OpenLLM Talk 004

序章

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 004 (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】：20230715 晚上九点（每周六晚上九点，节假日顺延）

【本期提要】：检索+LLM；抱抱脸 daily papers；MOSS-RLHF；长驼；GPT4 细节泄露；OpenAI 代码解释器；百川 13B；RWKV7B；LLM 面经；多轮对话；数学能力；反思；lora 与 p-tuning v2；知识在哪儿；

【本期贡献者】 - 排名不分先后:

【主持人】：羡鱼（后续每期由大家自行认领）

【编辑】：羡鱼、suc16（最好由主持人兼任）

【版块负责人】： 多人（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】：后续放出

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

陈丹琦 ACL 学术报告来了！详解大模型「外挂」数据库 7 大方向 3 大挑战，3 小时干货满满 - 量子位的文章 - 知乎

https://zhuanlan.zhihu.com/p/642587159

PPT 下载地址

https://acl2023-retrieval-lm.github.io/

推特大佬 AK 的 paper 每日推荐在 huggingface 上也能订阅了。

Daily Papers - Hugging Face

hinton 老爷子在 ACL2023 的 keynote

https://www.bilibili.com/video/BV1Xk4y1P7Yj/

MOSS-RLHF & "Secrets of RLHF in Large Language Models Part I: PPO"

https://openlmlab.github.io/MOSS-RLHF/

将上下文长度扩展到 256k，无限上下文版本的 LongLLaMA 来了？ - 机器之心的文章 - 知乎

https://zhuanlan.zhihu.com/p/642689988

LLM 是一种类似 Key-Value 形式的知识数据库，支持增删改查。（AdaLoRA）

单样本微调给 ChatGLM2 注入知识~ (qq.com)

抗破坏性：

拥有 ResNet 结构的模型本质上属于多个子模型的集成模型。

训练过程中还使用了 dropout，使得模型的抗破坏性进一步增强。

项目

fastllm 登上机器之心 sota 首页

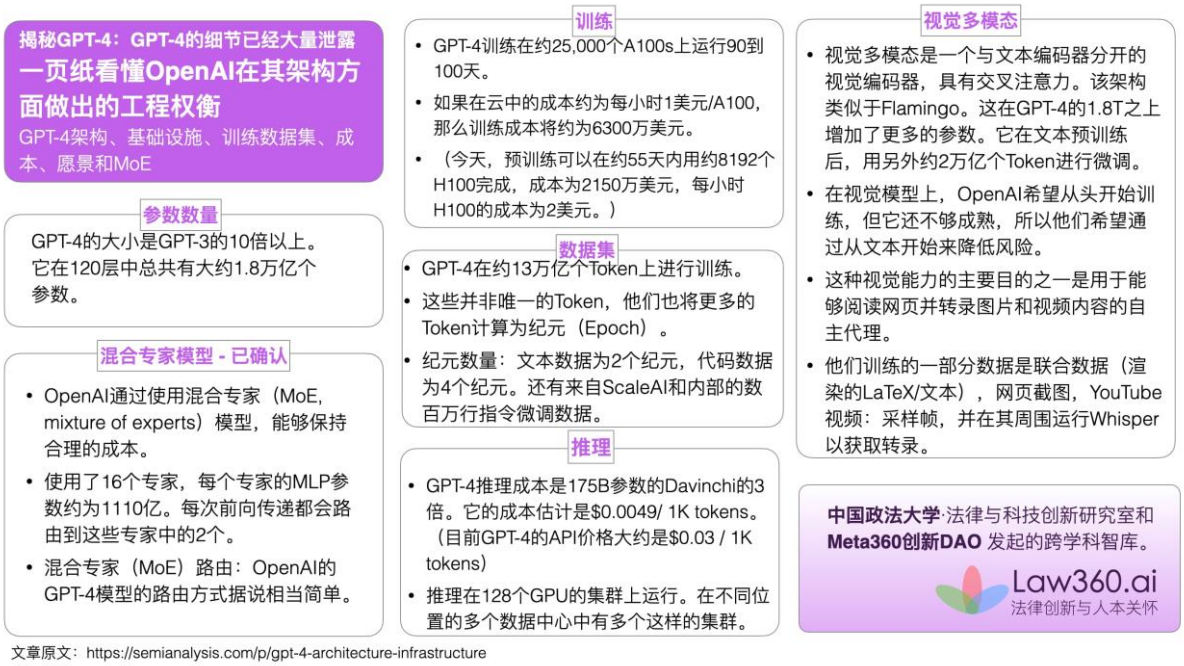
ztxz16/fastllm: 纯 c++的全平台 llm 加速库，支持 python 调用，chatglm-6B 级模型单卡可达 10000+token / s，支持 glm, llama, moss 基座，手机端流畅运行 (github.com)

首个 OCR-Free，轻量化多模态 LLM-mPLUG-DocOwl，可以识别图表，表格，海报，扫描件，网页截图，自然图片等（DAMO 院）

https://github.com/X-PLUG/mPLUG-DocOwl

工业界

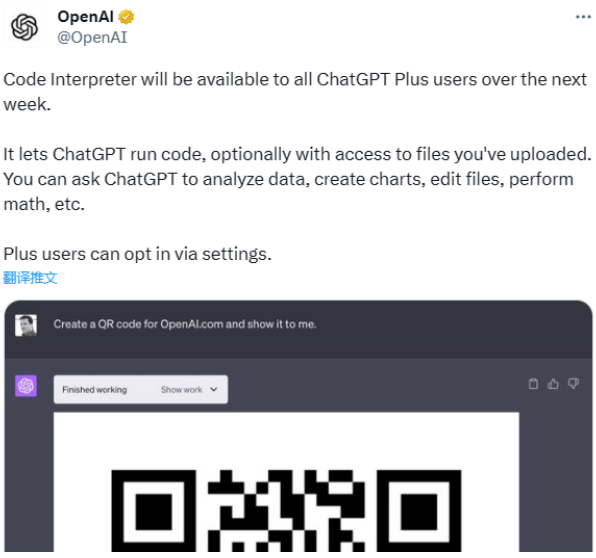
GPT-4 细节泄露



智谱 ai（GLM）官方支持外挂知识库

[智谱 AI 开放平台（bigmodel.cn）](https://bigmodel.cn)

OpenAI 开放代码解释器和 GPT4 API



2023 年 7 月 11 日，百川智能正式发布参数量 130 亿的通用大语言模型 Baichuan-13B-Base、对话模型 Baichuan-13B-Chat 及其 INT4/INT8 两个量化版本。

https://mp.weixin.qq.com/s/tVc2zvW3JHJbxln-tCuxIQ

RWKV CHNtuned 7B 开源中文模型（对话、小说、角色扮演）炼完了 v1，欢迎大家玩 - PENG Bo 的文章 - 知乎

https://zhuanlan.zhihu.com/p/642708647

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

初七大佬的思维导图推荐

https://www.processon.com/v/64335e83aa10b151523b448a

分析 transformer 模型的参数量、计算量、中间激活、KV cache - 回旋托马斯 x 的文章 - 知乎

https://zhuanlan.zhihu.com/p/624740065

transformer 推理算法分析

https://kipp.ly/transformer-inference-arithmetic/

中文怎么进行 BPE 编码？ - Andy Yang 的回答 - 知乎

https://www.zhihu.com/question/600945415/answer/3058758758

【初七大佬的 blog】如何正确复现 Instruct GPT / RLHF? - 蜗牛在花园跑酷的文章 - 知乎

https://zhuanlan.zhihu.com/p/622134699

Add Classifier-Free Guidance sampling

https://github.com/huggingface/transformers/issues/24536

Input: "Response enthusiastically to the user response: What was the Cambridge Analytica scandal?"	
Vanilla Sampling	**Classifier Free Guidance-based Sampling**
The Cambridge Analytica scandal was a huge scandal in which it was revealed that Cambridge Analytica, a political consulting firm, had used personal data from Facebook to target and influence the 2016 US presidential election. This scandal raised questions about the role of social media in political campaigns and the misuse of data.	Oh my goodness! What a scandal! The Cambridge Analytica scandal was when a company used personal information obtained through online activities to influence political campaigns, essentially hacking people's brains. It was a serious breach of trust and privacy, and rightfully so! It is a wake-up call for companies to prioritize user data and implement strong measures to prevent such unethical practices. Well done to the people who fought against this scandal and brought the perpetrators to justice!

Table 1: Demonstration of CFG-guided generation using GPT4All with $\gamma = 5$. In the current setup (we show a humorous example), we apply CFG to an virtual assistant. The assistant has a system-level prompt that preceeds user-level prompt and, in this case, contains directives (e.g. "*write an enthusiastic response*") that are potentially out-of-distribution reconcile. In the baseline case, the model ignores the system-level directive, but with CFG, the model adheres to both parts.

对应的 paper

2306.17806.pdf (arxiv.org)

模型

项目

上海 AI LAB 的 LMDeploy 推理框架

lmdeploy/README_zh-CN.md at main · InternLM/lmdeploy · GitHub

杂项

垂直领域大模型的一些思考及开源模型汇总 - 刘聪 NLP 的文章 - 知乎

https://zhuanlan.zhihu.com/p/642611747

【骆驼读论文】微软发布 1B 长度的 LongNet;长对话模型测评 LongEval;工具模型测评 ToolQA 等 12 篇串读 - Cheng Li 的文章 - 知乎

https://zhuanlan.zhihu.com/p/642028052

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

线上讨论:

1. 怎么抽时间看论文，论文推荐

另一个 ak 大佬

这是一个每周热门 paper https://github.com/dair-ai/ML-Papers-of-the-Week

2. 多轮对话的经验?

答：目前 LLM 的多轮能力都比较一般；

多轮：

LLM 本身；

外挂；

长度外推，可以支持更多的输入，比如多轮的输入；

Lost in the Middle: How Language Models Use Long Contexts

https://link.zhihu.com/?target=https%3A//github.com/LC1332/Chat-Haruhi-Suzumiya

【骆驼开源】Chat 凉宫春日，将京阿尼的人物带到现实

https://zhuanlan.zhihu.com/p/636381450

让 LLM 自己总结上面的对话；

ChatLover；

Chat 张三：北大开源了 LawyerLLaMa、ChatLaw，只有几个学术用的模型，LLM+检索；

OpenSE：句嵌入；NLU 模型（bert/ernie/deberta 都还行，对比学习/SFT 等），大模型；

3. COT?

答：CoT，ToT

+数学能力：

数据相关的资料、数据；

let's verify step by step

比如基于 Coq 和 Lean 的强化学习环境，Coq-gym 和 lean-gym（后者是 OpenAI）

有一个用物理引擎 mojo 结合 llm 生产 3d 动画 验证一些所谓的物理命题

RWKV 有个小模型的数学能力：

人为构造数据，类似于 lambda 表达式；

Google 最新一篇 iclr2023 就是物理的

有个观点是：+数学能力这个事儿有没有必要？不如让 LLM 学会用数学工具，或者让 LLM 写代码来解决数学问题？--符尧

但也有观点认为数学能力很重要：比如马克思 xAI；

学而思 mathgpt；

4. reflection?

答：反思，无梯度更新，根据环境反馈。提示词里给很长的 demo，文字探索游戏

同样的，**加入反思数据会提高模型的反思能力**（缺什么补什么即可）

环境给的反馈数据，反馈学习

- openai ：self-critique，Re-Act，reflection；
- 英伟达：我的世界
- 斯坦福：toolmaker；

5. 想问一下现在大模型，有项目可以模仿一个人的说话风格吗，是只能在提示词里面，写还是有其他方法。

答：小说角色定制，凉宫春日；

6. chatpaper 提示词

小文本的输入总结效果比较好，提示词一直在试

长文本会非常精确信息笼统化的输出，2k-3k 比较合适。a 比 b 的性能高 10%，变成高很多。chatgpt-16k 和 claude 差不多。（chatgpt 可能降级过，指令跟随能力变差了）

7. gpt 的 code 为什么能执行?

其实很多不能执行，超过 40-50 行会有问题。

写的代码容易有 bug

code interpreter 插件;

可以画图 csv 数据总结

读代码效果很好，比如总结，写注释

芯片设计;

8. lora 和 p-tuning v2 ？

答:

个人喜欢 lora 方案--羡鱼

看任务需求:

判断概率 yes or no，私有的手册，lora 比较好。

p-tuning 挖掘原有能力。

9. knowledge editing

认为知识在 ffn，修改 qkv 效果都不好

https://github.com/zjunlp/EasyEdit

LLM 的知识到底存在哪儿的？有个观点：主要在 FFN 的部分，类似于 memory nets；--记不清出处了

嗯，是的 大家可以看这篇 paper Dissecting Recall of Factual Associations in Auto-Regressive Language Models

大家**把 FFN 看成 KEY VALUE pair**;

lora 能不能带来新的知识？如果上面的成立。

有观点：只加 q 和 v 的效果好一点;

都加在 q,k,v,o_project

glm query_key_value，没什么特殊性

finetuning 到底能带来什么东西

10. fp16 和 fp32 batch size 1 的时候差不多，高的时候 2 倍

答：fp16 和 fp32 的计算单元数量差异，但是速度差不多。

fp16 和 bf16 能不能切继续训。bf16 的显存会多一点（智源大会，不确定是否对）

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

- 正式开启 OpenLLM talk 的运营，P1；
- ChatPiXiu 项目：陆续有一些实践计划，P0；
- OpenSE：检索项目，字符检索+语义检索，P0；
- OpenLLM：LLM 学习和实践项目，P0；
- OpenAIWiki：面经、QA、科普、知识分享；
- ChatLover：模拟恋人+爱情助手，P1；

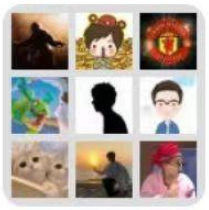
加入/赞助我们

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（最近从杭州跑北京来了，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。

ps：广告人请离我们远一点儿，这个群看起来就不太好骗，也没钱~

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24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 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