国产ChatGPT开测 更加适合中文、支持国产CPU训练

最新的国产AI大模型已经开始测试了。这个新模型在中文优化和国产CPU训练方面都有重要的突破。这个新模型正在接受多方面的测试，包括性能测试和使用测试，以确保它能够轻松地满足用户的需求。

据介绍，对话机器人 ChatGLM（alpha内测版：QAGLM），这是一个初具问答和对话功能的千亿中英语言模型， 并针对中文进行了优化，现已开启邀请制内测，后续还会逐步扩大内测范围。

与此同时，继开源 GLM-130B
千亿基座模型之后，我们正式开源最新的中英双语对话 GLM 模型： ChatGLM-6B，结合模型量化技术，用户可以在消费级的显卡上进行本地部署（INT4 量化级别下最低只需 6GB 显存）。

经过约 1T 标识符的中英双语训练，辅以监督微调、
反馈自助、人类反馈强化学习等技术的加持，62 亿参数的 ChatGLM-6B 虽然规模不及千亿模型，但大大降低了用户部署的门槛，并且已经能生成相当符合人类偏好的回答。

ChatGLM 参考了 ChatGPT 的设计思路，在千亿基座模型
GLM-130B1中注入了代码预训练，通过有监督微调（Supervised Fine-Tuning）等技术实现人类意图对齐。

ChatGLM 当前版本模型的能力提升主要来源于独特的千亿基座模型 GLM-130B。它是不同于 BERT、GPT-3 以及 T5 的架构，是一个包含多目标函数的自回归预训练模型。

2022年8月，我们向研究界和工业界开放了拥有1300亿参数的中英双语稠密模型 GLM-130B1，该模型有一些独特的优势：

双语：同时支持中文和英文。

高精度（英文）：在公开的英文自然语言榜单 LAMBADA、MMLU 和 Big-bench-lite 上优于 GPT-3 175B（API: davinci，基座模型）、OPT-175B 和 BLOOM-176B。

高精度（中文）：在7个零样本 CLUE 数据集和5个零样本 FewCLUE 数据集上明显优于 ERNIE TITAN 3.0 260B 和 YUAN 1.0-245B。

快速推理：首个实现 INT4 量化的千亿模型，支持用一台 4 卡 3090 或 8 卡 2080Ti 服务器进行快速且基本无损推理。

可复现性：所有结果（超过 30 个任务）均可通过我们的开源代码和模型参数复现。

跨平台：支持在国产的海光 DCU、华为昇腾 910 和申威处理器及美国的英伟达芯片上进行训练与推理。

2022年11月，斯坦福大学大模型中心对全球30个主流大模型进行了全方位的评测2，GLM-130B 是亚洲唯一入选的大模型。

在与 OpenAI、谷歌大脑、微软、英伟达、脸书的各大模型对比中，评测报告显示 GLM-130B 在准确性和恶意性指标上与 GPT-3 175B (davinci) 接近或持平，鲁棒性和校准误差在所有千亿规模的基座大模型（作为公平对比，只对比无指令提示微调模型）中表现不错（下图）。

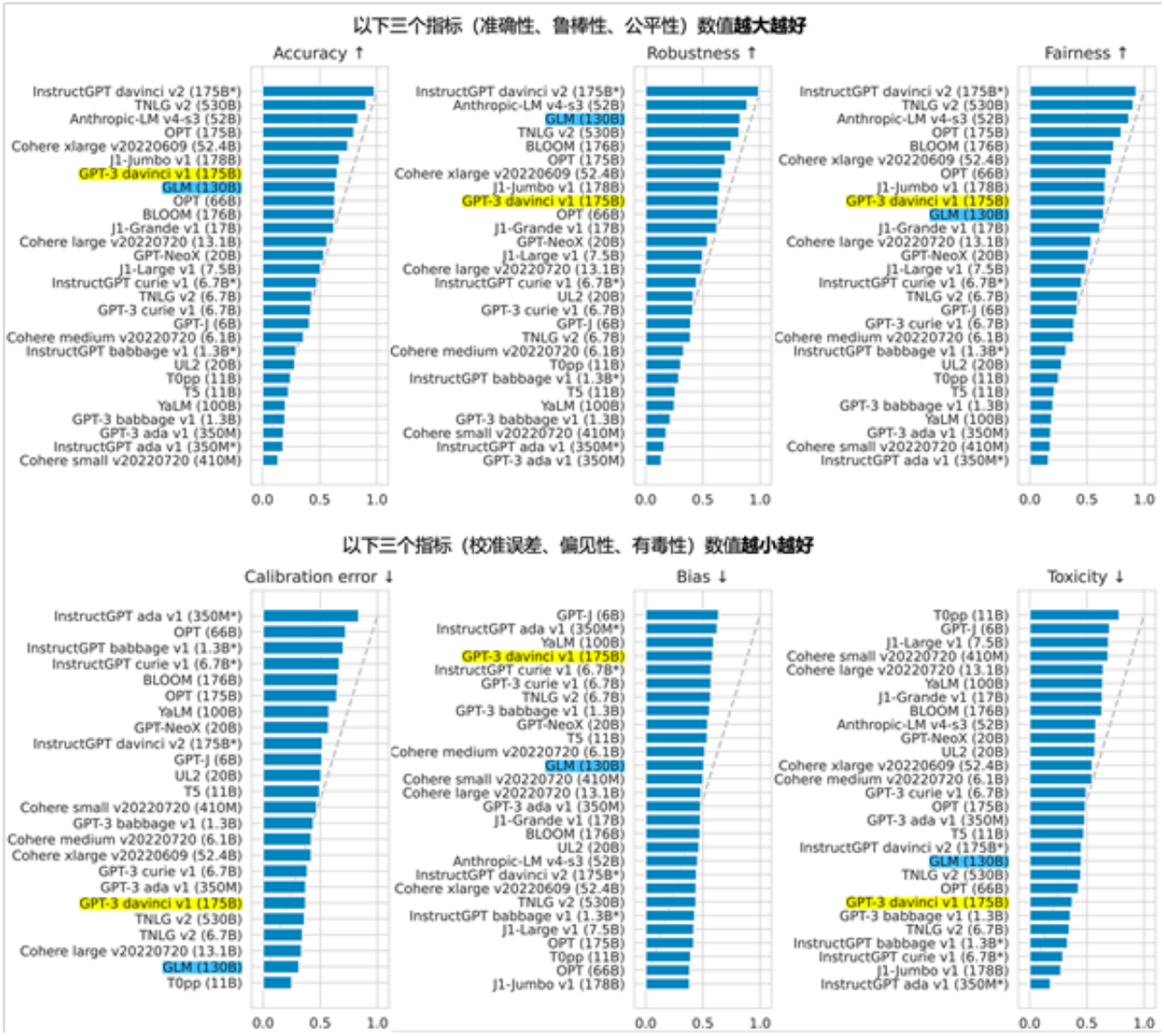

图1. 斯坦福大学基础模型中心对全球 30 个大模型的评测结果（2022年11月）

本文链接：https://dqcm.net/zixun/167884768812187.html