2025-01-12 LLM News

```

title: LLM News

一：LLM 新闻

1. OpenAI 目前处于亏损状态，而阿里巴巴的通义千问模型正在计划推出语音模式。用户猜测通义千问是否能达到 O1（可能是指 OpenAI 的某个模型）的水平。

2. 有用户询问是否存在可以将模型上下文协议（MCP）工具与本地 LLM 结合使用的项目。

3. Moondream 2b 新增注视检测功能。

4. 用户正在探索将多个独立的输入/输出通道与 LLM 集成。

5. 用户正在探索构建基于 Markdown 文件的本地 LLM 知识库。

6. 一项新的研究提出了一种新的 RLHF PPO 训练方法，该方法为每个语义完整的文本段分配奖励，而不是每个 token 或整个文本。

7. Negative\_LLAMA\_70B 模型发布，该模型基于 LLAMA 3.3 70B 微调而来，旨在解决大型语言模型的“正向偏见”问题。

8. DeepSeek 团队缺乏公开信息，包括采访、社交活动和团队照片等，引发社区对其背后团队身份的好奇。

9. 众多初创公司和大企业正在构建 AI 员工，例如 AI 销售代表、软件工程师、营销代理和客户成功代理。

10. NovaSky 发布 Sky-T1-32B-Preview 模型，这是一款开源推理模型，在流行的推理和编码基准测试中表现与 qwen1-preview 相当。

11. 拜登将进一步限制 Nvidia AI 芯片出口。

12. 用户在探索如何让 LLM 理解表格数据 (CSV) 之间的关系。

13. 虽然 Qwen, Mistral, 和 Llama 3.x 等模型宣称具有多语言能力，但用户报告说，在英语以外的语言（如法语）中的表现往往不尽如人意，尤其是在开源模型中。

14. “助手”模式限制了创造力，用户希望看到 LLM 在问答之外的更多应用，如场景生成和创意文本操作。

15. GMK 宣布将推出全球首款基于 AMD Ryzen AI 9 Max+ 395 处理器的迷你电脑，预计将于 2025 年上半年上市。

16. Deepseek 的 API 服务似乎出现问题，通过 OpenRouter 和 Deepseek API 都无法正常使用。

17. AI 员工兴起，Salesforce 主席 Mark Benioff 甚至表示不再雇佣人类员工。

18. 句子嵌入的翻译不变性：一个句子的嵌入向量与其翻译版本（例如英语和德语）在嵌入空间中是接近的。

19. 用户在寻找仅使用非版权数据（如许可代码、知识共享、公共领域等）训练的 LLM。

20. 用户提出对大型语言模型在代码文档和解释能力方面的基准测试的需求。

二：实用技巧：

1. 4070 显卡可以用于训练小型模型。

2. Moondream 模型的视线检测功能非常出色。

3. 开发一个具有语音交互、情绪识别和屏幕内容分析功能的 AI 助手需要掌握微积分及之前的数学知识，C#、C++、JavaScript、Vue 和 Python 等编程语言，预计需要 3.5-4.5 年时间才能实现所有目标。

4. 构建本地 LLM 知识库的一种方法是将 Markdown 文件作为独立的文档进行处理。

5. 为了实现多通道输入输出的分离，一种方法是对每种类型的输入进行标记，并将它们连接成一个单独的流输入到模型中；在输出端，模型需要对所有输出进行标记，并在输出前进行分离。

6. 可以使用网络爬虫工具将 Github 或网站上的文档内容抓取下来，并转换为 PDF 或其他格式，以便于 RAG 系统使用。

7. 对于包含纯文本的 PDF 文档，可以使用现有的 OCR 工具进行文本提取。对于包含图像或 SVG 中文本的 PDF，需要使用支持图像 OCR 的工具。

8. 可以开发工具或脚本，根据 PDF 文档的结构和内容，自动提取指定的章节、注释等内容，并将其保存到新的 PDF 文件中。

9. 当使用 `nvidia-smi` 命令无法识别 Tesla P40 显卡时，可能是由于显卡未正确供电。

10. 确保已安装 Unsloth 和 Hugging Face 库，并已获取访问 Llama 模型的权限。在代码中，需要提供正确的 Hugging Face token。

11. 用户分享了使用 vllm-openai 镜像运行 Qwen/Qwen2-VL-7B-Instruct-GPTQ-Int8 模型的 Docker Compose 配置，通过调整 `--gpu-memory-utilization` 参数 (设置为 0.85) 解决了内存不足的问题。

12. 双 24GB GPU（如 3090）配置 可运行 LLama3.3 Q4 模型。

13. 具有更大显存（如 96GB 或更高）的配置对微调模型、使用更大提示以及进行代码/数据生成、特定领域问答和创建角色等任务更有利。

14. Asus ROG G14（64GB RAM + 16GB vRAM）可以运行部分卸载到 GPU 的大型模型，具体模型大小取决于量化级别。

15. 选择 GPU 时，需要根据实际需求权衡显存大小和数量。

16. 在具有大内存但无独立显卡的设备上运行大型模型时，需要注意性能瓶颈。

17. Mac Mini M4 可以作为本地 AI 服务器的一个经济实惠的选择。

18. 用户寻求一个 UI 来管理多个已部署的 LLM，可以探索开源替代方案。

19. 将对话式讲座记录转换为 Markdown 中的教科书格式，可以使用 ChatGPT 并提供关于语气、风格和数学格式的特定说明。

20. 可以编写 Python 脚本访问本地 LLM 的 API 来进行基准测试。

21. 可以将多个 PDF 文件的文本内容作为上下文一次性输入给现成的模型，然后进行提问，无需微调即可实现简单的问答功能。

22. 用户购买了 3 块翻新的 RTX 3090 显卡，但发现它们都不稳定。建议谨慎选择翻新硬件。

23. ASRock Radeon RX 7900 XTX Creator 显卡是一款双槽单风扇显卡，适合用 Threadripper 主板搭建多 GPU 的 AI PC。

24. 用户可以使用一些模型来进行prompt的优化。

25. 对于少于 4b 参数量的模型，EXAONE-3.5-2.4B-Instruct-Q8\_0.gguf 是一个可行的选择。

26. 用户可以使用老旧的 Powermac G3 电脑构建 AI 工作站。

三：其他：

1. 有一个用户想找一个类似 O1 的模型。

2. 用户期待 RTX 5090 显卡的 LLM 基准测试结果。

3. 使用 Python 实现 Ollama 函数调用的高级指南。

4. 深度伪造技术（deepfakes）正在快速发展。

5. 随着 NVIDIA 的 Project Digits 等本地 AI 硬件的发布，本地 AI 推理有望在未来普及。

6. 如果使用手机和笔记本进行推理，可以考虑云同步方案同步个人上下文数据。

7. 用户目前在询问如何将非语言输出添加到 LLM 中，或者更一般地说，如何将单独的 I/O 通道分开。

8. 用户在关注硬件对 LLM 性能的影响，特别是新一代显卡的性能表现。

9. 用户有构建本地 LLM 知识库的需求，以提高问答系统的准确性和实用性。

10. 用户因不明原因被 r/chatGPT jailbreak 社区封禁。

11. 用户计划使用二手硬件搭建预算为 50,000 美元的 GPU 服务器，并从头构建自己的 Transformer 模型。

12. 用户计划创建一个专门针对伦理、哲学、人类意识和心理学等领域知识训练的“LLM 伦理内核”。

13. YC 公司对 AI 工作流程的兴趣浓厚。

14. 最新一代 Nvidia GPU (50x0 系列) 更注重内存升级 (GDDR7) 而不是 GPU 性能的显著提升。

15. Oculink (高达 PCIe 4x4) 理论上比 USB4 Gen1 (PCIe 3x4) 快，USB4 可能会引入额外的延迟。

16. 远程卸载 LM Studio 中的模型可能不直接支持。

17. 用户正在寻找适合通过 API 连接的远程 LLM 的前端。

18. ToGODer 的文章 "Why AI Agents Are Fundamentally Broken" 对当前的 AI 代理架构提出了批评。

19. 用户有兴趣微调语音转文本 (STT) 模型以支持本地语言。

20. 出现了用于模拟和玩 D&D 的多个本地 LLM 模型的 MVP 项目：GitHub - tegridydev/dnd-llm-game。

21. 有人整理了 MCP 客户端的集合：awesome-mcp-clients。

22. 出现了在真实 LLM 硬件上运行的 LocalGLaDOS。

23. 用户正在寻找构建代理工作流和 RAG 应用的最佳 AI 技术栈。

24. 用户在使用 stable-diffusion.cpp 工具时，发现上下文大小被裁剪，正在寻求控制该参数的方法。

25. 用户在使用 LLaMA 模型时遇到了提示模板的问题。

```