2025-01-10 LLM News

```

title: LLM News

一：LLM 新闻

1. Phi-3.5-MoE 模型已合并到 llama.cpp 中。

2. Dir-assistant 1.2.0 版本发布，支持使用 API 嵌入模型，提高 RAG 查找质量。

3. Agent Laboratory 框架发布，利用 LLM 代理作为研究助手，降低研究成本并提升研究质量。

4. Moondream 2B 视觉语言模型发布。

5. 提示的图灵完备性被提出。

6. Qwen2.5-32B-Instruct-AWQ 模型在 RTX 4090 上运行出现内存溢出问题。

7. 用户正在寻找适用于 Qwen2.5-Coder:32B 模型的通用提示词或系统提示词。

8. LLM Pools 发布，是一个端到端的 LLM 部署环境。

9. LongTalk-CoT v0.1 数据集发布，提升模型推理能力。

10. 模型识别为竞争对手产品，如 Deepseek 模型将自己识别为 OpenAI 模型。

11. Braina AI 推出了 LocalLibrary 本地 RAG 功能和 Deep Web Search 功能。

12. rStar-Math 证明小型语言模型可以在数学推理能力上媲美甚至超越 OpenAI 的 o1 模型。

13. 传闻即将发布一个名为 Phi 4 的新模型，参数规模为 14B。

14. 用户发现了一个名为 QwQ-LCoT-7B-Instruct 的微调模型，在复杂、长上下文的编程任务上表现出色。

15. 英伟达即将发布 Project Digits，这是一款人形机器人产品。

16. TGI v3 与 vLLM 的性能对比引关注。

17. Phi-4 模型发布多个修复版本，并提供动态 4bit 量化版本。

二：实用技巧：

1. RTX 4090 48GB 版本被发现，售价 4700 美元。

2. DeepSeek API 的使用方法：对于大型文本或 PDF 文件，需要分割成块进行翻译。

3. 将音频转录成 LaTeX，利用本地模型去除口语化表达。

4. 微调模型选择参考：Phi 模型有时过于“死板”，探讨了微调后的 Llama 和 Phi 模型的对比结果。

5. 小型模型（3B）进行 Roo/Cline 的效果不佳。

6. 多 GPU 环境下 VRAM 未充分利用问题，可能需要进一步排查 GPU 设置。

7. RTX 4060 Ti 16GB 与 A4000 16GB 性能对比，差异可能不大。

8. 运行 70B 模型的硬件选择：Digits、AMD AI max、Apple M4 Max、3090/4090/5090。

9. 显卡购买建议：双 3090 和单 5090 的选择，考虑显存和游戏需求。

10. 运行大型语言模型的硬件新思路：使用 Ampere Altra 处理器。

11. O1 模型在处理非结构化数据时表现优异。

12. 使用 Ollama, Anthropic 或 DeepSeek 运行开源 AI Web Agent。

13. 构建 AI 驱动的文档搜索系统，用于搜索产品手册。

14. 双 4090 显卡在特定场景下更具性价比。

15. Mac Mini M4 Pro 上 Llama 3.3 70B 的性能：仅使用 CPU 运行，预测速度可达 6.51 tok/s。

16. Braina AI 的 Deep Web Search 使用技巧：增加模型的上下文窗口长度。

17. 在本地运行大型语言模型的显存优化：研究通过硬盘读取内存。

18. 使用 TODO 驱动开发进行开源贡献。

19. KoboldCpp API 的 Python 脚本编写指南。

20. 双 GPU 运行 LLM 的硬件考虑：PCIe 通道和电源问题。

21. Ollama 模型下载路径配置：用户尝试修改下载路径，但模型仍然下载到默认路径。

22. GPU 选择：RTX 4060 Ti 16GB 比 RTX 3060 12GB 更值得考虑。

23. 本地 LLM 运行：可以通过购买二手 RTX 3090 显卡和升级 RAM 来在本地运行中等规模的 LLM 模型。

24. 知识图谱与向量搜索结合：构建 RAG 应用改进电影搜索体验。

25. Phi-4 模型现已支持 llama.cpp。

26. Phi-4 模型微调速度提升 2 倍，VRAM 使用减少 70%。

三：其他：

1. 用户创建了一个新的 LLM 挑战，旨在测试参与者找到正确函数调用的能力。

2. 关于本地 LLM 是否可以搜索互联网的讨论。

3. Moondream 2b 的新视线检测脚本。

4. 2025 年最佳信息检索嵌入模型（标题内容为空）。

5. Phi-4 的快速测试（标题内容为空）。

6. 无线去中心化推理：探索本地网络中多台机器的算力进行无线分布式计算。

7. 开源 AI Web Agent 的更新，可以与 Ollama、Anthropic 或 DeepSeek 一起使用。

8. LLM Pools 的开源地址和公共 LLM Pool 地址。

9. 开源提示库的需求。

10. 语音对话和开源语音模型的现状。

11. 开源、基于 Python 的数据连接器的需求。

12. 金融聊天机器人的意图分类挑战。

13. 测试 LLM 能力的方法：使用井字棋或四子棋等游戏。

14. AI 代理 24/7 运行：需要一个可以提醒日历事件并定期搜索互联网信息的 AI 代理。

15. 生成式 AI 工程师面试：关于生成式 AI 工程师的面试问题和建议。

16. Twitter 上的 AI 领域专家：用户询问了在 Twitter 上值得关注的 AI 领域专家。

17. 用户正在寻找能够总结医学或生物领域研究论文的预训练模型。

18. 一个 MERN 栈开发者 (大学生) 正在寻找学习 AI 的资源和建议。

19. 有人提出开发一个名为 “TransPixar” 的新生成模型。

20. 用户寻求基于 Ollama 的翻译工具。

21. 用户对 Phi-4 模型在实际应用中的表现进行评估。

```