复旦课程实践-LLM技术挑战

# 背景

随着大语言模型（LLM）技术的飞速进步，Intel开发了一系列在Intel硬件平台上支持LLM的AI工具，其中之一就是Intel® Extension for Transformers：这是一个创新工具包，旨在各种Intel平台（包括Intel Gaudi2、Intel CPU 和Intel GPU）上实现基于 Transformer 的模型的最佳性能，从而在任何地方加速 GenAI/LLM。 该工具包提供以下主要功能和示例：

* 通过扩展 Hugging Face Transformer API和Intel® Neural Compressor，在基于 Transformer 的模型上提供无感的模型压缩优化功能， 结合自带的纯 C/C++ 开发的大型语言模型 (LLM) 推理引擎，对常见的LLM提供推理加速功能。
* ITREX/NeuralChat: 一个可定制的聊天机器人框架，可利用知识检索、语音交互、查询缓存和安全防护等丰富的插件在几分钟内创建您自己的聊天机器人。该框架支持Intel Gaudi2/CPU/GPU。

基于LLM的应用技术方兴未艾。是LLM技术走进生产生活的关键环节。由于Intel提供的硬件到处可得，基于海量Intel平台开发的LLM应用，更具有实践意义。

# 挑战时间

2024年3月23日-5月15日

# 参与方式

个人或组队参加，每队成员人数最多不超过3人

# 奖项设置

1. 最佳实践奖3名，奖金每支队伍500元人民币或等额京东卡，由校企双方共同评审
2. 优秀分享奖若干名，奖金每支队伍100元人民币或等额京东卡，需要完成以下任务并提交至[腾讯文档](https://docs.qq.com/form/page/DWmlpeXhXZW9iemtG)[https://docs.qq.com/form/page/DWmlpeXhXZW9iemtG](https://wx.qq.com/cgi-bin/mmwebwx-bin/webwxcheckurl?requrl=https%3A%2F%2Fdocs.qq.com%2Fform%2Fpage%2FDWmlpeXhXZW9iemtG&skey=%40crypt_9790c688_28f70640b31a27d2baf6827634d1c287&deviceid=e512559582578472&pass_ticket=IsrjO7%252BthdnwZ7tNGVWxImlWoTUNR0xXcJ%252BNe5Iaq%252FN6ji%252F7%252B%252Be%252BN%252FiN6XO2suOmij4MD2AfNSg%252FHLNKUoNJdA%253D%253D&opcode=2&scene=1&username=@81b397c35c6a8f56fdfa9234240ef8898857c611c2b4ef8810fd3530c305ad44) ，审核合格后发放：公开发布LLM 模型文件：提交到魔搭社区或者hugging face网站上（CSDN也可以）。在总结报告和腾讯文档中提供公开可访问链接。

# 挑战内容

开发者利用开源的 Intel® Extension for Transformers或其组件(Neural Chat)，在魔搭社区云环境，或者自备Intel硬件或云端，开发基于大语言模型的新型应用。

应用的种类包括但不限于：智能语音助手、智能助理、聊天机器人、智能客服进行不同领域应用，比如健康服务，适老服务，教育等等。

备注：魔搭社区，参赛者可以到该社区，免费注册并申请使用基于Intel第三代可扩展Xeon处理器的云环境，进行开发测试。

# 结果提交方式：

1. 提交代码给助教或授课老师：可以编译并运行，包含环境安装文档。

2. 总结报告：项目说明文档（包括方案介绍，技术特点，使用到的Intel软硬件技术，成果说明和对照），还可以包含更多材料，如：演示PPT、演示视频等。

3.（可选，有加分和奖励）公开发布LLM 模型文件：提交到魔搭社区或者hugging face网站上（CSDN也可以）。在总结报告和腾讯文档中提供公开可访问链接。

# 评分标准：

满分：120分 + 10分（可选）

|  |  |  |  |
| --- | --- | --- | --- |
| 标准 | 无 | 低 | 高 |
| 1.开发测试环境 | 0分  没有使用基于Intel硬件的本地或云环境， | 5分  提供在Intel平台上测试的数据。  只有一个平台的数据。 | 10分  提供在Intel平台上测试的数据。包含不同平台的比较数据。 |
| 2. 开发工具和库 | 0分  没有使用任何Intel提供的AI工具或者软件 | 15分  使用了Intel提供的工具，但不包含Intel® Extension for Transformers或其组件(Neural Chat) | 30分  使用但不限于Intel® Extension for Transformers或其组件(Neural Chat) |
| 3.代码质量 | 0分  代码根据环境安装文档无法运行。  代码抄袭明显。 | 10分  代码可以运行,但无法重现总结报告中的结果（近似）。 | 20分  代码可以运行,可以重现总结报告中的结果（近似）。 |
| 4.代码优化 | 0分  未使用任何Intel优化方法，措施。 | 10分  使用了一些Intel优化的库或者工具，方法。  未提供优化结果对比数据。  优化结果存疑。 | 20分  广泛使用多种Intel优化的库和工具，方法。  提供了优化结果对比数据。  优化结果可信。 |
| 5.文档完备 | 0分  只有文档，或者只有代码。  文档抄袭明显。 | 5分  有文档和代码。但文档内容和代码不一致，文档内容有矛盾和不合理之处，结果存疑。 | 10分  有文档和代码。文档内容和代码对应。文档中的结果，通过代码可验证，合理。材料可以证明项目真实，结果可信。 |
| 6.创意的技术价值 | 0分  通过文档，无法理解创意。  创意描述过于简单，宽泛。 | 5分  缺乏新颖性，技术实现风险高。 | 10分  创意新颖，技术实现风险低。 |
| 7.创意的商业价值 | 0分  通过文档，无法理解创意。  创意描述过于简单，宽泛。 | 5分  缺乏新颖性，商业价值小。可操作性弱。 | 10分  创意新颖，商业价值较高。可操作性强。 |
| 8.代码实现的完整性 | 0分  实现与创意描述不符合。  不能单独演示。 | 5分  实现了部分创意的内容，但非核心内容。  可以单独演示。 | 10分  实现了全部创意，或者重要核心内容。  可以单独演示。 |
| 9.（可选）公开发布模型文件 | 0分  未提供公开发布模型文件 | 5分  公开发布的模型文件和提交的代码不配合。  自研模型文件存疑。 | 10分  公开发布的模型文件和提交的代码配合正确。  自研模型文件可信 （通过附加材料，代码自证）。 |

# 相关培训资料

1. 大语言模型的微调：

<https://www.bilibili.com/video/BV1Wu4y1c794/>

1. 使用Intel 大模型推理达到40倍加速：

<https://www.bilibili.com/video/BV1nu4y137JT/>

1. 大语言模型的4bits量化神器GPTQ：

https://www.bilibili.com/video/BV1R94y1c7eC/

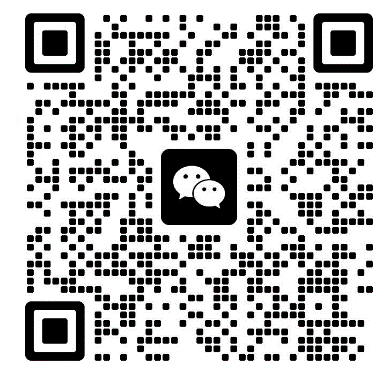
1. CPU上运行图像生成：

<https://www.bilibili.com/video/BV1HG411r7ZV/>

1. 大模型剪枝技术：

<https://www.bilibili.com/video/BV1gK4y1z7V4/>

# 交流&答疑微信群(0326有效期后或加群主微信15821798740入群）



# 注意事项

**原则上所有获奖作品需要开源，如有特殊情况需经终审评审专家集体决策同意可部分开源；大赛最终解释权归英特尔所有。**

# 英特尔学生大使项目正在招募中

欢迎离毕业年份一年以上的在校学生申请，详见<https://software.seek.intel.com/oneapi-academia-community-program-application>