会议摘要

微信群里的学习与交流

主要讨论了多智能体的工作，包括微软的A和Open AI等主流模型。同时，介绍了大语言模型的chatbot如何工作，以及如何通过训练对话数据来提高模型的效果。此外，还探讨了语言模型中的token和标记问题，以及如何在标记中区分用户和系统角色。

智能对话系统的模型与框架

主要讨论了语言模型和聊天机器人的工作原理，以及多智能体框架的应用。语言模型类似于从文本到上下文生成后续的文本，而聊天机器人则是根据上下文生成相应的回答。多智能体框架包括GPT、GBT等，它们在开发过程中发现了一些问题，因此开发了V0.4版本，引入了team的概念，实现了类似于群聊的功能。

多智能体编程与部署的探索

主要讨论了异步编程模式在多智能体部署中的应用，以及Open AI的Swan框架。Swan将agents视为同等地位，允许agent选择调用工具或让出发言权给其他agent。此外，还介绍了本地部署语言模型的选择和思考，包括欧拉玛、lava CPP和VLM等方案。最后，讨论了量化模型与全量模型的区别，指出量化模型在内存消耗和生成速度上有所优势，但相关性会有一定损失，一般在99%左右。

基座模型选择与量化影响分析

讨论了基座模型选择的问题，提到了量化模型和反馈不好可能的原因。同时，提到了阿里的千问和LAMA模型，强调了LAMA模型提供了详细的训练技术文档，值得大家去学习。此外，还提到了多智能体的问题，以及UC团队的大猩猩项目，以及deep stick模型的写代码能力。

国产模型与千万团队的体验分享

主要讲述了讲者使用过的两个国产模型，以及在使用过程中遇到的问题。讲者提到，虽然这两个模型是国产的，但使用频率不高。此外，讲者在小红书上找到了千万团队，并询问实习生招聘事宜，但对方对问题回答非常固执，不愿意提供详细信息。最后，讲者提到了拉马3.2模型，虽然官方说不支持中文，但实际上使用中文表现不错。此外，讲者还分享了自己实现run的过程，包括如何从网上获取数据、查重、抓取内容、生成文档等。

代码实现与优化分享

主要讲述了一个简单的代码实现，使用本地存储和本地目录进行持久通知。定义了agent的角色，包括发言人、话筒等，并通过传递function的方式来实现。同时，提到了腾讯推出的a四框架，它基于函数角度构建系统，每个方面都可以单独拿出来做测试和优化。最后，讨论了如何界定agent的边界，以及如何在对话中维护信息。

优化用户代理与语言模型的交互

讨论了用户代理和检索代理之间的对话，以及如何通过调用搜索引擎的方式来提高模型的稳定性。同时，提到了contextual retrieval和上下文过滤等技术。此外，还讨论了将常见功能集成到技术模型中的做法，以及如何通过模块化来提升系统性能。最后，提醒大家关注两个课程页面，一个是针对校内课程的，另一个是针对网上慕课公开课的，并介绍了证书的获取方式。

论文写作与软件开发分享会

讨论了如何撰写文章、做quiz和写post，以及如何使用CBD生成内容。同时，提到了软件开发领域的agent开发是一个热门话题。会议还介绍了新成员李普，他来自天津大学，主要研究方向是联盟学习。最后，提醒大家有任何问题可以在群里沟通。

会议待办

暂无会议待办