2025-01-08 LLM News

```

title: LLM News

一：LLM 新闻

1. Llama 4 早在 10 月 28 日就已经开始训练，使用了 10 万块 H100 GPU，计算量是 Llama 3 的 10 倍。根据估算，预训练阶段可能已经完成，目前可能处于微调阶段。

2. DeepSeek 仅用 540 万美元的预算就训练出了一个高性能模型，这让人们对 Meta 使用类似方法训练 Llama 4 充满期待，尤其是考虑到 Llama 4 可能是原生多模态模型。

3. Cosmos-1.0-7B-Text2World 发布，此模型可以根据文本提示生成视频。使用单个 3090 显卡即可运行。

4. Autopod 开源工具发布，利用 n8n、OpenAI 和 Raindrop.io 将用户的“稍后阅读”链接转化为播客，实现自动化阅读列表到音频的转换。该工具成本效益高（20 分钟播客成本低于 0.20 美元），且支持高度自定义。

5. Notate 开源 AI 研究助手发布，支持本地大型语言模型。

6. 用户提出通过增加计算资源（例如使用高端 GPU 运行更长时间）来提升模型输出质量的设想。

7. 用户寻求针对数学问题的最佳基准测试/排行榜，并期待出现类似 AlphaProof 的开源工具。

8. 用户尝试使用视觉语言模型从图表中提取数据，但仍面临数据流缺失、模型幻觉和数据流不正确等问题。

9. 英伟达的 Digits 在 AI 推理方面表现优于苹果的 M4 芯片。

10. 嵌入模型缺乏开源或通用的后训练量化技术。

11. 用户正在权衡购买英伟达 5090 显卡还是等待 Digits。

12. 有关于大型语言模型评估的电子书章节可供参考。

13. 存在无审查机制的大型语言模型可供下载。

14. CrewAI 与 NVIDIA 合作，或将成为企业级 AI Agent 框架的新标准。

15. Phi-4 模型在处理多轮 RAG 问题时，展现出卓越的上下文重构能力。

16. 硬件性能持续提升，为 AI 发展奠定基础。

17. Nvidia 推出“个人 AI 超级计算机” Project Digits。

18. DeepSeek-V3 因其低成本和高性能（超越 GPT-4o 和 Claude3.5）而备受关注，但其服务条款中存在一些潜在风险。

19. 英伟达发布 Cosmos 世界模型平台，旨在加速机器人和自动驾驶领域的物理 AI 开发。

20. NVIDIA 发布了新的 open nemotron 模型。

21. NVIDIA 宣布推出名为 Digits 的个人 AI 超级计算机，售价 3000 美元。

22. NVIDIA 发布了 Grace Blackwell，并将其应用于每个人的桌面和每个 AI 开发人员的指尖。

23. MSI 在 2025 年的 CES 上发布了 Cubi NUC AI 系列迷你电脑。

二：实用技巧：

1. 在单张 3090 显卡上运行 Cosmos-1.0-7B-Text2World：可以使用以下命令生成视频：`PYTHONPATH=$(pwd) python cosmos1/models/diffusion/inference/text2world.py --checkpoint\_dir /workspace/checkpoints --diffusion\_transformer\_dir Cosmos-1.0-Diffusion-7B-Text2World --prompt "water drop hitting the floor" --seed 547312549 --video\_save\_name Cosmos-1.0-Diffusion-7B-Text2World\_memory\_efficient --offload\_tokenizer --offload\_diffusion\_transformer --offload\_text\_encoder\_model --offload\_prompt\_upsampler --offload\_guardrail\_models`

2. 在双 3090 显卡上使用 vLLM Docker 运行 4 位量化的 70B Llama 3.3 模型。

3. 利用 Autopod 工具，可以将在线文章等内容自动转换为音频格式。

4. 增加计算资源（例如使用更强大的 GPU 并延长运行时间）有可能提升模型输出的质量和准确性。

5. 针对 8 个 GPU 的配置，用户正在寻求 6U 机架式机箱的建议。

6. 针对提示工程师，可以开发类似 HackerRank 或 LeetCode 的平台来评估其技能。

7. 寻找适用于 Firefox 的本地浏览器智能代理。

8. 可以通过使用两块 GPU（例如 RTX 3090 和 3090TI）并在同一台电脑上运行来合并显存，从而运行更大的 LLM。

9. 利用 Ollama 和前端工具进行代码仓库的检索增强 (RAG)。

10. 使用提示词工程引导 LLM 进行自我反思和改进。具体提示词请参考: [提示词链接](https://gist.github.com/Maharshi-Pandya/4aeccbe1dbaa7f89c182bd65d2764203)。

11. 利用“沉思式推理” (Contemplative reasoning) 提升 LLM 回答质量。

12. 通过修改参数 `n\_ct\_per\_seq` 来调整上下文长度。

13. 利用 [continue.dev](http://continue.dev) 与本地模型交互。

14. 5090 显卡以 1999 美元的价格成为性价比极高的选择。

15. 相较于单个 5090，购买两个 5080 可能是一个更好的选择。

16. 用户可以尝试使用 5090 显卡的 FP4 精度进行模型推理。

17. 如果用户需要让 LLM 遵守特定的语法规则，可以尝试对模型进行微调。

三：其他：

1. 有用户正在探索在无 GPU 的服务器上，使用小型 RAG 解决方案来增强或替代现有的文档界面，以提高开发效率。

2. 一位用户正在寻找一个小于 9B 参数、至少 Q4\\_K\\_M 量化、擅长作为离线维基百科、能解决填字游戏、并且在葡萄牙语表现良好的模型。

3. 用户询问了 Ray-Ban Meta 眼镜的技术原理，包括数据处理流程和 Llama 模型的更新计划。

4. 用户分享了自己使用 ChatGPT 的经验，并计划转向使用 OpenAI/Anthropic/DeepSeek 的 API 积分进行临时查询，以及使用 Cursor 进行代码编写。

5. 用户猜测 Digits 可能是 Jetson AGX Thor 的改版。

6. 在 2 万美元的预算下，用户寻求针对 Llama 3.3 70b 或 405b 或 Deepseek V3 模型，能够实现最高性价比的硬件配置方案。

7. 目前社区对嵌入模型量化技术的需求较低，这可能是其发展缓慢的原因之一。

8. 对提示工程师的需求正在增加，市场上出现了相关的课程和职位空缺。

9. 关于 AI 立法在美国的推进情况目前尚不明朗。

10. 有用户正在比较 M4 Max 和 5090/4090 运行 AI 的性能。

11. 用户正在探索使用 Ollama 结合 OpenWebUI 和 1filellm 对本地代码仓库进行检索增强，并寻求更高效、更少手动索引文件的方式。

12. 用户展示了 ChatGPT 根据图像生成贝塞尔曲线方程的能力，并询问是否可以使用开源模型和工具实现类似的功能。

13. 用户对于AI行业表达的准确性表示关注。

14. Together AI 提供 DeepSeek V3 API，但价格较高。

15. 有用户询问是否有可靠的 LLM/AI/ML 模型可以通过录音来评估英语口语水平。

16. 有用户提出了一个基于混合专家模型、GUI 任务数据集、分层记忆系统和多合成语言的 AI 智能体概念。

17. 有用户询问 Deepseek v3 性能优异且价格低廉的原因。

18. 有用户询问如何通过提示或其他方式使 Qwen SmallThinker 模型不受审查。

19. 有用户对英伟达 5090 显卡配备 32GB VRAM 而 Project Digits 桌面电脑配备 128GB VRAM 表示困惑。

20. NVIDIA 的统一内存可能不是传统的 HBM，但具体类型尚未明确。

21. 有用户质疑 NVIDIA 是否会推出售价 3000 美元以上、配备 128GB 显存的 5070 显卡。

22. 对于追求高性能计算的用户来说，NVIDIA 宣布的个人 AI 超级计算机 Digits 可能是一个 Mac Mini 的有力竞争者。

```