2025-01-09 LLM News

```

title: LLM News

一：LLM 新闻

1. Pixtral-12B 模型在单个 3090 显卡上运行出现内存溢出问题，即使使用 unsloth bnb 4 位模型也无法成功。

2. Phi-4 模型发布修复版本，解决了分词器和聊天模板相关的 4 个 bug，并进行了 Llamafied 改造，使其可以与 Unsloth 等框架兼容。使用 Unsloth 微调，速度提升 2 倍，显存占用减少 70%，上下文长度提升 9 倍。同时发布了 GGUFs、4 位、16 位版本以及动态 4 位量化版本。

3. 用户创建了一个 LLM 挑战赛，通过微调 unsloth/Llama-3.2-1B-Instruct 模型，使其遵循特定模式，挑战目标是让 LLM 输出预设的密码。

4. 约 7 个月前发布的 AI 模型种子追踪网站 AiTracker.art 目前无法访问，具体原因不明。

5. Chipper Ollama Haystack RAG 工具箱发布 1.0 版本，提供 Web 界面、CLI 和可定制的架构，用于嵌入管道、文档分块、网络抓取和查询工作流程。

6. DEEPSEEK-V3 的标记符中发现了 800 个占位符标记以及中间填充标记。

7. 英伟达 Project DIGITS 的内存带宽预计将达到 273 GB/s。

8. Zeit：一个 2 分钟的每日 AI 简报上线，网址为[https://readzeit.com/](https://readzeit.com/)。

9. Phi 系列模型在实际应用中表现不佳，例如，当前 Ollama 版本不支持函数调用。

10. 基于 CES '25 的公告，用户正在探讨构建高性能 GPU 运算平台以运行大型模型。

11. 用户搭建了四路 P40 GPU 系统并进行了性能测试。

12. 社区对英伟达的 Project Digits 持怀疑态度。

13. Seemless M4T Unity 小模型已支持英语、西班牙语、葡萄牙语、印地语和俄语的离线翻译应用。

14. 用户们在讨论本地部署的LLM的多语言能力与GPT-3.5是否能够匹敌。

15. LLaMA 3.1 405B 与图像生成和语音的结合被用户探讨。

16. 微软发布了 Phi-4 模型，并采用 MIT 许可协议。

17. Nvidia 发布了 Llama Nemotron 和 Cosmos Nemotron 模型系列，涵盖 Nano、Super 和 Ultra 三种规模。

18. 用户 ngxson 发布了其首个微调的小型推理模型 MiniThinky 1B。

19. ARM 架构在 AI 领域仍存在挑战。

20. RTX 5090 笔记本电脑可用于本地运行中小型模型。

21. AMD Strix Point 笔记本电脑有望运行 70B 模型。

22. Qwen2.5-coder:1.5b (Q4\_K\_M) 在 Strawberry 测试中表现出色。

23. Gemini Coder VS Code 扩展集成 AI Studio。

24. 本地 LLM 在 Android 手机上的运行成为可能。

25. Anthropic 的博客被阿里巴巴 Qwen 团队的技术负责人强烈推荐作为理解智能体 (Agent) 的重要资源。

26. NVIDIA 发布了名为 Cosmos 的世界基础模型。

27. Meta 与 Apple 的合作被认为是一个潜在的强大联盟。

28. 惠普推出基于 AMD 的生成式 AI 机器，配备 128GB 统一内存。

29. 科学创意生成模型 QwQ-32B-Preview-IdeaWhiz-v1 发布。

30. DIGITS 的真正用例是 SLM 训练。

31. Llama 4 训练进度推测：预训练可能已经完成，或已进入微调阶段。

32. Cosmos-1.0-7B-Text2World 模型可根据文本提示生成视频。

二：实用技巧：

1. 在 llama.cpp 中使用 Phi-4 模型，可以使用以下命令：`./llama.cpp/llama-cli --model unsloth/phi-4-GGUF/phi-4-Q2\_K\_L.gguf --prompt '<|im\_start|>user<|im\_sep|>Provide all combinations of a 5 bit binary number.<|im\_end|><|im\_start|>assistant<|im\_sep|>' --threads 16`。

2. 对于 Phi-4 等模型，使用 Unsloth 进行微调可以显著提升训练速度并减少显存占用。

3. 如果使用的是 iOS 系统且依赖 LLM 应用（如 LLM Farm），暂时不要更新到 iOS 18。

4. 目前，在创意写作方面，用户发现较小的 Mistral Large 模型（经过 bartowski 量化）表现相对较好。

5. 用户以低价购入了 4 张 3090 显卡，计划搭建一个包含 3 张 3090 的封闭式机箱系统，且无需水冷 GPU。建议使用 Lian Li O11 Dynamic XL 机箱。

6. 在公司内部开发 LLM 原型，建议使用 AWS EC2 实例，并通过 VS Code 连接。

7. 可以通过 `ollama run phi4` 下载 Phi4，然后通过 `git clone https://github.com/EliasPereirah/OrionChat.git` 克隆 OrionChat 仓库，或直接使用 GitPages：[https://eliaspereirah.github.io/OrionChat](https://eliaspereirah.github.io/OrionChat) 进行体验。

8. 搭建低功耗的 AI 迷你工作站，建议配置包括：AMD Ryzen 9 7900 CPU、Asrock A620I Lightning WiFi 主板、96GB DDR5 内存、2TB NVMe 硬盘、Quadro RTX 2000E ADA 16GB 显卡、SilverStone Milo ML07B 机箱和 500W 电源。

9. 3090 可以运行 8b 的 Llama 模型。

10. 私人电脑的Windows系统需要支持SSH服务，这样才能通过 VS Code 连接另一台机器。

11. 利用 Athina AI Flows，用户可以通过简单的拖拽操作，在两分钟内构建 AI 工作流。

12. 可以利用本地运行的 LLM（如 Llama）来执行文件操作。

13. 使用四路 P40 搭建的 GPU 系统，运行 Qwen-2.5-Coder-32B (Q8\_0) 模型，tg128 推理速度可达 13-16 t/s，pp512 预填充速度可达 197 t/s；运行 Llama 3.1-Nemotron-70B (Q8\_0) 模型，tg128 推理速度可达 8-10 t/s，pp512 预填充速度可达 128 t/s。

14. 若需在 2025 年构建高性能 GPU 运算平台，可考虑使用 AMD Ryzen APU (Strix Halo) 或 NVIDIA 5090 (推测) 等新一代硬件。

15. 针对卫星图像处理，可以探索前沿的本地部署图像模型，以识别哥伦比亚卫星数据中的特定标识符。

16. GGUF 模型直接从文件运行到 OpenWebUI，避免 Ollama 的重复下载和存储问题。

17. 考虑将 3070 Ti 升级到二手 4060 Ti 16GB，以获得更大的显存来运行本地 LLM。

18. 使用 `Dirty-Alice-Tiny-1.1B-V2-Chatml-Q8` 模型时，通过修改代码，添加系统提示 `system\_prompt`，并使用 `<|im\_start|>user` 和 `<|im\_end|>` 标记用户输入，以及 `<|im\_start|>assistant` 标记 AI 回答，最终解决了问题。

19. 分阶段进行模型微调，首先使用闭源模型开发完整解决方案并评估基线准确率；然后构建小规模数据集，选择合适的本地 LLM 进行微调并评估效果；接着利用微调后的模型生成大规模伪数据集，进行修正后再次微调模型，并迭代测试直到达到预期效果。

20. 当 RAM 不足以容纳整个模型时，可以使用 ZRAM（Linux 内核模块）创建压缩的内存块设备作为交换空间。

21. 可以通过 `ollama run qwen2.5-coder:1.5b` 命令运行该模型，并使用 `/show info` 查看模型信息。

22. 通过安装 Gemini Coder VS Code 扩展，可以方便地将代码上下文和自定义指令发送到 AI Studio。

23. Aider 被认为是在中大型代码库（超过 10k 行代码）中进行 AI 编码的最佳开源工具之一。

24. 可以使用 `ollama run 6cf/QwQ-32B-Preview-IdeaWhiz-v1` 命令本地运行 QwQ-32B-Preview-IdeaWhiz-v1 模型。

25. 惠普的 128GB 内存机器可以轻松运行 70B 模型（q8 量化）。

26. 对于仅需运行推理的用户，使用 2 个 4090 显卡可能更合适。

27. 用户分享了在双 3090 显卡上运行 `hugging-quants/Meta-Llama-3.1-70B-Instruct-AWQ-INT4` 模型的经验。

28. 针对 4070 Super 显卡，可以考虑 7B 或量化后性能更好的模型用于编码、PDF/图像扫描。

29. Llama 3.1 8B Instruct Q4\_K\_M 是目前测试过的 <9B 参数维基模型中最佳选择。

三：其他：

1. 用户正在寻找能够将大型文档分割成块并利用 Ollama 后端进行翻译，然后重新组装成完整文件的自托管前端工具。

2. ROG Flow Z13 2025 笔记本电脑搭载 Ryzen AI Max+ 395 处理器和高达 128GB LPDDR5X 内存，采用 8000MHz 四通道内存。

3. 用户分享了他们在 Groq 上使用模型的体验。

4. 有帖子提到了一个由 Llama 3.2, 3.3 & OS-Atlas 驱动的开源计算机使用代理。

5. 用户认为 AI Agent 将成为下一个风口，并计划构建一个能够执行内容创作任务的 Agent。

6. LLM 使用 Emoji 表情可能表示性能不佳。

7. 目前社区暂无针对新手入门的 LocalLLM 指导。

8. 用户正在寻找适用于移动设备并在 ChatterUI 上运行的 3B NSFW GGUF 模型。

9. 用户正在使用 AI 进行市场调研实验。

10. 一家医疗保健领域的软件/分析开发公司计划构建自己的服务器来托管大型模型。

11. 用户正在寻找适合儿童故事创作的大型语言模型。

12. 有用户反馈 Deepseek 提供了诈骗网站链接。

13. 用户用学习方式类比 AI 训练方式。

14. 一位用户向社区征求关于新 AI 架构设计的建议。

15. 一位用户正在寻求关于使用andrej karpathys的纳米GPT预训练模型进行指令微调的具体指导和资源。

16. Amurex 是一款自托管的会议记录工具，支持 Google Meet。

17. Razer Blade 16 笔记本电脑预计将配备 RTX 5090 (24GB GDDR7) 和 64 GB LPDDR5X-8000 RAM。

18. 有一个网站每天更新数百个 AI 职位信息。

19. 有开发者正在寻找构建 "agent builder" 框架的实际应用案例。

20. 有用户在探讨如何运行大于物理 RAM 容量的模型。

21. 用户询问是否有比 Dell 和 Lenovo 更便宜的 GPU 服务器供应商。

22. 介绍了如何跟踪 LLM、图像和视频生成等领域的最新进展。

23. 提到了对语音领域 AI 模型的需求。

24. 表达了对构建具有记忆能力的聊天机器人的兴趣。

25. 提到了在没有互联网的村庄中使用离线语音助手的需求。

26. 社区正在讨论编程是否已经超越了工作技能的范畴，而成为一种文化生活方式。

27. 有人在收集合成音频数据。

28. 测试表明，Qwen 2.5 Coder 32B 在中大型代码库中的编码能力远不及 DeepSeek 3。

29. 有人建议使用 CrewAI + Aider 的组合来进行企业级的 AI 编码工作。

30. 由于 Llama 3.2 不支持结构化输出，因此实现代理工作流程存在挑战。

31. 用户寻求托管 100-200GB 数据并能按需连接 4090 或 A100 显卡进行计算的方案。

32. 用户尝试在 Windows Server 2019 上安装 3090 显卡驱动失败，寻求解决方案。

33. 用户考虑在资源有限的服务器上使用小型 RAG 解决方案增强或替换现有的 Streamlit Web UI。

```