链滴

如何在矩池云上安装语音识别模型 Whisper

作者：matpool
原文链接：https://ld246.com/article/1669365138519
来源网站：链滴

如何在矩池云上安装语音识别模型 Whisper

Whisper 是 OpenAI 近期开源的一个语音识别的模型，研究人员基于 680,000 小时的标记音频数据行训练，它同时也是一个多任务模型，可以进行多语言语音识别以及语音翻译任务，可以将语音音频录为所讲语言的文本，以及翻译成英语文本。

查看论文：https://cdn.openai.com/papers/whisper.pdf

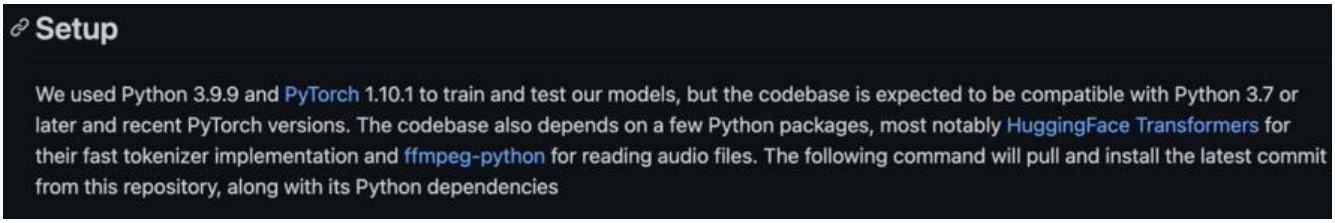
开源代码：https://github.com/openai/whisper

Whisper 的训练数据中65%为英语音频和相匹配的转录文本，大约18%为非英语音频和英语转录文本17% 为非英语音频和相应语言的转录文本。非英语的数据中包含了98种不同的语言，而某一特定语中的性能与所采用这一语言的训练的数据量直接相关，如在英语语音的识别中，模型已接近人类水平鲁棒性和准确性。

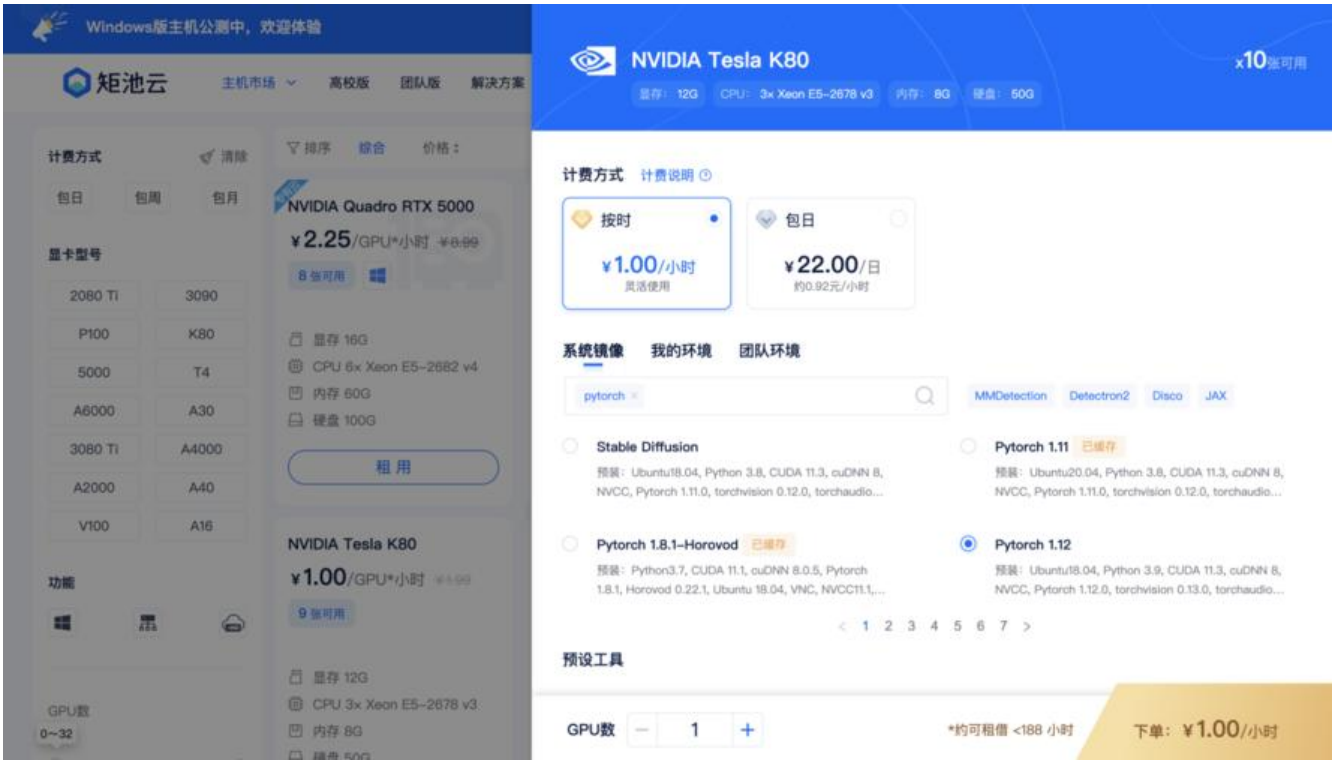
矩池云安装 Whisper 过程

环境配置&租用机器

在 Whisper 的 Setup 中，我们可以看到所需要的都是 Python 3.9.9和 PyTorch1.10.1，同时也兼容新的版本。



打开矩池云-主机市场，在此我们选中 K80 进行尝试，根据 Setup 可以选择 Pytorch 1.12系统镜像点击下单。

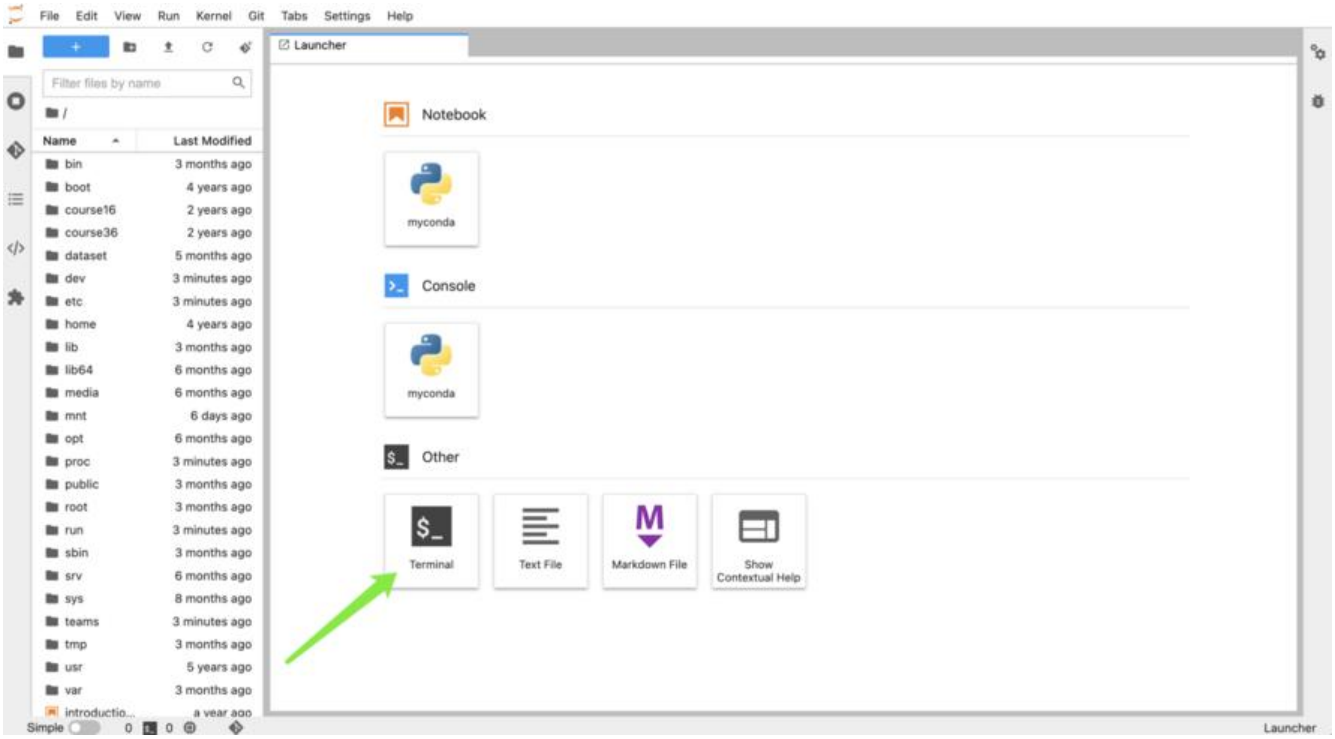


原文链接：如何在矩池云上安装语音识别模型 Whisper

运行后，点击 JupyterLab，进而“点击打开”。

下载代码&模型

进入页面后，点击 Terminal

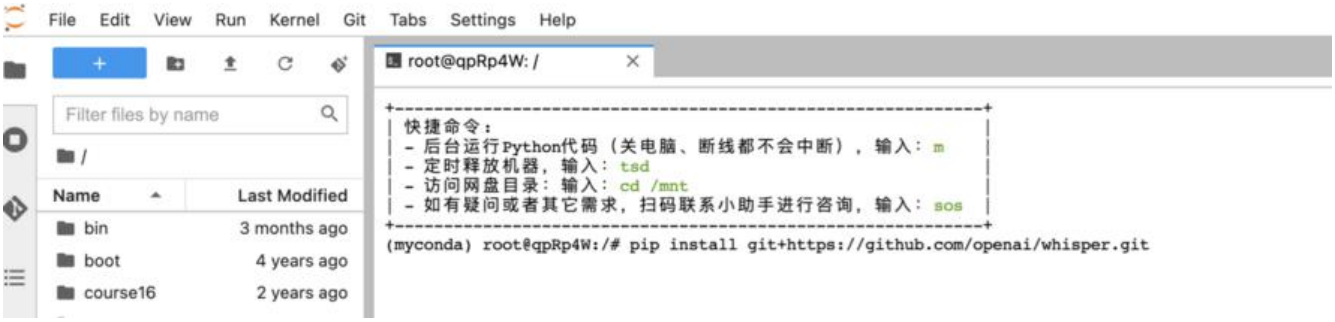


打开 Terminal 后，输入以下代码

```
pip install git+https://github.com/openai/whisper.git
```

如下

原文链接：如何在矩池云上安装语音识别模型 Whisper

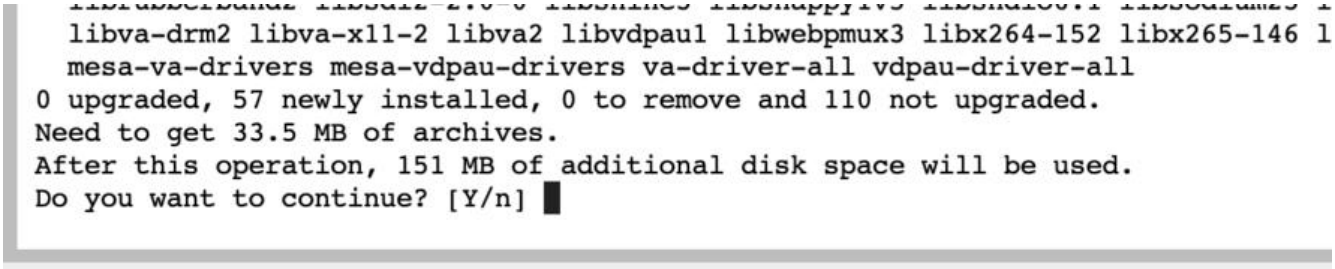


安装成功后，页面提示 successfully installed...

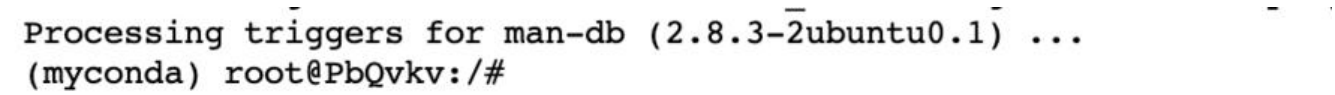
如果系统中没有安装过 ffmpeg，还需输入以下内容进行安装

sudo apt update && sudo apt install ffmpeg

安装过程中会提示 是否继续，输入 y ，回车即可



安装完成后，状态如下

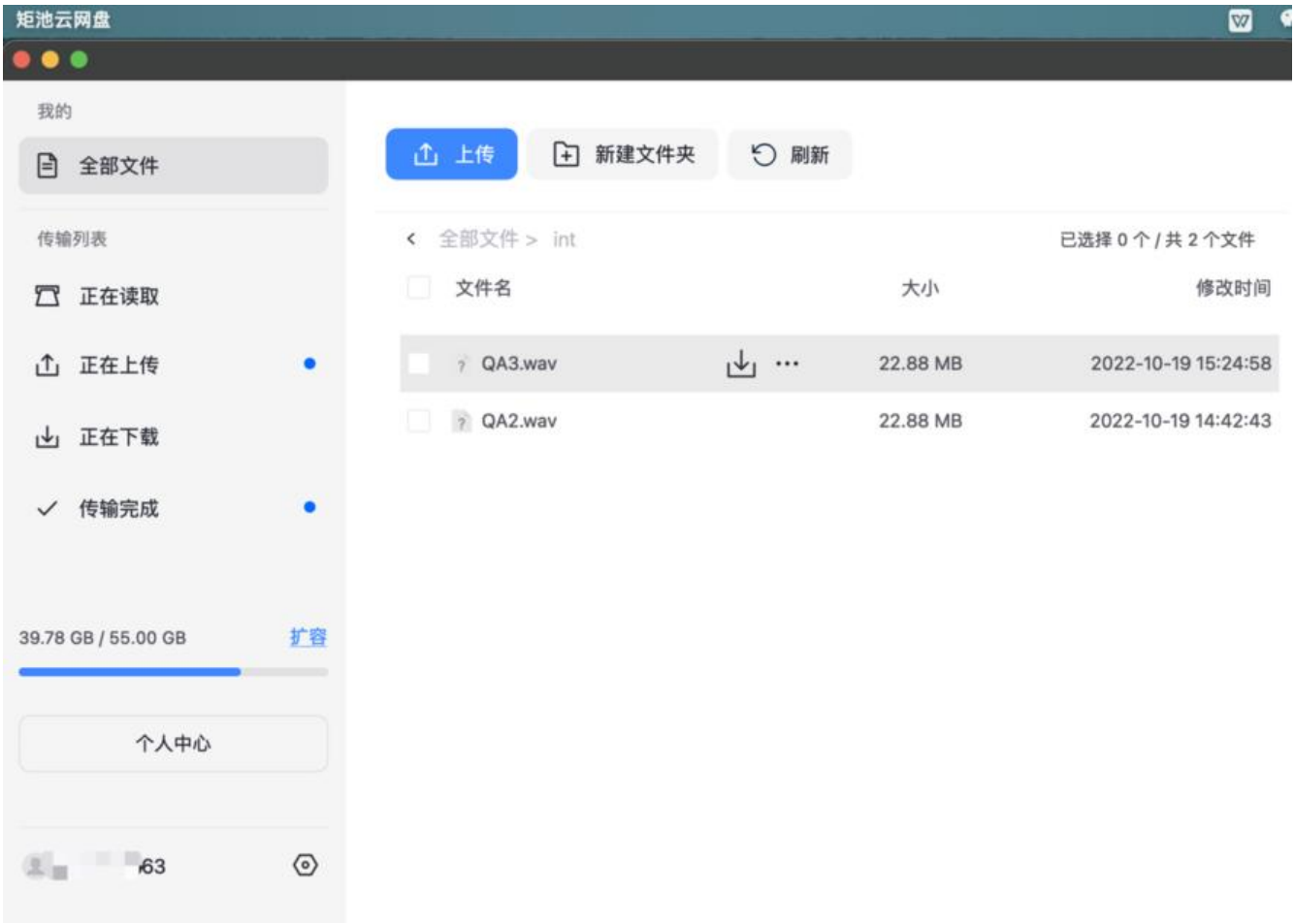


使用 Whisper 进行转录

准备文件

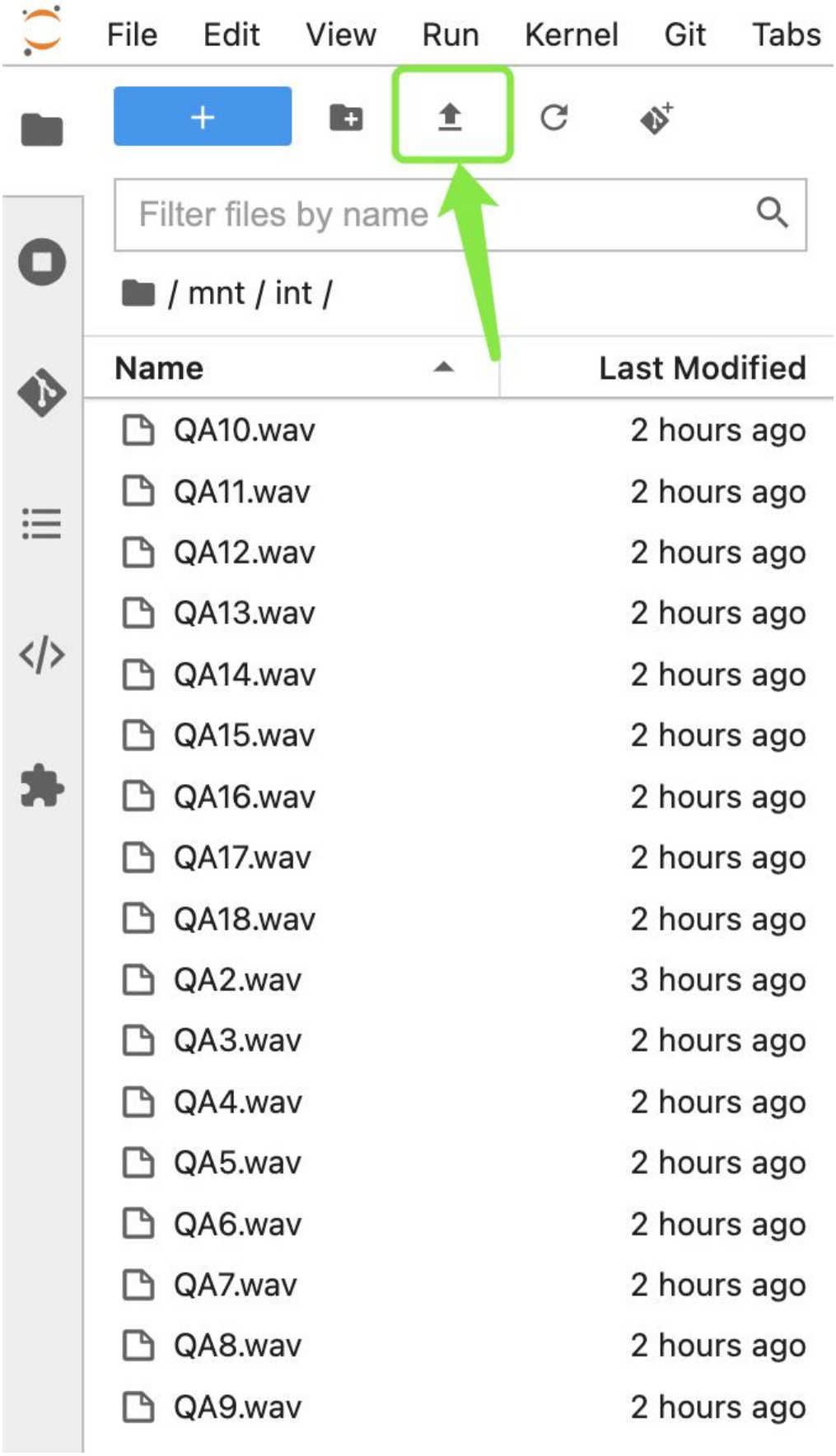
方法1: 通过矩池云网盘客户端上传文件

打开网盘客户端，可以点击上传，选择文件，或者直接将文件拖拽到客户端界面。



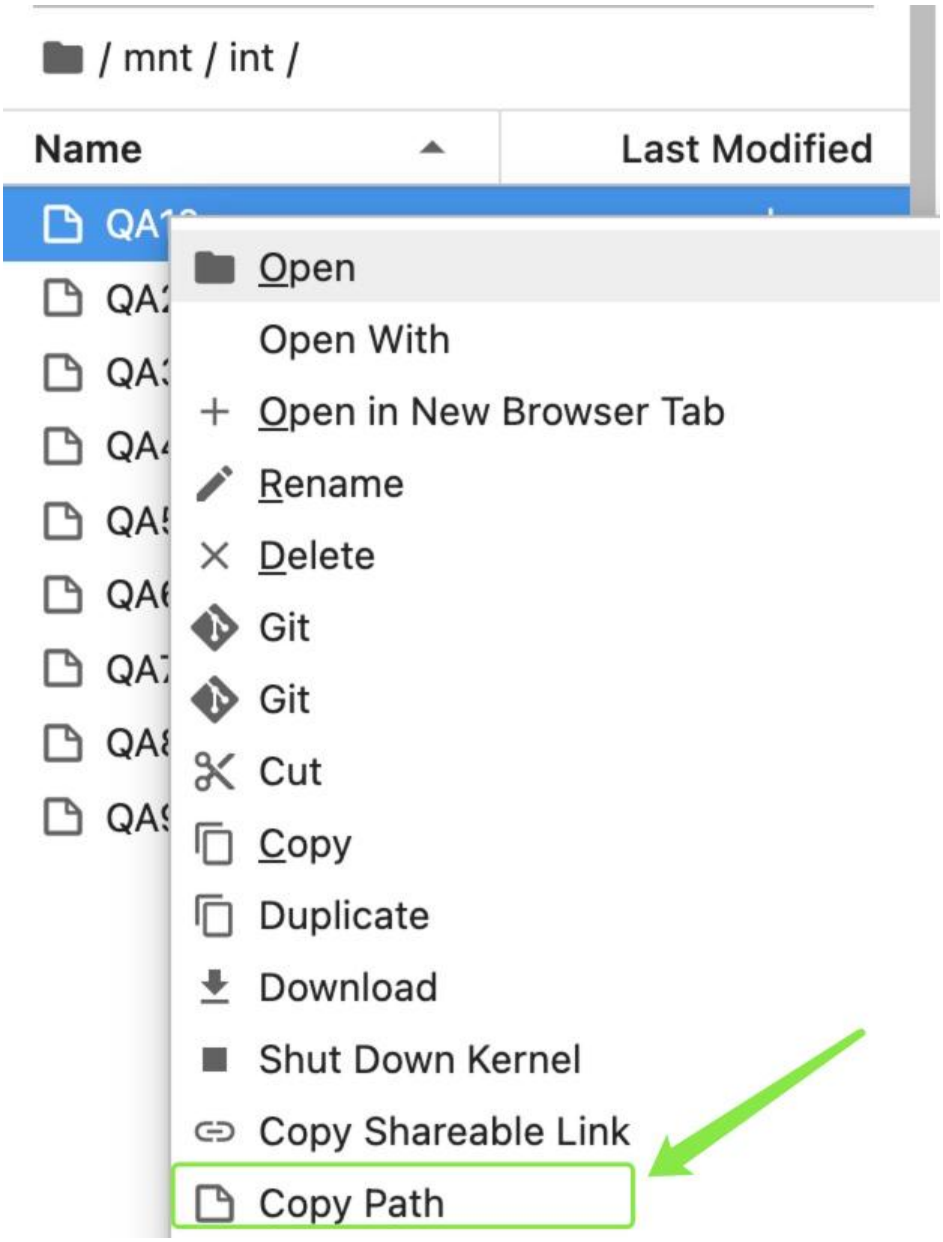
方法2:通过 JupyterLab 上传文件

在页面上点击，进入到/mnt，可以直接将音频文件在此进行上传。（此处我们自建了一个文件夹，大可以根据需要来进行操作）


File
Edit
View
Run
Kernel
Git
Tabs
Filter files by name
/ mnt / int /
Name
Last Modified
QA10.wav 2 hours ago
QA11.wav 2 hours ago
QA12.wav 2 hours ago
QA13.wav 2 hours ago
QA14.wav 2 hours ago
QA15.wav 2 hours ago
QA16.wav 2 hours ago
QA17.wav 2 hours ago
QA18.wav 2 hours ago
QA2.wav 3 hours ago
QA3.wav 2 hours ago
QA4.wav 2 hours ago
QA5.wav 2 hours ago
QA6.wav 2 hours ago
QA7.wav 2 hours ago
QA8.wav 2 hours ago
QA9.wav 2 hours ago

进行转录/翻译

Whisper 在默认条件下，输入音频是进行转录的 我们以转录为例子，在左侧文件夹，复制文件路径，



输入 whisper 路径如下

whisper mnt/int/QA2.wav

在不对其他项目进行设定的情况下，系统会自动检测语言，进而进行转录，

```
(myconda) root@PbQvkv:/# whisper mnt/int/QA2.wav
100%|                                       | 461M/461M [18:52<00:00, 427kiB/s]
Detecting language using up to the first 30 seconds. Use `--language` to specify the language
Detected language: English
```

以吴恩达 Andrew Ng 和 AI 科学家Christine Payne 的一次访谈 的音频转录为例，系统将自动生成间轴和文字内容，如下。

```
(myconda) root@ZWEXb8:/# whisper mnt/feifei.m4a --task translate
Detecting language using up to the first 30 seconds. Use `--language` to specify the language
Detected language: English
[00:00.000 --> 00:06.960]  So I'm really curious, when you think about the future of AI in terms of applications,
[00:09.600 --> 00:16.560]  you've mentioned medicine as one of the big ones. Are there others that are maybe as close to your
[00:16.560 --> 00:24.080]  heart or maybe at least as important in your mind as AI for medicine that you'll see happen in the
[00:24.080 --> 00:31.360]  next, I don't know, five, 10, 20 years? Yeah, Peter. So this is where I feel insecure answering
[00:31.360 --> 00:40.320]  because I'm talking to the world's greatest roboticist. So it's closer to your world. I am
[00:40.320 --> 00:47.680]  extremely excited by the world of robotics. As you said that 20 years ago, the world of robotics
[00:47.680 --> 00:56.640]  and AI were far further apart, right, that the tool sets and the problems we work on. But now
[00:56.640 --> 01:03.680]  that with the emergence of deep learning, machine learning, reinforcement learning,
[01:03.680 --> 01:09.840]  as well as I think as the maturation of computer vision, natural language processing,
[01:09.840 --> 01:17.120]  all this converged to a, in my opinion, very quickly going to be a watershed moment for
[01:17.120 --> 01:27.120]  robotic AI and robotic learning. And what that, what I feel extremely excited by is that that
[01:27.120 --> 01:35.760]  will fundamentally change the landscape of human labor. Of course, it's a very, very nuanced topic
[01:35.760 --> 01:41.840]  because human labor is about jobs, is about the livelihood of people, you know, even self-driving
[01:41.840 --> 01:51.440]  car is driving us pun intended into a deeper discussion of, you know, job changes of truck
[01:51.440 --> 02:01.680]  drivers and taxi drivers. But in the meantime, there's just so much productivity to be unleashed
[02:01.680 --> 02:13.280]  in human world that can make work safer, more efficient and more collaborative and possibly
[02:13.280 --> 02:22.640]  even break the boundaries of physical distance, thanks to robots. So I think that the imagination
[02:22.640 --> 02:34.000]  that's enabled by the future of robotics is just, it does excite me, excite me as a researcher,
[02:34.000 --> 02:46.080]  and they excite me as thinking about what world we can possibly imagine with the advent of robots.
[02:46.080 --> 02:53.680]  It also excites me because I think, you know, Peter, you and I have a role to play to ensure
[02:54.480 --> 03:01.920]  this will be going to the direction we want it, because it's so profound. It impacts people's
[03:04.240 --> 03:14.400]  work, people's dignity, people's agency, that technologies, technologists like us
[03:14.400 --> 03:21.360]  who understand this technology deeply and are making this technology have an incredibly important
[03:21.360 --> 03:44.720]  role to play with the rest of the society to make sure this will go where we would like it to go.
(myconda) root@ZWEXb8:/#
```

同时，在默认文件夹还会生成 srt txt vtt三种格式的文件，以方便使用者在不同情境下调用，指定文夹也可以通过指令 --output_dir 进行指定。

```
(myconda) root@ZWEXb8:/# whisper mnt/feifei.m4a --task translate
Detecting language using up to the first 30 seconds. Use `--language` to specify the language
Detected language: English
[00:00.000 --> 00:06.960]  So I'm really curious, when you think about the future of AI in terms of applications,
[00:09.600 --> 00:16.560]  you've mentioned medicine as one of the big ones. Are there others that are maybe as close to your
[00:16.560 --> 00:24.080]  heart or maybe at least as important in your mind as AI for medicine that you'll see happen in the
[00:24.080 --> 00:31.360]  next, I don't know, five, 10, 20 years? Yeah, Peter. So this is where I feel insecure answering
[00:31.360 --> 00:40.320]  because I'm talking to the world's greatest roboticist. So it's closer to your world. I am
[00:40.320 --> 00:47.680]  extremely excited by the world of robotics. As you said that 20 years ago, the world of robotics
[00:47.680 --> 00:56.640]  and AI were far further apart, right, that the tool sets and the problems we work on. But now
[00:56.640 --> 01:03.680]  that with the emergence of deep learning, machine learning, reinforcement learning,
[01:03.680 --> 01:09.840]  as well as I think as the maturation of computer vision, natural language processing,
[01:09.840 --> 01:17.120]  all this converged to a, in my opinion, very quickly going to be a watershed moment for
[01:17.120 --> 01:27.120]  robotic AI and robotic learning. And what that, what I feel extremely excited by is that that
[01:27.120 --> 01:35.760]  will fundamentally change the landscape of human labor. Of course, it's a very, very nuanced topic
[01:35.760 --> 01:41.840]  because human labor is about jobs, is about the livelihood of people, you know, even self-driving
[01:41.840 --> 01:51.440]  car is driving us pun intended into a deeper discussion of, you know, job changes of truck
[01:51.440 --> 02:01.680]  drivers and taxi drivers. But in the meantime, there's just so much productivity to be unleashed
[02:01.680 --> 02:13.280]  in human world that can make work safer, more efficient and more collaborative and possibly
[02:13.280 --> 02:22.640]  even break the boundaries of physical distance, thanks to robots. So I think that the imagination
[02:22.640 --> 02:34.000]  that's enabled by the future of robotics is just, it does excite me, excite me as a researcher,
[02:34.000 --> 02:46.080]  and they excite me as thinking about what world we can possibly imagine with the advent of robots.
[02:46.080 --> 02:53.680]  It also excites me because I think, you know, Peter, you and I have a role to play to ensure
[02:54.480 --> 03:01.920]  this will be going to the direction we want it, because it's so profound. It impacts people's
[03:04.240 --> 03:14.400]  work, people's dignity, people's agency, that technologies, technologists like us
[03:14.400 --> 03:21.360]  who understand this technology deeply and are making this technology have an incredibly important
[03:21.360 --> 03:44.720]  role to play with the rest of the society to make sure this will go where we would like it to go.
(myconda) root@ZWEXb8:/#
```

针对于多个文件，处理方式为直接将多个文件路径放置于 whisper 之后，即可逐个进行处理。

```
(myconda) root@PbQvkv:/# whisper mnt/int/QA5.wav mnt/int/QA6.wav mnt/int/QA7.wav mnt/int/QA8.wav mnt/int/QA9.wav mnt/int/QA10.wav mnt/in
t/QA11.wav mnt/int/QA12.wav mnt/int/QA13.wav mnt/int/QA14.wav mnt/int/QA15.wav mnt/int/QA16.wav mnt/int/QA17.wav mnt/int/QA18.wav
```

参数解析

Whisper 指定运行参数方式为：whisper 音频路径 --具体任务。 在 whisper 中，更多可运行参数如：

参数名	描述	默认值
[--model {tiny.en,tiny,base.en,base,small.en,small,medium.en,medium,large}]	-model 模型类型 从小到大的不同模型，分别为tiny.en,tiny,base.en,base,small.en,small,medium.e,medium,large	
[--model_dir MODEL_DIR]	储模型文件的路径	~/.cache/whisper

[--device DEVICE] 使用Pytorch的设备（CPU or GPU） CUDA

[--output_dir OUTPUT_DIR] output_dir 保存输出的路径 None

[--verbose VERBOSE] 是否打印过程和debg信息 True

[--task {transcribe,translate}] [--task {transcribe,translate}] --task 任务：是否执行 X->X 语音识别 ('transcribe') 或 X->英文翻译 ('translate') transcribe

[--language {af,am,ar,as,az,ba,be,bg,bn,bo,br,bs,ca,cs,cy,da,de,el,en,s,et,eu,fa,fi,fo,fr,gl,gu,ha,haw,hi,hr,ht,hu,hy,id,is,it,iw,ja,jw,ka,kk,km,kn,ko,la,lb,ln,lo,lt,lv,mg,mi,mkml,mn,mr,ms,mt,my,ne,nl,nn,no,oc,pa,pl,ps,pt,ro,ru,sa,sd,si,sk,sl,sn,so,sq,sr,su,sv,sw,ta,te,tg,th,tktl,tr,tt,uk,ur,uz,vi,yi,yo,zh,Afrikaans,Albanian,Amharic,Arabic,Armenian,Assamese,Azerbaijani,Bahkir,Basque,Belarusian,Bengali,Bosnian,Breton,Bulgarian,Burmese,Castilian,Catalan,Chinese,Cratian,Czech,Danish,Dutch,English,Estonian,Faroese,Finnish,Flemish,French,Galician,Georgian,Grman,Greek,Gujarati,Haitian,Haitian Creole,Hausa,Hawaiian,Hebrew,Hindi,Hungarian,Icelandic,ndonesian,Italian,Japanese,Javanese,Kannada,Kazakh,Khmer,Korean,Lao,Latin,Latvian,Letzebugesch,Lingala,Lithuanian,Luxembourgish,Macedonian,Malagasy,Malay,Malayalam,Maltese,Mari,Marathi,Moldavian,Moldovan,Mongolian,Myanmar,Nepali,Norwegian,Nynorsk,Occitan,Panjbi,Pashto,Persian,Polish,Portuguese,Punjabi,Pushto,Romanian,Russian,Sanskrit,Serbian,Shona,Sndhi,Sinhala,Sinhalese,Slovak,Slovenian,Somali,Spanish,Sundanese,Swahili,Swedish,Tagalog,Tajk,Tamil,Tatar,Telugu,Thai,Tibetan,Turkish,Turkmen,Ukrainian,Urdu,Uzbek,Valencian,Vietnamese Welsh,Yiddish,Yoruba}] --language 语：原音频中使用的语言

[--temperature TEMPERATURE] -temperature 温度参数：文章使用的是基于温度系数的采样，这个参数就是采样的温度系数

[--best_of BEST_OF] 温度非0时的抽样使用的候选词数 5

[--beam_size BEAM_SIZE] eam搜索中的beam数据的数目，仅在温度为0时可用 5

[--patience PATIENCE] beam解码是使的可选耐性系数
optional patience value to use in beam decoding, as in https://arxiv.org/bs/2204.05424, the default (1.0) is equivalent to conventional beam search (default: None) one

[--length_penalty LENGTH_PENALTY] length_penalty 惩罚系数：用于正则化的 optional token length penalty coefficient (alpha) as in ttps://arxiv.org/abs/1609.08144, uses simple length normalization by default (default: None) br />可选的惩罚系数\\alpha None

[--suppress_tokens SUPPRESS_TOKENS] 样期间要抑制的token ID的逗号分隔列表；“-1”时将抑制大多数特殊字符（常用标点符号除外） 1

[--initial_prompt INITIAL_PROMPT] 选文本，作为第一个窗口的提示。 None

[--condition_on_previous_ext CONDITION_ON_PREVIOUS_EXT] --condition_on_revious_text 先前文本使用状况：如果为 True，则提供模型的先前输出作为下一个窗口的提示； 禁可能会使文本跨窗口不一致，但模型变得不太容易陷入故障

[--fp16 FP16] 在fp16中进行推理
rue

[--temperature_increment_on_
allback TEMPERATURE_INCREMENT_
N_FALLBACK] --temperature_in
rement_on_fallback 回退温度系数：当解码未能满足以下任一阈值时的回退增加的温度
.2

[--compression_ratio_threshold COMPR
SSION_RATIO_THRESHOLD]
ompression_ratio_threshold 压缩率阈值：如果gzip
缩比高于这个值，则认为解码失败 2.4

[--logprob_threshold LOGPROB_THRE
HOLD] 如果平均对数概率低于此值，则将解码视为失败
1.0

[--no_speech_threshold NO_
PEECH_THRESHOLD] --no_s
eech_threshold 静音阈值：如果 <|nospeech|
标记的概率高于此值，并且解码由于“logprob_threshold”而失败，则将该段视为静音
.6

[--threads THREADS] 使用Pytorch CPU
推理时，使用的CPU线程数 0

保存环境，下次直接调用镜像

如果使用比较顺利，希望下次可以直接启动已经安装好的 Whisper 的镜像，可以在此处“保存到个环境”，如果是团队共享，则可以“保存到团队环境”

如果已经矩池云微信公众号上绑定过账户，则在手机上同时会收到保存环境成功的提醒。

保存环境后，下次使用该环境，可以直接在“我的环境”中迅速打开，无需再重复进行上一次的设置

原文链接：如何在矩池云上安装语音识别模型 Whisper

优势和局限性

我们针对一段在 CVPR 2022 会议上一段技术音频同时使用 Youtube 生成的字幕与 Whisper 生成的幕进行了比对。

句子完整性更好

Whisper 能按照speaker语气停顿断句，断句后有的甚至影响了精准性 vs 不破坏句子完整性，保持话轮、原语义群；

Youtube自动生成字幕

```
00:00:24,080 --> 00:00:25,840
who is a research scientist with google

12
00:00:25,840 --> 00:00:27,359
play
```

Whisper 转录

```
5
00:00:23,440 --> 00:00:27,800
Ruchi Gau, who is a Research Scientist with Google Brain.

6
00:00:27.800 --> 00:00:32.040
```

精准度更高

Whisper 在精准度上确实比较高，比如如下这个例子。

Youtube自动生成字幕

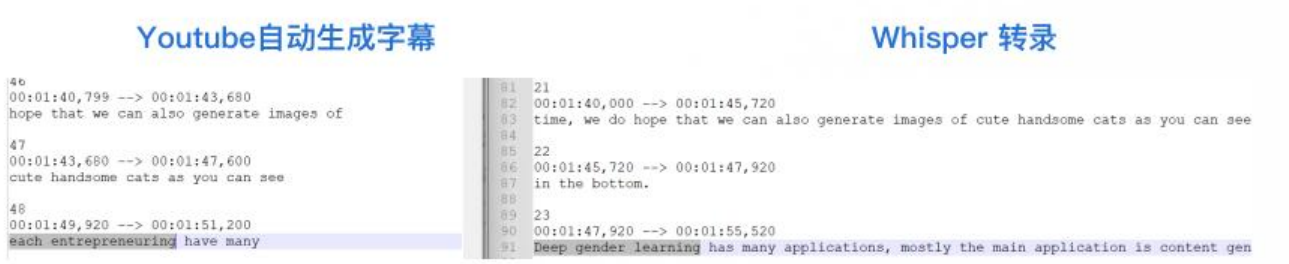
```
20
00:00:43,120 --> 00:00:44,960
after the conference and we'd like to

21
00:00:44,960 --> 00:00:46,480
share this friday with the research

22
00:00:46,480 --> 00:00:48,480
committee through youtube
```

Whisper 转录

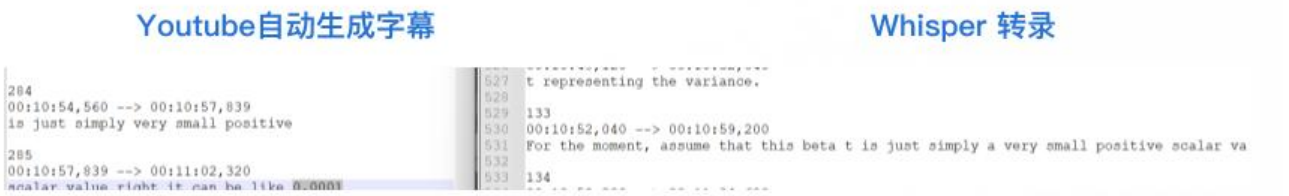
```
9
00:00:40,520 --> 00:00:45,000
we decided to record our presentation again after the conference and we would like to

10
00:00:45,000 --> 00:00:48,680
share this broadly with the research community through YouTube.

11
00:00:48,680 --> 00:00:52,720
If you happen to watch this video and you enjoy this video, we would like to encourage
```

这种精准度，同时体现在弱语气/低语调的插入语/状语的处理结果更优，如下。

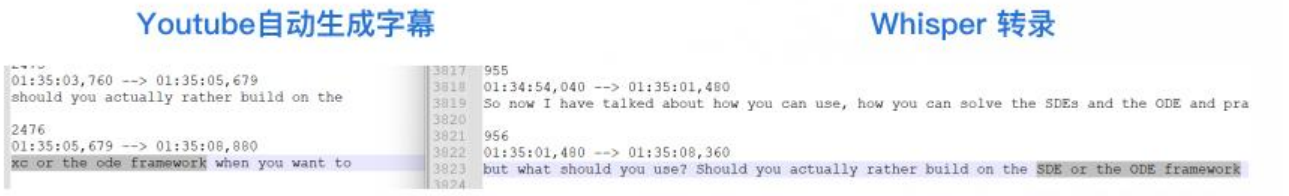
原文链接：如何在矩池云上安装语音识别模型 Whisper



在数字方面，精准度似乎也更胜一筹。



更重要的是，我们发现一些专业术语的转录方面，Whisper 也呈现出更精准的状态。

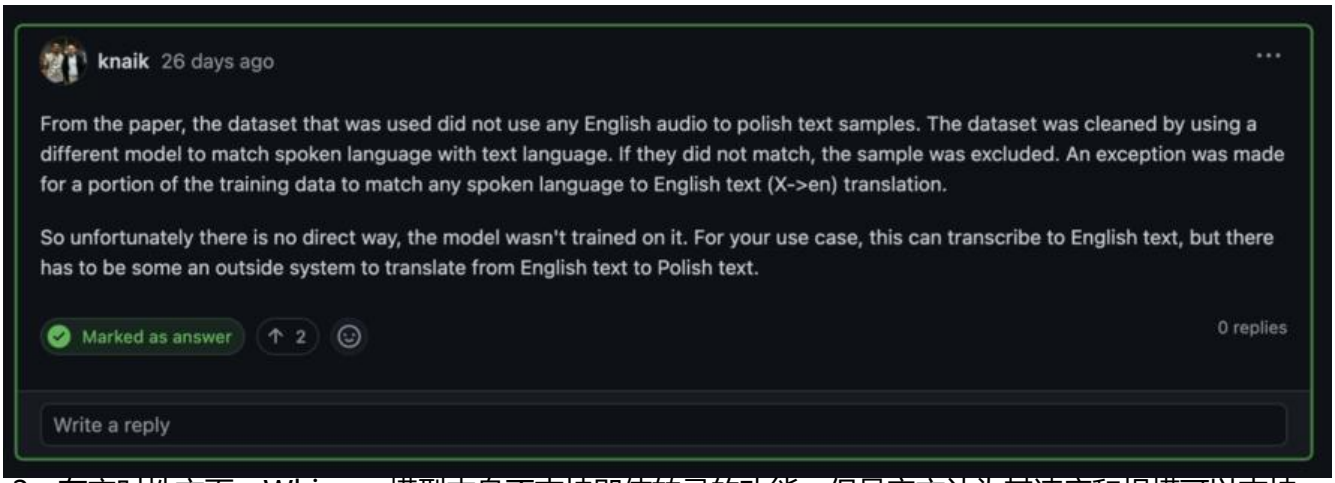


当然，以上并具有统计学意义，只是我们在做尝试的时候发现的一些 Whisper 优秀之处。

局限性

当然，Whisper 也有其局限性，我们也汇集了一些如下情形。

1、目前 Whisper 模型只能对语音识别后，转换为对应语言的文本，或将其翻译为英语，则意味着在译这一层面，最终无法实现由英语转换为其他语言，在这一方面，其他模型在多语言方面可能去的了多的进展；



2、在实时性方面，Whisper 模型本身不支持即使转录的功能，但是官方认为其速度和规模可以支持时转译，但仍需在此基础上进行二次开发； 3、如输入的音频中为多语言混合，Whisper 对于这种情

原文链接：如何在矩池云上安装语音识别模型 Whisper

也暂无解决方案； 4、此外，对于环境音比较嘈杂的情况（比如有噪音，或者有背景音乐），如不设具体的 temperature ，有一定可能转录结果会有所不同，所以如有这种情况可以进行设置，关于 Temperature 的一些信息可以参考https://algowriting.medium.com/gpt-3-temperature-setting-101-1200ff0d0be。