2025-01-13 LLM News

```

title: LLM News

一：LLM 新闻

1. Llama 3 的继任者 Llama 4 正在开发中。

2. SWE-Fixer 出现，旨在训练开源大型语言模型以有效且高效地解决 GitHub 问题。

3. QwQ-32B-Preview 模型在被限制使用自身 token 预测选择下一个词后，表现出混乱、自我分析、假设检验，甚至哲学思考等行为。

4. 中国人民大学提出 Search-o1，这是一个增强型的、具有代理能力的搜索增强型大型推理模型。

5. Radiologist copilot 的出现及其潜在应用，通过学习大量的放射学 PDF、教科书和 Radiopaedia.org 文章，旨在提供比 ChatGPT 更准确和专业的辅助诊断。

6. Webui 新工具 visual-text-generation-webui 出现，以支持 LLaMA 3.2 11b 等视觉语言模型的图像发送功能。

7. Kokoro 在 TTS 排行榜上名列第一，其模型尺寸更小，许可协议也更具优势。

8. Volo 发布，使用Kiwix创建的维基百科副本作为离线数据库,Qwen2.5:3B作为LLM,旨在通过使AI模型可以访问脱机维基百科副本来解决其幻觉问题。

9. 针对特定任务的模型选择成为讨论热点，用户正在寻求针对特定任务进行优化的模型，而不仅仅是通用的基准测试模型。

10. 非指令微调的文本模型用例探索，用户对 Llama 3.2 等非指令微调文本模型的实际应用场景表示好奇。

11. 合成数据集的创建与应用，有用户计划使用 DeepSeek 在其所在语言环境中生成合成数据集。

12. Meta 计划用 AI 工程师替代中级工程师，Mark Zuckerberg 表示，到 2025 年，Meta 可能会拥有能够编写代码的中级工程师 AI。

13. DeepSeek V3 模型持续带来惊喜。

14. 目前存在支持视觉功能的非聊天或指令微调的基础模型。

15. Daniel Han 发布了关于微软 phi-4 模型 Llama-fication 的帖子，解释了将模型转换为 Llama 格式的步骤。

16. Qwen 发布了在线版本的 Qwen Chat。

17. Gemma-2 Ifable 在故事讨论中展现出不俗的推理和理解能力。

18. 针对学术论文写作，具有 16GB VRAM 的 4060 Ti 显卡用户正在寻找合适的模型。

19. 视觉语言模型 (VLM) 领域涌现出更多选择，用户在寻找 14B 以下参数量的优秀 VLM 模型。

20. 有一家公司宣布转型为 AI 公司。

二：实用技巧：

1. Llama.cpp 分布式推理：可以通过在多个 GPU 服务器上运行 `build-rpc-cuda/bin/rpc-server`，并使用 `build-rpc/bin/llama-cli` 命令指定多个 RPC 服务器地址来实现分布式推理。

2. 利用特定领域的知识训练本地模型：通过使用特定领域的文本数据（如放射学 PDF、教科书）训练本地模型，可以显著提高模型在该领域的准确性和实用性。

3. 利用 Webui 工具增强视觉语言模型的功能：visual-text-generation-webui 填补了现有 Webui 工具在视觉语言模型图像发送功能方面的空白。

4. 利用现有显卡运行本地模型：一张 5090 显卡可以用于运行定制化的专业领域大语言模型。且2x P40可以运行qwen 2 audio模型。

5. 低配设备上运行小型 LLM：存在能够在移动设备或低配置笔记本电脑（集成 GPU + 8/12GB 内存）上运行的小型 LLM。

6. 生成特定格式的输出：可以使用 RAG 模型生成特定格式的输出，例如 JSON。

7. 大型编码项目的 AI 辅助开发技巧：使用 macOS 上的“Combine Files”应用将多个文件合并成一个文件，以绕过文件数量限制；需要仔细审查 AI 提交的代码。

8. 本地部署大语言模型的硬件配置：针对政府部门等需要处理敏感数据的情况，有用户咨询了在 2 万美元 GPU 预算下，本地部署大语言模型的最佳硬件配置方案。

9. LM Studio 使用外部存储 GGUF 模型的配置方法：需要更改 LM Studio 的配置文件，以指向正确的模型文件路径。

10. 寻找 Phi-4 的在线聊天界面：用户正在寻找能够提供 Phi-4 非量化模型在线聊天界面的平台。

11. 利用本地 AI 助手提高生产力：用户寻求能够协助制定计划、处理消息/邮件等琐碎任务的本地 AI 助手模型。

12. VLC 计划添加离线实时 AI 字幕：VLC 正在开发一项新功能，将利用 AI 技术提供离线、实时的字幕。

13. 利用 MCPAdapt 将 650+ MCP 服务器作为工具：一个名为 MCPAdapt 的开源库允许将 MCP 服务器作为工具集成到各种智能体框架中。

14. Agentless：无需智能体也能解决复杂的代码问题：Agentless 是一个非智能体框架，OpenAI 已经使用它在 SWE Bench 上用 o3 获得了高准确性。

15. 在 6GB 显存的显卡上运行本地 LLM：可以使用 GTX 1660 Ti (6GB 显存) 运行一些经过量化的较小模型。

16. 构建本地聊天机器人：使用 Ollama 服务器和 8GB 显存、32GB 内存的配置，可以尝试运行一些较小的模型。

17. 了解运行 LLM 的系统要求：有学习资料可帮助了解不同类型模型所需的显存大小、量化技术及其与内存的关系等基础知识。

18. 通过光学输出或 HDMI ARC 将电视音频连接到 PC：可以使用 Creative G8 DAC 作为中介。

19. Dell 电脑选购建议：在 2000-3000 美元的预算下，可以考虑配备第 14 代 i7 处理器、32GB 内存以及 8GB 或 12GB 显存显卡的 OptiPlex 或 Precision 系列工作站。

20. 实现 LLaMA 多模态：理论上可以将 LLaMA 13B 与附加组件结合，实现具有长期记忆的 TTS 对话。

21. 在当前阶段，推测性解码 (speculative decoding) 功能还不太可能被集成到 Ollama 中。

22. 6 台 AMD Instinct Mi60 AI 服务器可用于运行 Llama 405B 模型，并结合 vLLM 和 Open-WebUI 使用。

23. 选择合适的模型进行学术论文写作：拥有 16GB VRAM 的用户可以尝试一些针对学术写作优化的模型。

24. 在资源受限的情况下运行 VLM：寻找 14B 以下、性能优异且兼容 Ollama 的 GGUF 格式 VLM 模型。

25. 排查 RAG 问题：使用 AnythingLLM 和 Phi4 进行 RAG 时，需要检查 AnythingLLM 的配置。

三：其他：

1. 寻找最佳的本地模型以进行陪伴和闲聊：社区正在积极探讨哪些本地模型最适合作为孤独人群的陪伴。

2. API 提供商支持语法引导采样：用户寻找能够提供语法引导采样功能的 API 提供商。

3. 不同模型在不同任务上的表现差异：有用户发现 Claude sonet3.5 在结构化输出方面优于 GPT4-o。

4. 用户正在讨论是否应该升级显卡：用户正在讨论升级到 50xx 系列显卡的可行性和必要性。

5. 结合参与度指标的向量数据库应用：有用户提到了结合 LinkedIn 帖子和参与度指标的向量数据库。

6. 寻找音乐艺术家相似性模型：有用户询问是否存在用于查找相似音乐艺术家的模型。

7. 终结者视觉中的“ANALYSIS”可能是图像嵌入：用户推测电影《终结者》中机器视觉的“ANALYSIS”功能可能是指图像嵌入技术。

8. 欧盟 AI 服务器供应商：一位用户正在寻找欧盟地区提供 AI 服务器构建和销售服务的个人或公司。

9. 一些用户需要寻找支持视觉功能的非聊天或指令微调的基础模型。

10. 一位用户询问关于将微软的 phi-4 模型进行 Llama-fication 的实现细节。

11. 一位用户询问关于如何实现 LLaMA 多模态，并提出了具体的功能需求。

12. 有用户测试了 Moondream 模型在解决验证码 (CAPTCHA) 方面的能力。

13. 有用户正在寻找涵盖 LLM 深度细节的现代机器学习 (ML) 基础知识认证课程。

14. 用户在使用 AnythingLLM 进行 RAG（检索增强生成）时遇到问题。

```