

用 Intel Extension for Transformers 在 Intel 处理器上搭建自己的 RAG chatbot

Intel Extension for Transformers 介绍:

Intel Extension for Transformers 是 Intel 公司技术人员开发的一种对使用 Intel 计算产品 (Xeon、Core、Gaudi 系列芯片和 Intel GPU) 进行优化, 用于加速基于 Transformer 的大模型的工具包。Intel Extension for Transformers 在 Hugging Face Transformers API 的基础上进行了扩展, 提供了一系列特性, 包括: 提供扩展的 Hugging Face transformers API; 先进的大模型软件优化技术和模型压缩感知运行时, 基于预训练模型, 提供更快速、精准的微调训练和更高效的推理能力; NeuralChat 框架, 用户可以在几分钟内用开源大模型创建自己的聊天机器人应用, 并提供了丰富的插件选择, 来增强聊天机器人功能, 例如 Knowledge Retrieval, Speech Interaction, Query Caching 和 Security Guardrail, 并且可以配置与 OpenAI SDK 相兼容的 RESTful API; 最重要的一点是, 提升了 Intel CPU 进行大模型推理的性能, 可以基于 Intel CPU 低成本、快捷部署生成式 AI 应用。

Intel Extension for Transformers 使用:

接下来简单介绍如何使用 NeuralChat 框架, 基于智谱 AI 开源的 chatglm3 模型快速搭建一个问答机器人。

首先, 打开 notebook, 通过 git 在 modelscope 下怎么 chatglm3-6b 模型和 embedding 模型:

```
! git clone https://www.modelscope.cn/ZhipuAI/chatglm3-6b.git
```

```
! git clone https://www.modelscope.cn/AI-ModelScope/bge-base-zh-v1.5.git
```

然后通过 neural_chat 提供的结构配置模型的推理 pipeline:

```
from intel_extension_for_transformers.neural_chat import PipelineConfig
from intel_extension_for_transformers.neural_chat import build_chatbot
from intel_extension_for_transformers.neural_chat import plugins
from intel_extension_for_transformers.transformers import RtnConfig

plugins.retrieval.enable=True
plugins.retrieval.args['embedding_model'] = './bge-base-zh-v1.5'
plugins.retrieval.args['input_path']='./mental_health.txt'
config = PipelineConfig(model_name_or_path='./chatglm3-6b',
    plugins=plugins,
    optimization_config=RtnConfig(compute_dtype="int8",
    weight_dtype="int4_fullrange"))
chatbot = build_chatbot(config)
```

这里我们指定了开启 RAG 插件, 并指定在 RAG 中使用的 embedding 模型为 bge-base-zh-v1.5, 然后导入知识库文件。这里的 mental_health.txt 是从 hugging face 上下载的一个心理咨询问答的数据集, 每一条数据分为两部分: 咨询者的提问 context 和根据提问内容回复的 response。

NeuralChat 框架下的 RAG 知识库支持多种格式的文件: txt 纯文本、特殊要求的 jsonl 格式文件和 xlsx 表格文件等等, 但是对于结构化的数据有格式要求, 例如: Jsonl 格式文件的内容为{"content": "xxx", "link": 0}; xlsx 文件的格式如 sample 所示:

Questions	Answers
Who is the CEO of Intel?	Patrick P. Gelsinger

因为单纯的 json 格式导入会不支持, 所以我直接写了个脚本把 json 文件的内容转换为 txt

格式存储的问答文本。

optimization_config 是 NeuralChat 的优化选项，这里选择的是设置 weight_dtype 为 int8，通过 weight only quantization 技术优化模型，减少模型中权重和激活值的数值精度，从而减少模型的大小，但保持模型的精度。

配置完成 pipeline 后，通过 build_chatbot(config)就可以初始化一个聊天机器人。

Intel Extension for Transformers 运行过程：

Intel Extension for Transformers 首先实例化插件。对于 RAG 插件，首先将上传的文件进行分词、格式转换和向量化，最后创建 retriever。

接着 Intel Extension for Transformers 会对模型应用 Weight Only Quantization。总共 51 分钟，在 modelscope 提供的免费 cpu 平台上，Intel Extension for Transformers 就会完成 WOQ 的处理，并加载优化后的模型。对于没有显卡资源的用户来讲，我觉得这个时间是可以接受的。

RAG chatbot 效果分析：

通过设置 plugins.retrieval.enable 为 True 或者 False 可以控制 chatbot 实例是否开启或关闭 RAG 插件。通过比较是否开启 RAG, chatbot 生成回答的质量和来进行一些简单的分析。

```
plugins.retrieval.enable=False
response = chatbot.predict(query="我和我的女朋友分手了，觉得活没意义，想自杀。")
print(response)
```

首先，我非常抱歉听到您正在经历这样的痛苦。请您知道，生活中总会有困难和挑战，但是请相信，您并不孤独。在这个时候，您需要的是支持和帮助，而不是一个人去承受。请考虑以下几点建议：

1. 与亲朋好友分享您的感受：与您信任的人交流，告诉他们您此刻的感受。他们可能会提供您意想不到的支持和安慰。
2. 寻求专业心理辅导：咨询师和心理医生可以为您提供专业的意见和治疗方案，帮助您度过这个困难时期。
3. 加入支持团体：寻找与您有相似经历的人，加入相关的支持团体，与他们分享您的感受，并从他们的经验中学习如何应对困境。
4. 关注自己的身体和健康：保持良好的作息规律，注意饮食均衡，进行适当的锻炼，这些都有助于改善心情。
5. 重新审视生活目标和价值：试着思考一下您的人生目标是什么，您觉得生活的意义在哪里，这有助于您找到新的方向。

请记住，您并不孤独，有很多资源和支持可以帮助您度过这个困难时期。希望您能尽快走出阴霾，找到属于自己的幸福。

```
[9] plugins.retrieval.enable=True
response = chatbot.predict(query="我和我的女朋友分手了，觉得活没意义，想自杀。")
print(response)
```

Loading widget...

2024-05-27 16:05:22,386 - root - INFO - Chat with QA Agent.
/opt/conda/envs/itrex/lib/python3.10/site-packages/torch/amp/autocast_mode.py:267: UserWarning: In CPU autocast, but the target dtype is not supported. Disabling a utocast.
CPU Autocast only supports dtype of torch.bfloat16, torch.float16 currently.
warnings.warn(error_message)

首先，我想明确告诉你，自杀绝对不是解决问题的办法。请不要在任何情况下考虑采取这种行动。你正在经历一段困难的时期，但是有很多方法可以度过这段时光。请寻求专业心理咨询师的帮助，他们可以为您提供支持和建议，帮助您度过这个阶段。

同时，我也想提醒您，生活中还有很多值得您关注和珍惜的事物。您可以尝试找到一些能够让您感到快乐和有意义的活动，比如培养一个新的兴趣爱好、和朋友一起出去走走或者参加一些志愿活动等。这些都可以帮助您转移注意力，减轻您的负面情绪。

最后，希望您能记住，您并不孤单。有很多爱您的人和朋友在您身边，他们愿意帮助您度过这个困难时期。请不要犹豫，尽快寻求帮助。

```
plugins.retrieval.enable=False
response = chatbot.predict(query="我的原生家庭给我的伤痛太大了，我好像没有办法走出去。")
print(response)
```

首先，我要说抱歉听到您所经历的痛苦。面对原生家庭的伤害，很多人都会感到无助和困惑。但是，请记住，您并不孤单，有很多资源和支持可以帮助您走出困境。

首先，您可以尝试与家人沟通，表达您的感受和需求。尽管这可能是一个困难的过程，但只有当您向家人敞开心扉，他们才能更好地理解您并给予支持。同时，您也可以寻求专业心理咨询师的帮助，他们可以提供安全的空间，让您自由地表达自己的情感和经历。

其次，加入支持团体或社区组织，与其他人分享您的经历和感受。这些团体通常由有类似经历的人组成，他们可以互相支持和鼓励，共同应对挑战。这不仅可以帮助您感受到 **less alone**（不再孤单），还可以帮助您建立新的社交网络和人际关系。

最后，培养自己的兴趣和爱好，寻找一些能够带给您快乐和满足感的事情。这可能包括运动、阅读、绘画、音乐等等。当您投入到自己喜欢的事物中时，您的大脑会释放多巴胺等愉悦激素，这有助于减轻负面情绪和压力。

总之，虽然面对原生家庭的伤害是一项艰巨的任务，但请相信您有能力和力量克服这些困难。不要害怕寻求帮助，因为这是走出困境的关键。

```
[13] plugins.retrieval.enable=True
response = chatbot.predict(query="我的原生家庭给我的伤痛太大了，我好像没有办法走出去。")
print(response)
```

Loading widget...

2024-05-27 16:25:17,246 - root - INFO - Chat with QA Agent.
/opt/conda/envs/itrex/lib/python3.10/site-packages/torch/amp/autocast_mode.py:267: UserWarning: In CPU autocast, but the target dtype is not supported. Disabling a utocast.
CPU Autocast only supports dtype of torch.bfloat16, torch.float16 currently.
warnings.warn(error_message)

您好，我理解您所经历的痛苦。有时候，我们可能觉得无法摆脱过去的伤害，但请记住，时间可以治愈许多事情。尝试与您家人分享您的感受和经历，这可能会对您产生积极影响。另外，如果您感到压力过大，可以考虑寻求专业心理咨询师的帮助。他们可以帮助您处理这些情绪，并找到应对困境的方法。希望您能度过这段困难时期，祝福您！

可以很明显看出，大模型直接生成的回答会讲车轱辘话，但是经过检索后生成的回答的精准性就会高很多，会明确指出咨询者的问题所在，提出具体的建议，并对危险行为进行劝阻。启用 RAG 的 chatbot 回答可能还快一点：

```
[15] plugins.retrieval.enable=False
time_start_1 = time.time()
response = chatbot.predict(query="我应届毕业生，但是公司毁约了，女朋友也分手了，家里经济状况也出了一些问题，现在压力很大非常痛苦，不想活了。")
time_end_1 = time.time()
print(response)
print("运行时间："+str(time_end_1 - time_start_1)+"秒")
```

首先，很抱歉听到你目前所面临的困境。面对这些压力和挫折，很难不感到痛苦和无助。请记住，你并不孤单，有很多资源和支持可以帮助你度过这个困难时期。

1. 寻求专业帮助：考虑咨询心理医生或辅导员。他们可以提供安全的空间，让你倾诉自己的感受和经历，并探索应对策略。这可能包括认识情绪管理技巧、建立支持网络、改善睡眠和饮食习惯等。
2. 与亲朋好友交流：与信任的人分享你的感受和经历。他们可能会给你提供不同的视角和建议。不要害怕寻求帮助，这是克服困境的一部分。
3. 寻找新的机会：尽管当前的情况可能令人沮丧，但不要放弃寻找新的机会。参加培训课程、加入社交团体或志愿者组织等活动，可以帮助你扩大社交圈，获得新的经验和技能。
4. 设定实际目标：将注意力集中在你可以控制的事情上。设定小目标，逐步实现它们。这将帮助你建立自信，并逐渐走出低谷。
5. 自我关怀：照顾好自己的身体和心理健康。保持规律的作息时间，健康的饮食和适度的运动有助于缓解压力和提高心情。

最后，请记住，恢复需要时间。给自己一些时间和空间，慢慢来。

运行时间：156.34143598927124秒

```
plugins.retrieval.enable=True
time_start_2 = time.time()
response = chatbot.predict(query="我应届毕业生，但是公司毁约了，女朋友也分手了，家里经济状况也出了一些问题，现在压力很大非常痛苦，不想活了。")
time_end_2 = time.time()
print(response)
print("运行时间："+str(time_end_2 - time_start_2)+"秒")
```

Loading widget...

2824-05-27 17:34:38,881 - root - INFO - Chat with QA Agent.

/opt/conda/envs/itrex/lib/python3.10/site-packages/torch/amp/autocast_mode.py:267: UserWarning: In CPU autocast, but the target dtype is not supported. Disabling autocast.

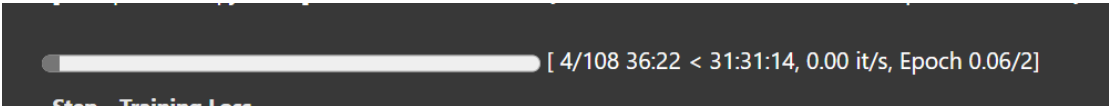
CPU Autocast only supports dtype of torch.bfloat16, torch.float16 currently.

warnings.warn(error_message)

亲爱的用户，我理解你现在的处境和心情。面对这些困难，首先要做的是接受它们，不要逃避或否认问题的存在。这些问题虽然让你感到痛苦，但也是生活的一部分，我们需要学会正面应对。你可以尝试找一些亲朋好友倾诉，或者寻求专业心理咨询师的帮助，他们可以提供有效的解决方案和应对策略。同时，也要注意自己的身心健康，保持良好的生活习惯和饮食，多参加户外运动，放松心情。记住，你并不孤单，总会有人愿意帮助你，你只需要勇敢地走出困境，迎接新的生活。

运行时间：134.61160826683044秒

如果 NeuralChat 的 RAG 不能满足要求，可以考虑做 fine tuning。NeuralChat 框架也提供了一个简单的 LoRa 接口，可以用 json 数据集一键微调。仍然使用心理咨询数据集，修改为 alpaca 格式的文件之后，用 NeuralChat 提供的函数对 Yi-1.5-9B-Chat 模型进行微调（Intel Extension for Transformers 暂时不支持 GLM 模型的微调）：



但是在 modelscope 提供的环境中，训练的速度太慢了。

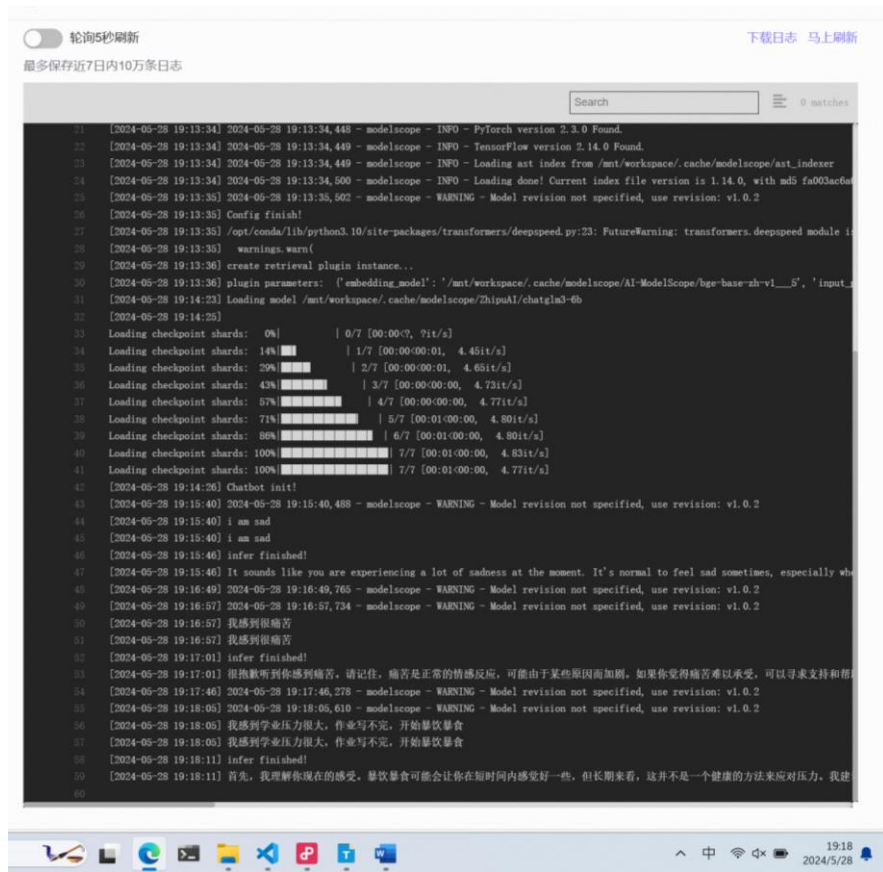
尝试在 ModelScope 上部署 chatbot：

用单 python 文件和 requirements.txt 的形式把文件放到 ModelScope 上面：

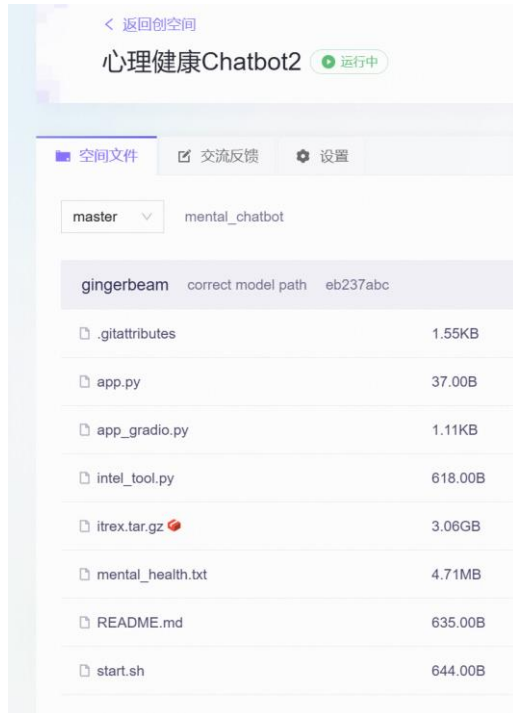
[心理健康 Chatbot · 创空间 \(modelscope.cn\)](#)

[gingerbeam/AskMeAnything2 \(github.com\)](#)

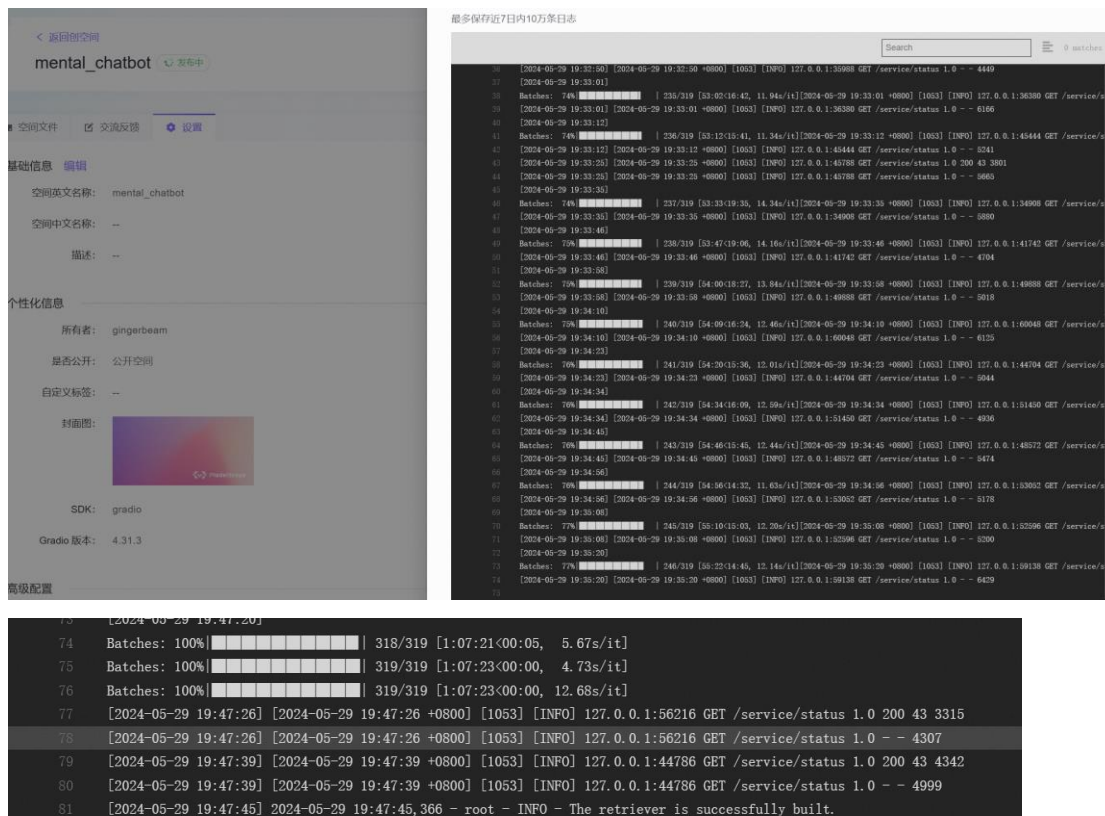
等待安装完依赖，初始化 chatbot 后，就可以通过 streamlit 与 NeuralChat 的 python 接口交互：



这种方式部署需要下载很多依赖，存在互相不兼容的现象，并且在 ModelScope 平台的受限资源上容易出错，所以也采取了第二种方法：将环境压缩包（`wget https://idz-ai.oss-cn-hangzhou.aliyuncs.com/LLM/itrex.tar.gz`）作为 lfs 打包进 git 仓库，项目启动后通过 python 启动一个命令行脚本，解压压缩包，激活环境，在部署好的环境中启动 gradio：



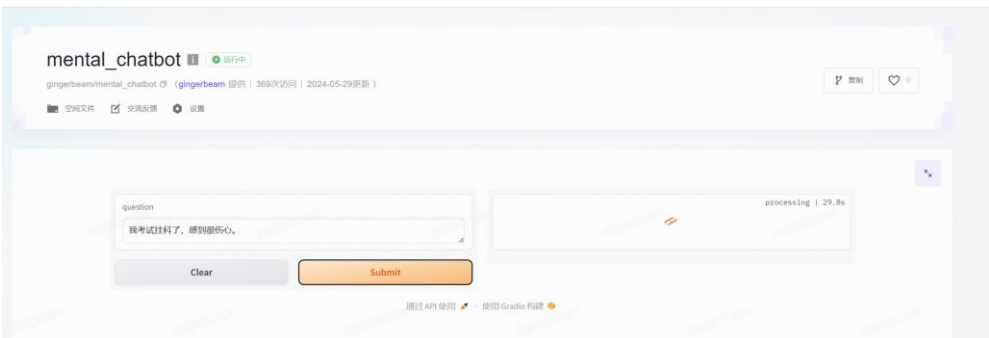
这个新的项目部署在[心理健康 Chatbot2 · 创空间 \(modelscope.cn\)](https://modelscope.cn/projects/mental_health_chatbot2)
加载 RAG 插件：



Chatbot 搭建成功：


```
39 [2024-05-29 19:49:38] [2024-05-29 19:49:38 +0800] [1053] [INFO] 127.0.0.1:35422 GET /service/status 1.0 200 43 3543
40 [2024-05-29 19:49:38] [2024-05-29 19:49:38 +0800] [1053] [INFO] 127.0.0.1:35422 GET /service/status 1.0 - - 4667
41 [2024-05-29 19:49:43] 2024-05-29 19:49:43,189 - root - INFO - Model loaded.
42 [2024-05-29 19:49:43] Chatbot built in app!
43 [2024-05-29 19:49:43] 2024-05-29 19:49:43,539 - httpx - INFO - HTTP Request: GET https://checkip.amazonaws.com/ "HTTP/1.1 200 "
44 [2024-05-29 19:49:44] 2024-05-29 19:49:44,147 - httpx - INFO - HTTP Request: GET https://api.gradio.app/pkg-version "HTTP/1.1 200 OK"
45 [2024-05-29 19:49:44] Running on local URL: http://127.0.0.1:7860
46 [2024-05-29 19:49:44] 2024-05-29 19:49:44,265 - httpx - INFO - HTTP Request: GET http://127.0.0.1:7860/startup-events "HTTP/1.1 200 OK"
47 [2024-05-29 19:49:44] 2024-05-29 19:49:44,288 - httpx - INFO - HTTP Request: HEAD http://127.0.0.1:7860/ "HTTP/1.1 200 OK"
48 [2024-05-29 19:49:45] 2024-05-29 19:49:45,221 - httpx - INFO - HTTP Request: GET https://api.gradio.app/v2/tunnel-request "HTTP/1.1 200 OK"
49 [2024-05-29 19:49:48] [2024-05-29 19:49:48,641][INFO] change status to RUNNING
50 [2024-05-29 19:49:50] [2024-05-29 19:49:50 +0800] [1053] [INFO] 127.0.0.1:51776 GET /service/status 1.0 200 42 4707
51 [2024-05-29 19:49:50] [2024-05-29 19:49:50 +0800] [1053] [INFO] 127.0.0.1:51776 GET /service/status 1.0 - - 5297
52 [2024-05-29 19:50:15]
53 [2024-05-29 19:50:15] Could not create share link. Missing file: /opt/conda/envs/itrex/lib/python3.10/site-packages/gradio/frpc_linux_amd64
54 [2024-05-29 19:50:15]
55 [2024-05-29 19:50:15] Please check your internet connection. This can happen if your antivirus software blocks the download of this file.
56 [2024-05-29 19:50:15]
57 [2024-05-29 19:50:15] 1. Download this file: https://cdn-media.huggingface.co/frpc-gradio-0.2/frpc_linux_amd64
58 [2024-05-29 19:50:15] 2. Rename the downloaded file to: frpc_linux_amd64_v0.2
59 [2024-05-29 19:50:15] 3. Move the file to this location: /opt/conda/envs/itrex/lib/python3.10/site-packages/gradio
60
```

在 Gradio 搭建的界面上进行交互：



```
57 [2024-05-29 19:50:55] 2024-05-29 19:50:55,865 - matplotlib.font_manager - INFO - generated new font
58 [2024-05-29 19:53:11]
59 Batches: 0%|          | 0/1 [00:00<?, ?it/s]
60 Batches: 100%|██████████| 1/1 [00:00<00:00, 7.02it/s]
61 Batches: 100%|██████████| 1/1 [00:00<00:00, 7.00it/s]
62 [2024-05-29 19:53:11] 2024-05-29 19:53:11,826 - root - INFO - Chat with QA Agent.
63
```

经过推理后生成了回答：

```
55 [2024-05-29 19:50:55] 2024-05-29 19:50:55,697 - matplotlib.font_manager - WARNING - Matplotlib is building the font cache; this may take a
56 [2024-05-29 19:50:55] 2024-05-29 19:50:55,865 - matplotlib.font_manager - INFO - generated new fontManager
57 [2024-05-29 19:53:11]
58 Batches: 0%|          | 0/1 [00:00<?, ?it/s]
59 Batches: 100%|██████████| 1/1 [00:00<00:00, 7.02it/s]
60 Batches: 100%|██████████| 1/1 [00:00<00:00, 7.00it/s]
61 [2024-05-29 19:53:11] 2024-05-29 19:53:11,826 - root - INFO - Chat with QA Agent.
62 [2024-05-29 20:00:32] Based on your symptoms and the information available, it seems like you may be experiencing some form of mental heal
63
```



上线的项目还存在两个问题：

一是语言问题，这可能因为采用的数据集是英文数据集，也可能是 RAG 无法准确理解“挂科”推理 pipeline 配置的过程可能有一些问题；还有一个是 pipeline 没有配置模型优化，这主要

是因为免费云环境只有 16GB 内存，担心会爆内存。

Anyway，这是一个使用 Intel Extension for Transformers 轻松搭建聊天机器人应用的成功实践，可以看到 Intel Extension for Transformers 在 Intel CPU 上起到的加速效果和工程实践上的便利之处。这个项目未来可以调整的地方还有很多，例如数据集的调整和进一步发掘 Intel Extension for Transformers 的能力，在 Intel 设备上更简单高效地搭建个人大模型定制服务。