2025-01-07 LLM News

\_标题：LLM News\_

\_一：LLM 新闻\_

\_1. DeepSeek V3 模型展现了与以往中国大语言模型不同的审查机制，其审查主要在推理阶段，且可通过非中国地区的 API 服务规避。\_

\_2. Nvidia Triton 推理服务器在配置特定模型时面临诸多挑战，反映了当前 AI 工具链的复杂性和文档质量参差不齐的问题。\_

\_3. GPT4All 推出了具备推理能力的系统，该系统专为结合迭代式 LLM 输出、思维链和工具调用而设计，旨在解决更复杂的问题。\_

\_4. 持续训练 Llama 3b 可显著提升数学能力，通过在 1600 亿个高质量 tokens 上持续训练，Llama 3b 模型的数学能力可提高 2-3 倍。\_

\_5. Hugging Face 对 Llama 3.2 3B 模型进行了持续预训练，显著提升了数学能力，相关模型、数据集及复现代码均已开源。\_

\_6. Qwen2.5 14B 可以在树莓派上运行。(注：原文标题如此，但内容为空，此处仅作字面转述。) \_

\_7. AMD 在 CES 2025 上宣布推出新型 Ryzen CPU (AMD Ryzen AI Max+ 395)，专为笔记本电脑设计，在运行 Llama 3.1 70B-Q4 模型时，速度比 RTX 4090 独立桌面 GPU 快 2.2 倍。\_

\_8. RTX 5090 显卡：据传 RTX 5090 将拥有 1.8 TB/s 的显存带宽和 512 位显存总线，以及 32GB GDDR7 显存。\_

\_9. 用户正在探索使用 NVIDIA GH200 超级芯片进行本地部署，并分享了 Llama 3.1 70B 在 vLLM 上的基准测试结果。\_

\_10. 有用户询问是否有人在仅使用 CPU 的 AWS EC2 实例上成功运行了 Deepseek V3。\_

\_11. Sam Altman 的最新博文提到 2025 年可能会看到首批 AI 智能体“加入劳动力队伍”。\_

\_12. Hugging Face 发布了其内部使用的评估框架 Lighteval。\_

\_13. DeepSeek v3 数据隐私问题引发关注。\_

\_14. DeepSeek v3 在 2x M2 Ultra 上使用 MLX.distributed 实现了 17 tps 的推理速度。\_

\_15. 结构可能成为 AI 模型的枷锁: AI 结构越少，AI 方法的可扩展性越强。\_

\_16. Qwen 2.5 支持 128k 上下文长度，但输出仅支持 8k。\_

\_17. OpenAI 的 o1 模型使用强化微调技术构建。\_

\_18. DeepSeek v3 在 MBP 128GB 上运行情况不明，但可以运行 v2.\_

\_19. DeepSeek V3 模型发布并广受好评，拥有 6000 亿参数，展现出强大的性能，尤其在长文本生成和问题诊断方面表现出色。\_

\_20. Dolphin 3 模型发布: 作为一种新型的非审查模型，Dolphin 3 通过在“有害”数据集上训练来减少拒绝回答的情况。\_

\_21. VITA 开源项目: VITA 项目致力于开发一个开源的交互式全能多模态大型语言模型。\_

\_22. Kotaemon-papers 应用发布: Kotaemon-papers 是一款开源、可自托管的 Web 应用，允许用户与论文进行交互。\_

\_二：实用技巧：\_

\_1. 使用多个 GPU 运行大语言模型：通过 PCIe 转接卡将多个 GPU 集成到一台计算机中，并利用 GGUF 等格式将模型层分布到各个 GPU 上。\_

\_2. 利用多台计算机运行大语言模型：理论上，可以通过局域网将模型层分布到多台计算机上运行。\_

\_3. Llama 3.3 Q5 或 Q8 GGUF 模型可以在 4090/i9 14900k 平台上运行，可通过 llama-cpp-python 库将模型部分加载到 GPU 上。\_

\_4. 创建用于编程模型的自定义分词器：使用自定义分词器将每个关键字/标准函数作为一个 token 可以提高编码大语言模型的效率。\_

\_5. 本地部署 LLM 可节省开支：相比于订阅服务，本地部署 LLM 可节省开支。\_

\_6. 利用 RTX 3090 运行本地模型：NVIDIA GeForce RTX 3090 显卡能够支持本地运行大型语言模型。\_

\_7. 处理 SMPL/AMASS 数据集：SMPL/AMASS 数据集用于人体动作相关的微调项目，以 npz/pkl 格式存储，包含三维数据，需要特别注意处理方式。\_

\_8. LM Studio 在 AMD 8945h 处理器上运行情况：用户报告称，虽然 8945h 处理器的内置 GPU 和 NPU 性能不强，但 LM Studio 显示支持 Vulcan，不支持 ROCm。\_

\_9. 用户寻求推荐低代码/无代码 AI 自动化工具：用户希望找到合适的工具，以便利用 AI 代理构建概念验证 (POC)，且只需少量或无需编写代码。\_

\_10. DeepSeek-V3 模型可以在具备 96GB VRAM 和 256GB RAM 的 Linux 系统上运行，用户分享了详细步骤。\_

\_11. 用户可以通过修改配置文件，优化 RTX 3090 上运行的本地模型。\_

\_12. 即使使用拥有大量科学论文的 RAG 知识库，仍然可能会因为上下文窗口限制而丢失部分信息，应选择更大上下文窗口的模型，或尝试其它方案解决此问题。\_

\_13. 本地部署 Web 应用：用户正在寻找托管运行本地大型语言模型的 Web 应用的方法，并考虑使用 VPS 或云服务提供商。\_

\_14. IRC 问答机器人的免费 API：用户正在开发一个 IRC 问答机器人，并寻找免费的 AI API 用于问题生成。\_

\_15. 代码分析工具：用户正在寻找能够分析整个代码库的工具，要求能够理解代码库的上下文，并提供代码质量反馈、潜在 bug 识别、单元测试生成和文档生成等功能。\_

\_16. VSCode 插件 Continue 调试：用户在使用 Continue VSCode 插件与 Ollama 结合运行 Llama 3.1 8B 时遇到问题，正在寻求帮助。\_

\_17. Lighteval 的使用：可以通过 `pip install lighteval[vllm,litellm]` 安装，并使用类似 `lighteval vllm "pretrained=meta-llama/Llama-3.1-70B-Instruct,dtype=bfloat16" "lighteval|gsm8k|5|1" --use-chat-template` 的命令进行评估。\_

\_18. 模型共享：一位助理教授询问如何与不熟悉计算机的同事共享模型访问权限，他面临的挑战是建立一个带有用户账户和聊天功能的 Web 界面 (LLM + RAG)。\_

\_19. Qwentile 2.5-32B-Instruct 模型被认为是一个出色的全能型模型。\_

\_20. 用户寻求适用于细致语言理解任务的模型，要求模型至少能够处理英语、法语和德语。\_

\_21. 多 GPU 系统进行 LLM 推理的注意事项。\_

\_22. 本地运行 LLM 的优势和挑战：可以保护用户数据隐私，需要考虑硬件配置和模型选择，LM Studio 是一个可以考虑学习的工具。\_

\_23. 对于多图像分析，可以尝试比较 10 张图像，识别出不包含在 5 张产品图片中的特定产品视图。\_

\_24. Deepseek V3 可能是英语/中文国际化的最佳模型。\_

\_25. DataBridge：一个完全本地化、开源的多模态 RAG 解决方案，可通过修改 `config.toml` 文件进行定制化部署。\_

\_26. DeepSeek V3 Coder 模型部署: DeepSeek V3 Coder 可以在配备双至强 v4 处理器和 512GB 内存的服务器上运行。\_

\_27. 面试准备: 针对侧重于基础模型和视频理解的研究职位面试，应重点关注与视频理解相关的论文。\_

\_三：其他：\_

\_1. 用户正在寻求能够帮助准备考试的本地大语言模型软件（最好是开源的），例如可以从 PDF 格式的课程材料中创建抽认卡、测验等功能的软件。\_

\_2. 用户开发了一个基于遗传算法改进提示词的命令行工具。\_

\_3. 本地模型的重要性日益凸显：当前 AI 服务的低廉价格可能是由于风险投资补贴，这种模式可能不可持续。\_

\_4. OpenAI Pro 订阅正在亏损：Sam Altman 在 X 上表示，OpenAI Pro 订阅目前处于亏损状态。\_

\_5. 用户提供了关于在特定硬件上运行AI模型的具体信息和观察结果。\_

\_6 用户正在积极探索本地 LLM 的持续运行用例。\_

\_7. 用户对本地 LLM 在个人思考、日记记录和准治疗方面的应用表示赞赏。\_

\_8. 用户分享了利用 RTX 3090 运行本地 LLM 的经验。\_

\_9. 存在一个针对 LLM 创意故事写作的基准测试，但具体内容未提供。\_

\_10. 一位新手正在学习使用 Ollama、RAG 以及研究科学期刊，并寻求优化 RAG 知识库以避免信息丢失的方法。\_

\_11. 支持工具和视觉的本地模型：用户正在寻找同时支持工具和视觉的本地模型。\_

\_12. 最新的创意写作大型语言模型：用户正在寻找一个专门针对文本补全写作的大型语言模型。\_

\_13. 用户询问了 F5 TTS 模型在 Windows 或 Linux 上运行量化版本的情况。\_

\_14. 用户提议欧盟应该资助一个开源的欧洲医疗保健 LLM。\_

\_15. 寻找最佳的国际化模型：DeepSeek v3 可能是最佳选择，并寻找相关的基准测试。\_

\_16. Linux 上运行本地 LLM 的框架和工具：不同的 Linux 发行版对运行 LLM 的影响可能不大，LM Studio 可以作为学习的工具。\_

\_17. 询问了关于如何比较多张图像的问题，以识别电商网站产品图中缺失的特定视角。\_

\_19. 对于在 GPU 上运行 LLM 推理的最佳设置，在 1500 欧元的预算下，可以考虑 Tesla M10s, Tesla P40s, Instinct MI50s, RTX 3090s, 和 7900 XTXs。\_

\_20. 建议根据个人的需求和预算决定是否升级硬件。\_

\_21. 非审查模型的类型: 用户对 "dolphin" 和 "abliterated" 这两种非审查模型的区别进行了讨论。\_

\_22. 特征引导生成: 用户讨论了特征引导生成与推理引导生成的区别，并探讨了使用稀疏自编码器 (SAE) 内部引导生成的可能性。\_