2025-1-4 LLM News

一：LLM 新闻

DeepSeek-V3 GGUF 量化版本发布： DeepSeek-V3 模型的 GGUF 量化版本已上传至 Hugging Face，使得该模型可以在资源受限的设备上运行，降低了使用门槛。

清华大学团队提出 PRIME 和 Eurus-2 模型： 清华大学团队提出了 PRIME (Process Reinforcement through Implicit Rewards) 方法，并基于此训练出了 Eurus-2 模型，该模型使用仅为 Qwen2.5-Math-Instruct 十分之一的数据，却超越了后者在数学推理能力上的表现。

关于 o1、o1 preview、o1 pro 和 GPT-4o 的讨论： o1、o1 preview 和 o1 pro 实际上是相同大小的模型。o1 比 GPT-4o 更贵的原因在于序列长度的 KV 缓存开销。o1 pro 和 o1 是同一模型，只是在推理时进行了调整。

模型大小随上下文长度增加而增大： 使用 Ollama 时，不同模型的大小会随着上下文长度的增加而不同比例地增长，这可能是由于嵌入长度的不同导致的。

Deepseek V3 由 Fireworks 托管： Fireworks 宣布托管 Deepseek V3 模型，支持 128k 的完整上下文长度，但目前价格较高。

SmallThinker-3B-Preview 模型 WebGPU 版本： SmallThinker-3B-Preview 模型被重新编译为 WebGPU 格式，可以直接在浏览器中使用。

关于 1.58bit 模型“革命”的讨论： 关于 1.58bit 模型的“革命”似乎已经沉寂，最近很少听到相关信息。

基于 Ollama 的自改进翻译器： 有人使用 Ollama 构建了一个上下文感知的自改进翻译器，旨在解决传统翻译工具在处理创意媒体项目时丢失上下文和语气的问题。

LLM + 编程环境整合的挑战： LLM 在选择合适的工具方面表现不佳，例如，在实际中 shell 脚本或静态类型的语言可能更合适时，它们倾向于使用 Python。

寻找专门用于函数调用的模型： 有人正在寻找专门用于函数调用的小型模型，这些模型不包含文化知识。

eBay 上 4090 显卡缺少核心和 VRAM 的现象： eBay 上出现了大量缺少核心和 VRAM 的 4090 显卡在售。

Moshi 和 Hertz-dev 模型为何没有引起广泛关注： Moshi 和 Hertz-dev 这两个语音到语音模型没有引起广泛关注。

RAG (检索增强生成) 的局限性： RAG 系统在处理包含表格的文档并计数时，往往会给出错误答案，并从其他地方提取数据。

LLM 对列表的排序机制： LLM 如何排序导入语句，是使用近似方法还是外部脚本来完成精确排序。

本地 LLM 应用的挑战： 在 iOS 设备上使用 LLM Farm 应用运行本地 LLM 时，下载的 GGUF 文件会产生乱码响应。

本地模型应用： 用户们正在积极探索本地模型在各种任务中的应用，包括图像内容分类、文本推理和任务自动化。

推理能力： 当前本地模型在复杂的推理任务上表现不佳。

工具调用（Tool Calling）： 较小的本地模型在处理简单任务时（如识别 UUID）仍然存在问题。

XML 输出问题： 一些用户在使用 LLM 输出 XML 时遇到了丢失最后一个闭合标签的问题。

LoRA 微调 Moondream2 的挑战: 使用 LoRA 技术微调 Moondream2 模型时，遇到了损失曲线异常的问题。

上下文聚合图（CAG）技术的兴起: “上下文聚合图”（Context Aggregation Graph, CAG）技术可能在未来大模型优化和移动设备上的部署方面发挥重要作用。

Llama 2 7B 模型的增强: Llama 2 7B 模型添加了思维链（Chain-of-Thought）推理和网络搜索功能。

Llama 3 2 3B SpinQuant GGUF 版本: Llama 3 2 3B SpinQuant 模型的 GGUF 版本已发布。

关于推理和微调的问题： 关于运行完全量化 Llama 70B 模型进行推理所需 A100 80 GPU 数量的问题，以及微调该模型的相关想法。

漫画/漫画理解的视觉模型： 用户正在寻找能够理解漫画/漫画并描述所有对话和动作的本地视觉模型。

模型量化与故障： Mistral Large 2411 模型的量化版本（Q5/Q6 gguf）在特定硬件上出现分段错误。

多模态交互探索： 利用多模态模型结合摄像头实时画面，让 AI 模拟在“另一边”可能发生的事情，并通过图像进行互动。

模型融合与 MoE： 使用 Mergoo 工具将三个 Qwen 2.5 3B 模型组合成一个混合专家模型（MoE）时遇到问题。

模型基本知识： 用户在寻找一个拥有大学知识背景但偏见较少的“基本”语言模型。

模型审查限制： 用户在试图使用 Claude 聊天机器人获取恐怖场景写作建议时，因为审查限制而受阻。

终端交互工具： Twoie 终端用户界面（TUI）工具允许用户通过命令行访问来自多个提供商的模型。

量子计算影响： 量子计算可能会如何改变大型语言模型。

二：实用技巧

使用本地 LLM 作为生存知识库： 将本地大型语言模型作为生存知识库，根据用户描述的特定情况和可用的工具，提供即时建议。

多 GPU 运行问题： 使用 LM Studio 尝试运行模型时，添加第二个相同型号的显卡后，速度反而下降。

树莓派 5 上运行 LLM： 在树莓派 5 上使用 Ollama、Hailo AI Hat、Python 和 Visual Studio Code 运行大型语言模型。

为角色创作寻找最佳 LLM： 寻找最适合角色创作的 LLM 模型。

硬件选择与本地 LLM 使用： 为旧服务器升级 GPU，用于编码和本地运行 LLM。

GPU 选择： 选择合适的显存大小和计算能力，用于编码和运行本地 LLM。

RAG 系统优化： 针对 RAG 系统在处理表格数据时出现的问题，用户可能需要考虑更复杂的预处理步骤。

本地 LLM 模型的使用： 了解模型所需的提示模板、特殊标记以及其他配置信息，以确保模型正常工作。

微调决策框架: 分享一个用于判断是否需要微调 LLM 的简单框架。

多 LLM 对话： 探索如何让多个 LLM 互相交流。

本地个人助理： 尝试构建一个本地的个人助理，能进行语音交互。

本地模型组合应用: 利用 Ollama 服务结合多个本地模型构建实用工具。

截图管理工具的开发经验分享： 分享截图管理工具的技术栈和实现思路。

RAG 流水线中并行处理的方案： 使用 Ollama、LlamaIndex、Qdrant DB 和 Llama3.2 等技术构建 RAG 流水线，并处理大量 PDF 文档的并行处理。

购买二手 GPU 的注意事项： 购买二手 GPU 时，需要注意很多显卡都被用于挖矿，建议在购买前进行测试。

本地文章/论文写作工具： 寻找本地文章写作工具，能够像 ChatGPT 一样将文章反馈给 API 并利用新的提示进行修改。

本地运行 LLM： Mac 用户需要考虑硬件配置对模型运行效率的影响，并注意模型量化版本可能存在兼容性问题。

合规性 AI 解决方案： 银行合规官可以利用 AI 系统分析法规文件，找出潜在的违规行为，并生成简洁的报告。

终端工具： Twoie 这类工具对于需要频繁使用 LLM 的开发者很有帮助。

使用旧版库： 解决特定问题有时可能需要回溯使用旧版库。

三：其他

GGUF 在 2024 年的发展： “2024 是 GGUF 起飞的一年”。

请求性能测试: 有用户希望使用配备 512GB DDR4 内存和单个 3090 显卡的计算机进行 DeepSeek-V3 GGUF 的推理速度测试。

隐私问题： 用户在讨论 Fireworks 托管 Deepseek V3 模型时，对其隐私政策提出质疑。

关于 AGI 的期望： 用户对 AGI 的期望，认为 AGI 应该能够进行深度互联网研究，从模糊的问题开始，逐步学习，收集相关信息，最终提供决策方案。

Speculative Decoding： 用户询问是否只有 Fireworks AI 和 OpenAI 通过 API 提供推测解码。

服务器显卡选择： 用户正在考虑使用 AMD 显卡构建一个更便宜的服务器。

本地编码 LLM 需求： 用户询问本地运行的、适用于 VSCode 的优秀 LLM。

寻找性价比高的本地开源 LLM： 用户正在寻找最便宜或开源的大语言模型。

AI 行业需要像研究实验室一样运作： 随着 AI 研究和产品开发周期缩短，拥有强大内部研究能力的行业将在长期内更具竞争力。

低成本聊天机器人的需求： 用户寻求为大学活动构建经济实惠的聊天机器人解决方案。

社区对技术进展的兴奋： 社区对 AI 技术进步的积极态度和兴奋之情。

LLM 的越狱： 用户正在研究 LLM 越狱的自动化。

选择备份人类知识的方式： 用户在讨论备份人类知识时，考虑选择维基百科数据库还是开源 AI 模型。

构建 LLM 专用电脑： 用户想构建一台专门用于运行本地大型语言模型的电脑。

硬件选择： 用户在考虑为本地运行大型语言模型购买 Mac 电脑。

二手 GPU 服务器： 有人正在考虑购买二手 8x V100 DGX 服务器用于模型推理和微调。

多模态交互： 多模态模型与摄像头实时画面的结合，可能开启新的交互方式。

模型审查与限制： 模型审查可能会限制其在某些领域的应用，用户正在寻找替代方法。

模型选择： 不同用户对模型的偏好不同。