源模型保持了微妙平衡。以美国企业为主的前沿闭源模型在综合能力上仍然保持领先，而以中国企业为代表的开源模型，则在成本、开放性与生态扩散速度上形成了明显优势。在 LMSys Arena 等公开评测榜单上，2025 年 8 月底前三名仍由美国前沿模型占据，但前十名中已有五席由中国模型获得，数量和活跃度均显著提升。DeepSeek、通义千问、智谱、MiniMax、Kimi 等厂商在过去一年里频繁迭代开源模型与低价 API 策略，使其在全球开发者社区快速传播，并在海外市场取得可见进展，已经成为不少硅谷创业公司的主力模型选择之一。

大模型之外的其它模型也可以很有价值。2025 年行业最具存在感的意外，来自 Google 在 8 月推出的 Nano Banana。它在图像生成与编辑领域展现出极高的一致性与质量，直接打开了电商展示、广告创意、产品设计等严肃商业场景。它所传递的信息很清晰：除了通用大模型之外，围绕视觉、视

周枫
网易有道 CEO。他拥有清华大学计算机学士和硕士学位，后获得加州大学伯克利分校博士学位，在顶级国际学术会议和期刊发表超过10篇重要论文，主要研究领域是面向大规模服务的软件基础设施、操作系统和编程语言。他始终保持对领先技术的关注和热爱，主持有道多款产品开发。2023 年带领团队研发出全国首个教育大模型子曰，并率先落地一系列大模型应用。

频、音频、翻译等具体任务打造的专业模型，同样可以孕育出大规模应用与可观的商业价值。

这一方向，也是我们长期关注并持续投入的路径。2025 年，有道词典的 AI 同传功能在“子曰”翻译模型驱动下，使用量实现数量级增长，体现出非 LLM 的专业模型在垂直任务中的效果与性价比优势。因此，我们非常看好这类“专用能力模型”的发展潜力。2026 年，视频与语音等领域是否会出现类似级别的突破，值得持续关注。

应用与创业

AI 与移动互联网技术最大的区别在于，模型本身即是应用。移动互联网时代的生态结构相对分层清晰——基础设施（如通信网络、操作系统、应用市场）与上层应用之间存在明显分工，大部分基础技术公司并不会直接深入消费端产品。AI 时代的格局则有不同：到目前为止，头部企业大多选择从底层模型能力一路延伸到终端产品，推动“模型 + 工具 + 平台 + 应用”的一体化布局。

Anthropic 是典型代表：底层通过 Opus/Sonnet/Haiku 等系列模型构建推理引擎；中间层向开发者提供 Claude API 和 Claude Enterprise；上层则有面向终端用户的 Claude、面向开发者和工程师的 Claude Code。Claude Code 自 2025 年初推出后，上线 6 个月 ARR 即达到 10 亿美元，不仅证明了“模型能力本身的商业价值”，更意味着“围绕模型构建产品和工作流能够产生规模化收入”。这种从模型到产品的纵向整合，在 AI 市场占据了大体量的用户、数据和收入份额，是与移动互联网时代显著不同的结构特点。

场景比模型能力更重要，机会在于把大型需求用 AI 重做。在大型刚需场景中，我认为有一些是模型厂商的机会，有一些是纯应用团队的机会。可

SpaceOne AI 答疑笔（子曰多模态识图 / 解题界面）

以看到的一些方向包括：搜索、翻译、个人助理、娱乐与陪伴、办公自动化、知识管理、法律 / 医疗 / 编程等专业工作、个性化学习、教育内容生产、营销与获客、智能客服、决策支持、视频生产、图像设计与数据处理等。真正的门槛，正在从“能力演示”转向“重构工作流”。

“Cursor for X”是创业的好目标。关于寻找什么样的场景切入，我认为 Cursor 这个案例很值得研究。Anysphere（Cursor）是当前体量最大的“应用起家”AI 公司：以 AI 化 IDE 为切入口，重构开发工作流，从需求理解、代码生成到调试与重构形成闭环，从而带来强付费意愿与高使用粘性。Cursor 的成功，使“Cursor for X”（即在某一垂直领域打造一个 Cursor）逐渐成为投资人与创始人之间的常用提法——把 Cursor 的方法论复制到律师、金融、教育、创作等高价值流程中。沿着这一路径，已经出现了一批具有代表性的产品：Harvey 面向法律服务、Den 面向知识工作者工作台，NotebookLM 面向知识整理与学习，它们都不再只是“在原有产品

上加一个聊天框”，而是以 AI 为核心，对某一专业工具进行从工作流层面的重塑。这一类产品的共性在于：把模型、工具调用与上下文管理深度结合，真正提高单位时间产出，而不仅仅提供“回答问题”的界面。

会员是成熟的变现形式，广告是未知的变量。ChatGPT 已证明“生产力工具订阅”具有强商业可行性；国内头部产品则多保持 C 端免费，利用低价 API 与开源生态协同扩大影响力。中小型 AI 应用目前主要采用「基础免费 + 高阶会员 / 订阅」的 Freemium 模式。至于聊天式 AI 助手是否会走向广告化变现，仍存在较大不确定性：广告与“可信、专注、为我工作的助手”这一产品心智能否共存，既是产品设计问题，也是商业模式问题。这很可能会成为 2026 年 AI 应用生态中的重要观察变量之一。

算力

一个好应用往往消耗大量 token。从 2024 年到 2025 年，全球 token 消耗量持续指数级上升，复杂任务的平均调用深度和自动化比例都在显著提高。以中国市场为例，2024 年 1 月到 2025 年 6 月，日均 token 消耗量从约 1000 亿增长到约 30 万亿，增幅 300 倍（国家数据局数据）。

我个人判断，在未来五年内，国内 token 使用量保持“年十倍级增长”是大概率事件。我常和同事打趣说 Claude 是“编程 5 分钟，等待 5 小时”——token 消耗极快，用尽后只能等额度刷新。但这个现象的另一面，是应用形态的变化：越来越多工作不是“人工 + 辅助生成”，而是直接交由模型自主跑完整个流程。只要算力价格继续下降，那些今天看似“烧钱”的智能编程、智能体和图像视频生成应用，未来反而会成为最具护城河和商业价值的方向。

非 GPU 架构的芯片值得关注。从 Intel 拟收购 SambaNova，到 NVIDIA 授权 Groq IP 并吸纳核心团队，背后的共同信号是：在以推理为主的 AI 应用时代，算力格局不太可能只由 GPU 一种架构长期主导。SambaNova 采用数据流（dataflow）为核心的专用加速器路线，强调在特定模型上的高效率；Groq 则以“语言处理单元（LPU）”为特色，主打极高吞吐和确定性延迟，适合大规模低延迟推理。

国内本就存在 GPU 与 NPU、ASIC 等多架构并行发展的格局。如果未来非 GPU 架构在功耗与成本上持续拉开差距，并与特定应用（如推理、边缘部署）深度结合，那么完全有可能形成一个具有差异化优势的独立算力体系，而不是单一 GPU 的简单补充。

个人

程序员职业被颠覆是大概率事件。2025 年是编程模型大发展的一年，年初还无法看清是否机器真的会编程，年末它已经能接管很多开发流程。这种变化对行业的冲击是结构性的，我自己也在不断思考由此带来的生产方式调整与人才培养转向。

对于以编程为职业的人士，以及即将进入这一行业的学生，我有两点建议。第一，要尽量把能力上移到更高层级：系统设计、架构把控以及对业务的深入理解，逐步超越简单页面搭建、脚手架生成等重复性强且易被自动化替代的工作。第二，要尽快掌握并熟练使用新一代工具：Agent、Skills、MCP、LSP，以及 prompting 与 context engineering 等方法论，并把它们融入日常实践。能够指挥 AI 高效完成复杂任务，将会成为未来程序员最基础也最重要的能力。

愿我们都在变化的时代里，找到属于自己的答案。