引用格式：武延军．DeepSeek引发的AI创新和开源生态发展的思考．中国科学院院刊，2025, 40(3): 446-452, doi: 10.16418/j.issn.1000-3045.20250225007.
WU Yanjun. Thoughts on AI innovation and open source development: Lessons from DeepSeek. Bulletin of Chinese Academy of Sciences, 2025, 40 (3): 446-452, doi: 10.16418/j.issn.1000-3045.20250225007. (in Chinese)

DeepSeek引发的AI创新和开源生态发展的思考

武延军

中国科学院软件研究所　北京　100190

摘要　在人工智能（AI）领域竞争激烈的重要时刻，DeepSeek发布了V3/R1等基础大模型产品，性能比肩国际领先机构OpenAI，不仅展现了中国AI领域科技创新的实力，更为全球AI发展提供了来自中国的创新路径：一是低成本训练推理，打破高端算力的垄断封锁，降低研发应用门槛；二是全栈、全系列的开源开放，支持按需自主部署，普惠各行各业。这一来自中国的技术创新和开源实践，值得学习借鉴。文章将从“以软补硬”“开源传播”“生态优先”3个步骤归纳DeepSeek的开源模式创新之处。同时，也从大模型入口、开源软件供应链、开源基础设施3个方面，分析当前我国AI开源创新仍然面临的问题和风险。最后从大模型操作系统创新、软件供应链保障、开源基础设施建设、软硬件协同发展4个角度，提出加强我国AI创新与开源软件基础能力的建议。

关键词　人工智能，基础软件，开源，基础设施，软硬件协同

DOI　10.16418/j.issn.1000-3045.20250225007

CSTR　32128.14.CASbulletin.20250225007

2025蛇年春节前后，杭州深度求索人工智能基础技术研究有限公司（以下简称“DeepSeek”）发布的开源大模型引起了国内外广泛关注。首先是模型基准测试性能与世界领先的OpenAI闭源模型GPT-4o比肩，其次是训练成本相比其他模型大幅降低，且带思考链的推理模型R1及其蒸馏版本可以在多种计算能力设备上部署，最后是其代码、文档、模型权重等在MIT许可协议（极为宽松的一种开源许可协议）下完全开源。这一套集高性能、低成本、开源开放于一体的“组合拳”，使得DeepSeek在短时间内成为国内外人工

修改稿收到日期：2025年3月9日

智能（AI）领域的焦点，后续接踵而至的各行各业推广部署，让大模型应用在中国真正实现了“飞入寻常百姓家”。

大模型[①]从形态上是一种软件。虽然模型文件通过训练生成，通过参数和数据迭代，以概率性输出结果，无法精确断点调试，黑盒特征明显；但与传统软件一样，它可复制、可复用，需要操作系统提供运行环境，需要存储系统，需要处理用户输入并输出反馈。因此，DeepSeek大模型这一来自中国本土的技术创新和开源开放实践，也为中国软件行业提供了可深入分析并学习借鉴的模式。

本文将DeepSeek的创新模式归纳为“以软补硬”、“开源传播”和“生态优先”。同时，也从生态入口、开源软件供应链、开源基础设施3个方面，分析当前我国AI开源创新仍然面临的问题和风险。最后从大模型操作系统布局、软件供应链保障、开源基础设施建设、软硬件协同发展4个维度，提出加强我国科技基础能力的建议，以期更好支撑中国创新团队的长足进步发展，不断抢占AI和软件领域的全球科技制高点。

1 DeepSeek的创新模式分析

1.1 “以软补硬”开辟大模型创新路径

在算力资源受限的背景下，DeepSeek通过软件架构创新和算法优化，使其模型在保持高性能的同时，大幅降低了对硬件投入的依赖，并为全球开发者提供了可复现、可负担的“以软补硬”技术方案。这让近年来大模型领域普遍推崇的规模定律（scaling law）出现了拐点，依赖大规模硬件投资建立的算力垄断“高墙”出现了缺口，大模型研究和应用的门槛被大大拉低，资源有限的中小企业、研究机构甚至个人，都迎来了AI创新和AI赋能的可能性。

软件在这一轮大模型浪潮中往往被忽视。事实上，对于硬件架构确定、优化目标明确的场景，软件改进带来的总体收益通常大于硬件。2017年图灵奖获得者汉尼斯和帕特森于2018年4月在国际计算机学会（ACM）[②]做获奖演讲时，给出了用不同编程方法计算两个4096×4096矩阵相乘的性能对比，该数据引用了美国麻省理工学院（MIT）计算机科学与人工智能实验室（CSAIL）雷瑟斯等人发表在*Science*上的文章*There's plenty of room at the Top: What will drive computer performance after Moore's law?*[③]（《顶端仍大有可为：摩尔定律之后什么将驱动计算机性能发展?》，这里的“顶端”指代软件），具体对比数据见表1。从表中可以看到，用C语言编写比Python要快47倍，分治法并行优化后可得到6 727倍的加速，而采用SIMD指令集则可加速6万多倍。同样，DeepSeek使用英伟达PTX，即介于CUDA高级编程语言和实际GPU机器代码之间的中间代码表示语言，也起到了极大的加速效果。

在过去几年中，华为鸿蒙操作系统同样采用了“以软补硬”的方法，在处理器制程受限的情况下，通过操作系统、编译器、渲染引擎等多种软件优化手段，在手机上保持了良好的用户体验。

更重要的是，软件优化方案为快速传播奠定了基础。软件之于硬件的一大优势，就是传播的便捷和迅速，通过网络下载就可以快速到达最终用户。试想，如果这次DeepSeek发布的是“星际门”一样的硬件堆叠方案，又或是使用了某种硬件加速方案（如同当年谷歌为深度神经网络专门设计的TPU），将很难如此快速传播推广。

① 这里指模型文件、权重文件，以及配套代码工具、数据集和文档等的集合。

② Association for Computing Machinery，也是图灵奖的颁发机构。

③ “plenty of room at the Top”是在向费曼1959年的著名演讲*There's Plenty of Room at the Bottom*（《底部还有很大空间》）致敬。费曼当时是在讨论纳米技术的潜力，而这篇文章则是在讨论计算机性能提升的新方向。

表1　采用不同编程方法对两个4096×4096矩阵相乘的加速效果

Table 1　Speedups from performance engineering program that multiplies two 4096-by-4096 matrices

序号	编程语言	运行时间(s)	性能	绝对加速比	相对加速比	峰值分数(%)
1	Python	25 552.48	0.005	1	—	0
2	Java	2 372.68	0.058	11	10.8	0.01
3	C	542.67	0.253	47	4.4	0.03
4	Parallel loops	69.80	1.969	366	7.8	0.24
5	Parallel divide and conquer	3.80	36.180	6 727	18.4	4.33
6	plus vectorization	1.10	124.914	23 224	3.5	14.96
7	plus AVX intrinsics	0.41	337.812	62 806	2.7	40.45

1.2 以开源开放实现用户高速增长

软件的核心竞争力是用户。大规模、高质量、多样化的用户群体，不仅是软件价值变现的坚实基础，更是推动软件持续迭代创新的强劲动力。正如中国科学院计算技术研究所包云岗研究员所说，*在开源模式下，软件的价值计算和传播效应遵循梅特卡夫定律(Metcalfe's Law)，即网络的价值与网络中用户数量的平方成正比。*具体表现为两个方面，一是用户规模效应：用户越多，价值越大，反馈和改进更多，生态系统更丰富。二是网络效应：更多开发者参与，就会有更多的应用场景，继而更快的迭代速度。当众多用户转变为开发者和测试者，就会极大地降低软件开发测试成本，驱动软件升级演化和价值提升，继而吸引更多的开发者参与，形成持续的良性循环。

前面提到大模型本身也是一种软件，因此开源软件曾经创造的发展模式，完全可以被大模型所复用。然而，DeepSeek开源模式创造了比传统软件更为迅速的用户增长奇迹。据统计，DeepSeek连续登顶苹果App Store和谷歌Play Store全球下载榜首，上线18天累计下载量突破1 600万次，远超Chat-GPT发布首月的900万下载量。这其中固然有大模型概念热度的加持，但更有DeepSeek几乎毫无保留开放了模型文件、权重文件、核心代码和技术文档的原因。由此在短短半年内吸引了全球超过百万开发者，建立了活跃的开发者社区，不仅贡献了大量的代码和工具，还形成了自发的技术交流和学习氛围，例如GitHub上DeepSeek所维护的awesome-deepseek-integration页面。这种社区驱动的创新模式，为AI技术的快速迭代和应用落地提供了强大的动力。DeepSeek的经验也表明，即便在AI时代，开源开放仍然比封闭垄断更具竞争力。

1.3 以标准化接口和工具构建上下游生态

DeepSeek在建立生态方面同样展现出了很高的效率，在短短一个月内，DeepSeek R1从满血版671 B到70 B、32 B、7 B甚至1.5 B等大小不同模型得到快速部署，大到云服务厂商、互联网巨头、国资央企、高校院所，小到街道办、实验室、个人用户等，从制造业到服务业，从教育到医疗，DeepSeek渗透到各行各业，推动效率提升和智能化转型。

生态快速壮大背后则是其对调用接口和AI软件工具包的标准化，以及因此而快速聚集的上下游生态伙伴。标准化调用接口简化了AI应用的接入流程，使得DeepSeek很容易被Ollama、vLLM、SGLang等大模型服务框架所支持，也使得ChatBox、AnythingLLM等大模型入口应用能够很快接入DeepSeek。标准化软件工具包大幅降低了AI应用部署门槛，同时还提供了丰富的预训练模型和数据集，使得开发者可以通过领域

精调和检索增强生成（RAG）实现自身业务需求，进一步开展应用创新；同时，使得华为昇腾、寒武纪等其他非英伟达芯片能很快完成适配，形成百花齐放的国产软硬件协同适配景象。

从更宏观的生态视角看，DeepSeek已经在中国建立了事实上的大模型标准。自从2020年底Chat-GPT发布以来，无论美国还是中国都进入了“百模大战”的格局，尽管OpenAI引领了发展，建立了提示词工程（Prompt Engineering）等事实标准，但因其选择闭源策略，且其最大投资者微软公司的Windows操作系统同样闭源，使得“应用—模型—系统—硬件”生态链路参与者无法自主开展大模型和系统的适配，阻碍了参与者的参与意愿和创新动力。例如，对于大量非英伟达的硬件加速卡厂商来说，因为无法修改基础模型和相关代码，只能模拟与转译英伟达GPU指令集，无法实现与模型的原生适配；对于亚马逊、谷歌、阿里等云平台服务商来说，由于与微软Azure的竞争关系，也无法与OpenAI实现充分的业务整合。

DeepSeek开源发布之后，不仅出现了微信、WPS等应用整合，也出现了华为云、阿里云、腾讯云等服务集成，还出现了华为昇腾、寒武纪、沐曦、海光、申威等硬件原生适配，甚至出现了大量本地部署的一体机解决方案。以DeepSeek为大模型事实标准，中国正在形成“应用—模型—系统—硬件”全链路的生态聚集。长远来看，这一变化必将重塑中国乃至全球AI的发展格局。

2 我国AI开源创新面临的风险挑战

在看到DeepSeek成功一面的同时，还需要看到当前中国AI开源创新面临的一些风险挑战。

2.1 大模型入口程序的风险

所谓大模型入口程序，对于部署者是指Ollama、SGLang、vLLM等大模型服务框架程序，用来启动大模型服务进程；对于用户是指通过封装多个大模型服务，为用户提供更加方便易用、灵活可配置的交互界面程序，如ChatBox、AnythingLLM等。

以Ollama为代表的大模型服务框架，在启动大模型服务时通常以网络守护进程的方式出现，会打开某个端口并监听来自网络的服务请求。这样的守护进程一旦出现漏洞，攻击者很容易通过服务端口入侵服务主机。事实上，近期已经发现了Ollama导致的、可被利用的大模型服务漏洞。

而对于用户交互的入口程序来说，尽管ChatBox等通过开源来证明自身程序的安全性，但却无法证明用户隐私数据的安全性，毕竟所有的对话信息都会被入口程序转发和截取。

对主流入口程序的掌控和主导，会成为大模型竞争的焦点之一，但目前为止，大模型的入口程序还是运行在已有主流操作系统之上，因此操作系统不自主可控的风险将会延伸到大模型入口程序，毕竟操作系统很大程度上决定了谁能成为入口，20世纪90年代网景公司NetScape浏览器在与微软IE浏览器竞争中败北就是前车之鉴。

2.2 软件供应链的安全可靠风险

DeepSeek的开发依赖大量开源或闭源组件。例如：基础框架中的PyTorch深度学习框架、CUDA GPU加速库；训练相关的Megatron-LM分布式训练框架、Flash Attention高效注意力机制；推理优化相关的FasterTransformer推理加速引擎、TensorRT推理优化库、ONNX模型转换标准库；工具链中的版本控制Git、容器化部署Docker；数据处理中的NumPy数值计算库、pandas数据处理库，以及HuggingFace数据集管理工具等。

以上仅是基于公开信息的判断，实际使用的工具可能更多，有些专有工具可能未公开。而在这些互相高度依赖的软件供应链中，有些关键环节仍然被Meta公司等国际竞争对手掌控（如PyTorch开发框架，以及前面所述的Ollama入口程序），或属于某家公司私

有产品（如英伟达CUDA），均存在断供可能。此外，根据奇安信的最新报告，已出现一些专门针对DeepSeek的供应链伪造或投毒攻击。这些都构成了我国AI面临的软件供应链安全可靠风险。

健康的大模型生态需要一个同样健康的开源软件生态。对于软件供应链，特别是开源软件供应链关键节点的认真梳理和持续维护，仍然是企业和行业，甚至国家实现人工智能高水平科技自立自强必须要付出的投入。

2.3 开源基础设施的风险

不仅DeepSeek，国内主要开源大模型项目几乎都选择在美国微软公司旗下的GitHub平台发布，这是因为GitHub全球开发者集中度最高，有完整的开源基础设施能力、成熟协作工具链和已经发展壮大的程序员社交网络，因此国际影响力更大，更有利于项目推广。然而，选择GitHub未来也面临挑战和风险，包括但不限于地缘政治风险、数据主权问题、潜在的访问限制风险等。这并不是DeepSeek和国内开源项目维护者的问题，而是国内缺乏与GitHub竞争的开源基础设施，从设施完善程度、开发者聚集规模、国际化程度、运营能力等，国内现有基础设施与GitHub相比都存在较大差距。

Hugging Face近年来随着大模型爆发而异军突起，成为全球最流行的模型托管平台，国内的阿里魔搭等平台虽然已经起步并初具规模，但与Hugging Face相比，同样在功能、规模、国际化、运营等方面存在显著差异。

3 加强我国AI创新能力的建议

基于以上分析，本文提出加强我国AI创新能力的如下建议。

（1）尽快启动大模型操作系统的研发探索。大模型仍然以软件的形态存在于现有操作系统生态体系，虽然出现了ChatBox等新的入口程序，但不足以撼动Windows、iOS、Android的生态主导地位。美国苹果公司和我国华为公司先后提出了面向意图的开发框架，旨在整合大模型的能力，继续掌控用户入口。微软公司通过预装Copilot并与办公套件、浏览器等深度捆绑，巩固其桌面领域垄断地位。上海交通大学陈海波团队提出了大模型操作系统的3种技术路线，即渐进路线（大模型作为操作系统外挂组件）、激进路线（大模型即操作系统）和融合路线（大模型与操作系统深度融合），并建议采用融合路线，从而在利用大模型能力的同时，最大程度兼容现有操作系统应用生态。鉴于大模型带来的机器智能跃升和交互范式变革，无论采用何种路线，大模型操作系统研发工作都迫在眉睫。随着大模型和操作系统各自发展，不同技术路线会自然合并，然而一旦错过生态初始构建的机会窗口期，将面临新的、更难突破的生态垄断。

（2）加强开源软件供应链治理。开源软件已经成为组装大型复杂系统软件的“原材料”和“元器件”。一个Linux开源操作系统发行版（如Debian、openEuler等）往往包含上万个开源组件，通过这些组件的彼此依赖关系编译组装而成。一个大模型从开发、训练到部署、运行、推理，也依赖于大大小小的开源组件。随着大模型成为像操作系统一样的战略基础软件，其开源软件供应链的保障必不可少。中国科学院软件研究所从2019年发起“开源软件供应链点亮计划”，梳理全球开源软件知识图谱，找出操作系统等大型复杂基础软件的关键供应链节点，通过“开源之夏”等活动，持续培养能够看护关键开源软件的高水平人才。建议围绕大模型的开源组件依赖情况，持续梳理开源软件供应链，对其中关键节点进行重点布局，投入或培养相应的人力资源，确保具备持续开源维护的能力。

（3）加快对标GitHub和Hugging Face的开源基础设施建设。面对GitHub和Hugging Face托管平台的垄断局面，一方面继续完善现有国产代码托管平台，提

升平台稳定性和功能完整度，优化开发者体验。另一方面也要有过渡策略，采用多平台同步策略，建立战略备份机制。从2019年中国科学院软件研究所启动建设“源图”开源软件供应链基础设施，迄今已形成对全球关键开源软件的全量备份，并提供可信软件仓、可信编译构建环境等平台服务。后续还需要面向大模型的新需求、新场景，加快打造新一代开源开发基础设施，联合国内优势力量逐步培育本土开源基础设施生态，并以更加开源开放的模式，吸引国外机构和开发者参与，共同对冲潜在的地缘政治风险。

（4）加大开源软硬件协同力度。在新一届美国政府不断升级管控施压的背景下，英伟达GPU硬件供应限制和CUDA软件生态壁垒，已经成为中国实现AI领域高水平科技自立自强面临的最主要障碍之一。例如，DeepSeek训练优化所使用的PTX仍然属于CUDA生态体系。建议加大RISC-V开源指令集下软硬件协同，特别是AI相关扩展指令集的协同力度。RISC-V指令集的崛起，不仅为了从指令集层面打破x86/ARM的生态垄断，同时也有望打破英伟达GPU私有指令集和私有算子的垄断。随着RISC-V向量指令集、矩阵/张量指令集的制订和完善，新的软硬件接口标准规范有望取代CUDA私有接口规范，并配合编译器等在RISC-V专用AI加速卡上实现软硬协同。一旦某款RISC-V加速卡在性能功耗比上超越英伟达的旗舰GPU，整个RISC-V生态也将迎来“DeepSeek时刻”。

需要强调的是，以上风险分析和建议，并非为了形成封闭的、防御式的技术体系，而是为了中国乃至全球都有更为开源开放的选择，平等参与AI新技术、新产品、新服务的研发应用，共同打造AI时代的人类命运共同体。

Thoughts on AI innovation and open source development: Lessons from DeepSeek

WU Yanjun

（Institute of Software, Chinese Academy of Sciences, Beijing 100190, China）

Abstract At a critical moment of intense competition in the artificial intelligence (AI) field, DeepSeek has released foundational large language models (LLM) such as V3/R1, with performance comparable to leading international organizations like OpenAI. This not only demonstrates China's technological innovation capabilities in AI but also provides a Chinese innovative pathway for global AI development. Firstly, through low-cost training and inference, break the monopolistic barriers of high-end computing power and lower research and development thresholds. Secondly, through full-stack and comprehensive open-source strategies, support customizable and local deployment that benefits various industries. This technological innovation and open-source practice from DeepSeek deserves in-depth discussion and learning. This study summarizes DeepSeek's innovative model from three perspectives, namely, compensating hardware with software, acquiring users through open source, and ecosystem priority. Meanwhile, it analyzes the current challenges and risks for China's AI open-source innovation, in terms of ecosystem portal, open-source software supply chain, and infrastructure. The study concludes by proposing suggestions to strengthen China's AI technological foundation in four aspects, i. e., LLM operating system innovation, software supply chain governance, open-source infrastructure construction, and software-hardware collaboration.

Keywords artificial intelligence (AI), system software, open-source, infrastructure, software-hardware collaboration

武延军 中国科学院软件研究所副所长、研究员。开放原子开源基金会开源安全委员会主席，中国计算机学会开源发展委员会执行委员。主要研究方向为操作系统。E-mail: yanjun@iscas.ac.cn

WU Yanjun Deputy Director of the Institute of Software, Chinese Academy of Sciences (CAS), Chairman of the Open Source Security Committee of OpenAtom Foundation, Member of Open Source Development Committee of the China Computer Federation. His research focuses on operating system. E-mail: yanjun@iscas.ac.cn

■责任编辑：文彦杰