**OpenLLM Talk 010**

# 序章

注意事项：出于隐私保护和数据安全的考量，建议尽量不要在 talk 过程中涉及到自己的单位信息、自己的隐私信息、违反法律和道德的信息以及其他引起争议的内容，请保护好自己的马甲哈哈。

# 背景介绍

【缘起】：OpenLLM Talk 这个事情起源于 20230603 OpenLLM 交流群中大家的一场讨论，本着心动不如行动的想法，我们花了一点时间来将其落地，希望可以为大家提供一个 LLM/NLP 领域的**交流平台**。——**我们或许不够 AI，但尽量足够 Open；我们也不知道能走多远，但尽量比自己想的更远。**

【结构】：整体上分成本周新闻、本周推荐、本周经典（可选）、本周实践（可选）、free talk 等版块，建议后续最好采用每个版块每期由 1-2 人认领+多人参与贡献+自由讨论的形式。

# 本期记录

【编号】：OpenLLM Talk 010 (三位数是希望 LLM 的热度+我们的热情+读者的热情可以支撑我们做到三位数)

【时间】：20230826 晚上九点（一般每周六晚上九点，节假日顺延）

【本期提要】：AutoGPTQ；code llama；多轮对话；ceval；SFT 基座选择；ReST；Ilya Sutskever 的新 talk；

【本期贡献者】- 排名不分先后：

【主持人】：hope（后续每期由大家自行认领）

【编辑】：羡鱼（最好由主持人兼任）

【版块负责人】： （后续每期由大家自行认领）

【具体内容贡献者】：请查看具体内容后面的署名，比如问题、回答和观点的来源

【talk 视频】：

## 注意事项

【talk 模板】：https://zhuanlan.zhihu.com/p/640522290；可参考模板进行贡献

【小要求】：主持人及版块负责人认领之后尽量准时参加，其余同学可自行选择是否参与；

# 本周新闻

【本周新闻】：LLM/AI news，包括但不限于学术、项目、工业界新闻和进展；多人认领或者直接在此添加，由 **1-2 人认领并汇总**；建议大家都参与进来，相互补充，尽量**减少信息冗余和缺漏**；共~10 分钟；

【贡献者】：

【建议区】：可以考虑 GitHub 的讨论区，看个人习惯；论文可以写个摘要；

## 学术

注：论文+重点

## 项目

# 使用 AutoGPTQ 和 transformers 让大语言模型更轻量化

https://mp.weixin.qq.com/s/ualxZFpcVTsKE_uA-V37bQ

# 逼近 GPT-4，AI 编程要革命！Meta 开源史上最强代码工具 Code Llama

https://mp.weixin.qq.com/s/VkhClhJRKLnDjE1J0GBHOQ

# DeepMind 新研究：ReST 让大模型与人类偏好对齐，比在线 RLHF 更有效

https://zhuanlan.zhihu.com/p/651583621

# 关于 Ilya Sutskever 的新 talk 的笔记

https://zhuanlan.zhihu.com/p/651702408

## 工业

## 本周推荐

【本周推荐】：本周重点内容推荐和介绍，模型、开源项目、好的资料或课程，建议 1-3 项；共 15 分钟；

【贡献者】：

【提名区】：

【建议区】：

【本期主题】：

## 资料

## 模型

## 项目

## 杂项

## refs:

## 本周经典-optional

【本周经典】：NLP/LLM 领域的经典话题探讨；~15 分钟；

【贡献者】：

【提名区】：量化

【本周主题】：

# 本周实践-optional

【本周实践】：NLP/LLM 领域实践经验分享，可以分享自己的实践经验或者他人的实践经验，后面群里也会组织一些实践内容；~15 分钟；

【贡献者】：

【提名区】：

【建议区】：coding 搞起来；后续拉个 read_code/paper 分支，LLM 精读、注释；专门建一个**数据专题**；

# Free Talk

【Free Talk】自由提问，自由讨论；在文档里提问或者在群里提问，建议尽量在此汇总；如果群里已经有比较好的讨论结果，也可以将讨论结果搬运过来；时间不限；

【贡献者】：羡鱼（编辑）+OpenLLM 群友

1. 多轮对话的效果不好如何改善

可能是数据的问题，或者是模型的问题；可以通过在 llama 70b chat 上进行测试，看一下效果，判断是数据还是模型的问题

2. 虚拟人
3. 专门领域是否只需要 50b-70b 就会出现和 gpt4 一样能力的大模型，看 code llama 能力很强。
4. 数据量小的情况下，sft 可以在 chat 模型上进行微调；数据量大的情况，可以在 base 模型上微调。
5. ceval 上有些分数高的离谱，可能是加了一些测试集做训练；gaokao 这个数据集可能会好点，题比较多；在 gaokao 语文上，gpt 得分竟然很低。

| 模型名称 | 支持语言 | 组织机构 | TOTAL | 语文 | 英语 | 数学 | 生物 | 历史 | 物理 | 政治 |
|---|---|---|---|---|---|---|---|---|---|---|
| ChatGPT(with GPT-4 ~1000B) | 英文 | OpenAI | **0.604** | 0.100 | **0.614** | **0.444** | **0.929** | **0.710** | **0.625** | **0.810** |
| ChatGPT(with GPT-3.5-turbo 175B) | 英文 | OpenAI | 0.425 | 0.100 | **0.614** | **0.444** | 0.429 | 0.645 | 0.125 | 0.619 |
| AquilaChat-7B | 中英 | 智源 | 0.372 | **0.150** | 0.432 | 0.333 | 0.500 | 0.323 | **0.625** | 0.238 |
| ChatGLM2-6B | 中英 | 智谱/清华 | 0.255 | 0.000 | 0.500 | 0.222 | 0.381 | 0.516 | 0.167 | 0.000 |
| Chinese-Alpaca | 中英 | YimingCui | 0.247 | **0.150** | 0.515 | 0.074 | 0.214 | 0.204 | 0.333 | 0.238 |
| StableLM-Alpha | 英语 | StabilityAI | 0.245 | 0.050 | 0.242 | 0.370 | 0.310 | 0.226 | 0.292 | 0.222 |
| Alpaca | 英语 | 斯坦福大学 | 0.233 | 0.000 | 0.424 | 0.259 | 0.381 | 0.355 | 0.167 | 0.048 |
| MOSS-003-SFT | 中文 | 复旦大学 | 0.229 | 0.067 | 0.492 | 0.111 | 0.190 | 0.280 | 0.208 | 0.254 |
| BELLE-LLaMA | 中英 | 链家 | 0.116 | 0.000 | 0.394 | 0.000 | 0.000 | 0.054 | 0.125 | 0.238 |
| ChatGLM-6B | 中英 | 智谱/清华 | 0.099 | 0.000 | 0.386 | 0.000 | 0.048 | 0.215 | 0.042 | 0.000 |

大模型在 Gaokao2023 V1.0 评测集的结果

# 线上讨论:

# 群里讨论：

有空会同步，取决于人力，希望大家积极认领~

# 参考资料

# 后续计划

我们正式升级为一个不太正式的组织了！叫做 OpenLLMAI.

https://github.com/OpenLLMAI

- 正式开启 OpenLLM talk 系列的运营，P1；
- ChatPiXiu 项目：陆续有一些实践计划，现已分拆为各个项目，貔貅只做文档，P1；
- https://github.com/OpenLLMAI/OpenLLaMA2，P0，doing
- https://github.com/OpenLLMAI/chinese-llama2，P0，doing
- https://github.com/OpenLLMAI/OpenLLMData，P0，doing
- OpenSE：检索项目，字符检索+语义检索，P1；
- OpenLLM：LLM 学习和实践项目，P0；
- OpenAIWiki：AI wiki for everyone；
- ChatLover：模拟恋人+爱情助手，P1；

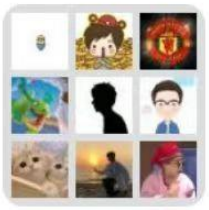
# 加入/赞助我们

蹲算力！！！

我们非常缺人，也非常缺时间和算力，希望能有越来越多的朋友参与进来，认领 talk 的组织者、主持人（最近工作比之前忙不少，不太可能每期都由我来组织了~）、版块的负责人；参与项目后续的开发和讨论等等。

微信群：（请优先加入微信群，如果失效则加入 QQ 群再私聊我进微信群）

（二维码过期了 !）



群聊：羡鱼智能-OpenLLM技术交流群

该二维码7天内(7月7日前)有效，重新进入将更新

QQ 群：

# 羡鱼智能-OpenLL...

群号：740679327

扫一扫二维码，加入群聊。

QQ

## 往期精彩

【OpenLLM Talk 006】本期提要：LLM 加水印；softmax 的 bug；llama2 汉化；多轮对话；DPO 论文阅读；LLM 评估；SE；量化；NOPE；长度外推；OpenLLMAI 与实践计划 - 羡鱼智能的文章 - 知乎

https://zhuanlan.zhihu.com/p/647879679

【OpenLLM Talk 005】本期提要：llama2；FreeWilly；LLM 推理与评估；LLM 八股；RetNet；DPO；数据配比 - 羡鱼智能的文章 - 知乎

https://zhuanlan.zhihu.com/p/645679737

【OpenLLM Talk 004】本期提要：外挂知识；抱抱脸每日论文；MOSS-RLHF；GPT4 细节；OpenAI 代码解释器；百川 13B；LLM 面经；多轮对话；数学能力；反思；LLM 中的知识 - 羡鱼智能的文章 - 知乎

https://zhuanlan.zhihu.com/p/643960837

【OpenLLM Talk 003】本期提要：SuperCLUE-Open；文心盘古；chatlaw；LLM 综述；NTK-Aware Scaled RoPE；10 亿上下文；InternLM；GLM 讲座 - 羡鱼智能的文章 - 知乎

https://zhuanlan.zhihu.com/p/642376781

【【OpenLLM Talk 003】SuperCLUE-Open；文心盘古；chatlaw；LLM 综述；NTK-Aware Scaled RoPE；GLM 讲座】 【精准空降到 10:10】
https://www.bilibili.com/video/BV1Kh4y1E7nX/?share_source=copy_web&vd_source=9e7882f0ef2735e23d66a6f128612943&t=610

【OpenLLM Talk 002】本期提要：chatgpt 增速放缓；gorilla-cli；RoPE 外推；vllm vs llama.cpp；lora 融合；模型参数和数据之比；OpenSE 计划 - 羡鱼智能的文章 - 知乎

https://zhuanlan.zhihu.com/p/641285737

【OpenLLM Talk 001】本期提要：长程记忆；OpenAI 上新；百川智能 7B 模型；State of GPT；位置编码；deepspeed-rlhf；RLHF 数据 - 羡鱼智能的文章 - 知乎

https://zhuanlan.zhihu.com/p/640275116

【OpenLLM Talk 000】我们做了一个 LLM 领域的交流平台 - 羡鱼智能的文章 - 知乎

https://zhuanlan.zhihu.com/p/636350755

【OpenLLM Talk 模版】兴趣和热爱胜过一切，OpenLLM 就从这里开始吧！欢迎加入！ - 羡鱼智能的文章 - 知乎

https://zhuanlan.zhihu.com/p/640522290