2025-01-06 LLM News

```

title: LLM News

一：LLM 新闻

1. DeepSeek V3 模型表现出色，参数量为6000亿，被认为是目前的一个“甜蜜点”，能够提供长篇且深入的回复，并可以通过简单的语言指令在长篇和简短回答之间切换。

2. 开源RAG解决方案DataBridge推出，具有高度可定制性和模块化，允许用户通过编辑`config.toml`文件来定制RAG部署。

3. 关于Dolphin和Abliterated模型的讨论出现，前者使用“有害”数据集进行训练以减少拒绝，而后者则通过“手术”方式来移除模型的拒绝。

4. VITA 项目，一个开源交互式多模态LLM项目被提及，但具体细节未提供。

5. Kotaemon-papers，一个开源的，可以自托管的网络应用程序，允许用户与论文进行对话，提供论文的摘要、思维导图，并支持基于论文内容的问答。

6. Dolphin 3.0 模型发布，基于 Llama 3.1、3.2 以及 Qwen 2.5 的组合。

7. Meta AI 推出 EWE (Explicit Working Memory) 模型，通过整合工作记忆来增强长篇文本生成的真实性。

8. LLaMA 3.1 405B Chatbot 发布，在 Hugging Face 上提供 API 测试空间。

9. GPU 与 RAM 共享内存以扩展上下文窗口的思路被提出，旨在突破硬件限制。

10. Kokoro-onnx TTS 发布，基于 \*kokoro\* 模型构建，并使用 onnxruntime 运行，支持多种平台。

11. Deepseek-v3 模型非常受欢迎，其 671B 版本的下载量预计将超过 QwQ-32B-preview。

12. 出现了利用 DeepSeek v3 模型将 GitHub 代码库转化为即时 Wiki 的工具。

13. o1 模型被用于故事个性化，通过简单的提示技术，让 o1 模型分析并选择最合适的文化元素。

14. URIAL: Untuned LLMs with Restyled In-context Alignment (Rethinking alignment) 论文被提及。

二：实用技巧：

1. 有用户询问关于在Linux上运行本地开源AI框架的建议，包括选择哪个发行版、是否应该投入时间学习LM Studio，以及哪个工具可以使本地LLM集成得尽可能无缝。

2. 关于运行本地AI是否需要强大的GPU的问题，以及付费和免费模型之间的差异也在讨论中。

3. 有用户询问如何在配备双Xeon v4 CPU和512GB内存的系统上运行Deepseek V3 Coder，以及是否有安装指南。

4. 用户分享了他们开发的论文聊天应用Kotaemon-papers，该应用提供论文总结、思维导图和基于论文内容的问答功能。

5. 用户基于“The Dark Matter of AI”视频提出了关于如何通过稀疏自编码器（SAE）来控制LLM内部生成的问题，并引入了GemmaScope和TransformerLens等工具。

6. 有用户提到，他们之前了解到的运行本地大型语言模型的工具是 `ooba booga text gen ui`，但他们认为现在情况可能有所变化。

7. 有用户询问如何为 `SillyTavern` 快速启动 `alltalk`，并具体提出了创建快捷方式的需求。

8. 一位用户分享了他们的硬件配置（i7-12700kf, 128GB RAM, RTX 3090 24GB），并指出显存受限的情况下，在显存中运行模型速度远快于在 CPU 上运行。

9. 有用户提出了对低延迟流式文本转语音 (TTS) 的需求。

10. 一位数据科学家希望扩展技能到本地 LLM 的应用，计划测试多数据点公司身份验证、环境抓取上下文相关问题，并使用 Python 交互。

11. 一位开发者正在本地托管 Ollama，并且在 OpenRouter 中有一些额度，正在选择合适的 JS 库。

12. 一位用户正在寻找可以本地运行的代码补全模型，该模型应能支持多种编程语言，并能与 VS Code 集成。

13. 有用户分享了一个“前沿”基准测试的示例，提醒大家要注意数据集中的潜在问题。

14. 有用户在询问当前运行本地大型语言模型的最佳硬件配置，特别是像 DeepSeek V3 这样的大型模型。

15. 有用户询问如何在 M 系列 Mac 的 Docker 容器中使用 GPU。

16. 有关于 DeepSeek V3 模型在 llama.cpp 中的 token 生成性能如何随着 prompt 长度变化而变化的讨论。

17. 有评论指出，可以使用 gitingest 扩展程序一键将 GitHub 存储库转换为提示。

18. 有人询问是否可以在较旧的 GPU（例如 GT 710）上运行 Ollama，并运行较小的模型。

19. 有用户提出利用随机化的奇异值分解 (SVD) 或主成分分析 (PCA) 来处理更长文本上下文的方法。

20. 有用户正在寻找一种便捷的方式来部署和测试最新的开源模型，而无需承担存储模型的费用。

21. 一位用户正在尝试构建一个廉价的 ARMv9 单板计算机 (SBC) 集群来运行 Deepseek v3 模型。

22. 有用户因为订阅了 ElevenLabs 的 scale 计划而获得了 600 万字符的积分，正在寻找最佳的利用方式。

23. 有用户意识到自己更需要的是本地集成，而不是本地托管模型，正在寻找可以依靠在线模型来处理计算密集型任务的工具。

24. 有用户讨论如何绕过 AI 检测软件，并询问在本地运行模型时，人们还在修改哪些参数来规避检测。

25. 有用户提出了一个关于本地运行 LLM 的有趣想法，即去除预加载的对话或共情能力等“辅助”功能。

26. 有用户在微调期间遇到了 Mistral Nemo 指令聊天模板的问题。

27. 有用户在构建一个简单的聊天机器人（访谈机器人）时遇到了问题，机器人过度关注上下文。

28. 用户已经开始思考“少即是多”的策略，这意味着在提供上下文时，可能需要更加精炼。

三：其他：

1. 有用户寻求关于LLM应用（如提示管理和防护）的每日或每周新闻简报。

2. 有用户正在准备视频理解方向的研究职位面试，寻求关于该领域重要论文的建议。

3. 有用户提出在极其有限的资源下（2-4GB RAM）运行模型的需求，用于每天分析 1-2 封电子邮件。

4. 一位用户创建了一个关于理解英国电视节目《Never Mind the Buzzcocks》中笑话的幽默分析基准。

5. 一位用户询问是否有本地工具或模型可以可靠地检测由 LLM 生成的文本，特别是像 ChatGPT 3 这样的早期模型。

6. 有用户比较了 AMD 7900 XT GPU 和 Apple M2 Ultra 芯片的 token 生成速度。

7. 有用户发起关于 LLM 未来发展最大瓶颈的讨论。

8. 有用户在探讨拥有一个能够与个人邮件、日历和待办事项列表交互的 AI 代理的未来。

9. 另一个讨论围绕着升级本地 AI 服务器以运行更大的模型，比如 Llama 3.3 70B。

10. 评论中提到结构化输出（例如匹配 Pydantic 类）对于 LLM 的重要性。

11. 有用户在使用 LLMLingua 进行代码块压缩方面进行了提问。

12. 一个用户分享了自己搭建的 AI/NAS/Plex/媒体服务器。

13. 有用户正在积极探索运行大型语言模型（如Llama 3 70b）的各种硬件配置。

14. 有用户展示了一个配置了12个3090显卡的服务器。

```