2025-01-05 LLM News

```

title: LLM News

一：LLM 新闻

1. Mistral Nemo Instruct Chat模板的微调：有用户对Mistral Nemo Instruct聊天模板在微调中的使用方式提出了疑问。该模板在推理时，仅在最后一个用户-助手交互且不存在最后一个助手消息时才应用系统消息。然而，在微调过程中，系统消息似乎总会被排除在外。用户因此疑问：为了准备用于Nemo模板微调的数据，是否应该将系统消息中的指令放入用户消息中。此外，用户还想了解为什么此模板不选择像其他模板一样始终将系统消息放在最前面。

2. 强大的服务器用于运行LLM：讨论中提到了一个名为“themachine”的服务器配置，该配置包含12个NVIDIA 3090显卡，表明一些用户正在探索使用大型服务器来运行和训练大型语言模型。

3. Llama3 推理引擎的 CUDA C 实现：一位用户基于 `llama.cpp` 和 `ollama` 等工具的灵感，使用 CUDA C 构建了一个针对 Llama3 8B 模型的推理引擎。该引擎完全使用 CUDA 运行时 API 和 `cuda\_fp16`，在 FP16 精度下运行，需要约 17-18GB 的显存。此实现未采用 `cuBLAS` 等抽象库，旨在更深入地了解底层 CUDA 编程。

4. DeepSeek-V3 支持合并入 llama.cpp：`llama.cpp` 项目已合并对 DeepSeek-V3 模型的支持。一位用户更新了他们 Hugging Face 仓库中的量化模型，方便用户测试。其中 Q4\\_K\\_M 量化版本表现良好，在 MMLU-Pro 计算机科学测试中取得了与 API 性能相当的结果。

5. 本地LLM的应用：有用户询问是否有人将工作代码库放入本地LLM进行处理，并希望得到使用反馈以及具体使用了哪些本地LLM。

6. LLM的聊天记录管理：有用户在构建基于Groq API的聊天应用时遇到了聊天记录存储和恢复的问题。他尝试使用不同的方法，包括Langchain的记忆功能、嵌入技术（如Pinecone、FIASS）以及简单的JSON文件存储，但遇到了理解困难和性能问题（如嵌入耗时）。

7. LLM层的功能识别：有用户希望了解如何识别LLM中每一层处理的特征，以便进行模型合并等操作。他寻求可以追踪信息在模型层之间传播的工具或方法。

8. 内存层扩展（Memory Layers at Scale）：一篇新论文探讨了内存层的实际应用。内存层使用可训练的键值查找机制，在不增加计算量（FLOPs）的情况下为模型添加额外参数。研究表明，配备改进内存层的语言模型在下游任务中的表现优于计算量多两倍的密集模型。

9. RAPTOR 用于 RAG（检索增强生成）：一篇关于RAPTOR的论文引起了一些关注。RAPTOR的核心思想是将嵌入聚类和LLM总结相结合，构建文档的语义树结构，用于检索任务。

10. DeepSeek V3 的卓越表现：Reddit 社区的讨论集中在 DeepSeek V3 模型在特定领域的优异表现上。有人通过实际案例证明了 DeepSeek V3 在医学诊断、数学问题和一些专业知识方面的卓越能力。

11. ScreenSpot-Pro 基准测试的推出：一项名为 ScreenSpot-Pro 的新基准测试被推出，旨在推动多模态大型语言模型 (MLLMs) 在高分辨率专业 GUI 代理和计算机使用环境中的应用。

12. Grok 2 可能开源：根据 Elon Musk 在 X（原 Twitter）上的帖子，Grok 2 模型可能即将开源。

13. iOS本地运行LLM的应用体验问题：用户在尝试使用iOS本地应用程序（如LLM Farm）运行下载的GGUF格式的LLM时遇到问题。

14. 一种名为CAG的新型AI技术：有评论提到一种名为“CAG”的技术，并指出它可能会带来变革性影响。

二：实用技巧：

1. 为LLM聊天机器人提供上下文：有用户尝试构建一个简单的聊天机器人，并通过提供一些上下文信息让其学习特定的提问和回答技巧。然而，该机器人似乎过于依赖提供的上下文，以至于只专注于讨论上下文本身，而不是开放式地回答问题。

2. 利用硬件进行LLM的成本效益考量：用户正在考虑如何利用现有的硬件资源或以更经济高效的方式构建硬件，以运行Llama 3.3 70B模型，以替代使用ChatGPT和Claude等服务。

3. 本地 LLM 检测：有评论指出，某些本地 LLM 模型（如 Mistral Small 22b）可以规避诸如 ZeroGPT 等检测工具，而 ChatGPT 和 Claude 则容易被检测到。

4. 小型 LLM 翻译：有用户在寻找 3B 参数或更小的模型进行英西翻译，特别针对推文和博客类短文本。

5. M2 Macbook 上 Continue.dev 和 oLLaMA 的推荐模型：有用户在寻找可以在 M2 Max MacBook 上运行的，速度和输出质量都较好的模型，用于在公司内部部署 `continue.dev` 扩展，并要求使用 `oLLaMA` 进行本地部署。

6. GPU扩展方案：有用户咨询如何在消费级硬件上安装第三张RTX 3090显卡。

7. 图形文本识别：有用户需要从新闻图片中提取叠加的图形文本。

8. LLM推理批处理（Batched inference）：一位用户希望在 Vega 56 (8GB) 显卡上使用小型 LLM（<3B）实现高吞吐量。他们了解到批量推理可能有效，因此询问是否可以在 LMStudio 中使用批量推理。

9. Mini PC 用于本地 LLM：一位用户正在寻找一台小型的迷你 PC，专门用于本地 LLM。

10. 低端 PC 上运行本地模型：一位用户寻求在低配置 PC 上运行本地大型语言模型（LLM）的方法。

11. 移动端语音聊天替代方案：另一位用户在寻找 OpenAI 语音聊天在移动端的替代方案，但更侧重于私有/自有基础设施的本地运行。

12. Llama-cpp-python 的 GPU 使用：有用户在使用 Llama-cpp-python 时遇到了 GPU 使用问题，虽然在命令行中可以使用 GPU 运行模型，但在 Python 脚本中却无法使用 GPU 进行文本处理。

13. LLM参数化的重要性：用户意识到，要使LLM正常工作，需要了解如何进行参数化，例如选择正确的提示模板、处理BOS/EOS/特殊标记等。

14. GGUF文件的局限性：用户提出的问题，比如GGUF文件是否应该包含运行模型所需的所有信息，引发了对GGUF文件格式的讨论，及其在实际应用中的局限性。

三：其他：

1. LLM开发的主要瓶颈：用户讨论了LLM开发未来面临的最大障碍。普遍认为，数据可用性和计算成本是主要的限制因素。

2. 浏览器在应用中的使用：有一条评论简单地提及了“浏览器在应用中的使用”，但没有提供更多细节。

3. NVIDIA 5080 显卡价格：有消息称，在西班牙，NVIDIA 5080 显卡售价为 1699.95 欧元（含税），换算成美元约为 1384 美元。

4. 大型 GPU 集群：有用户询问使用多个 GPU 构建大型集群运行大型 LLM 的可行性，以及模型大小增长对推理速度的影响。

5. Llama3.2-vision 的视频帧分析：评论中提到有人尝试使用 Llama3.2-vision 进行视频逐帧分析，但没有提供更多细节。

6. 显卡散热器捐赠请求：一位用户在欧盟地区寻找旧的 980Ti FE 或 Titan X 显卡上拆下来的散热器或损坏的显卡。

7. Memoir+ 在 RunPod 上的问题：一位用户在 RunPod 上使用 Ooba 运行 Memoir+ 时遇到问题。

8. 新的谋杀之谜数据集：一位用户创建了一种新的谋杀之谜推理谜题，并开源。

9. 混合专家模型的影响：有用户询问关于混合专家模型和参数数量相同的密集模型在输出质量方面的性能差异的论文。

10. 关于 DeepSeek 模型可信度的质疑：有帖子质疑 DeepSeek 的模型是否过于夸大，暗示 DeepSeek 可能存在欺诈行为。

11. LLM 如何排序列表：有用户对 LLM 如何排序导入列表的机制感到好奇。

12. 对AGI的期望：有用户提出了自己对“通用人工智能（AGI）”的期望，认为一个真正的AGI应该能够独立进行深入的互联网研究。

13. 当前LLM的不足：用户对现有LLM在互联网研究方面的能力表示失望，认为它们过于依赖用户提供的关键词和线索，无法独立深入研究并提出关键信息。

```