

亚马逊云科技



中国峰会

2025年6月19日-20日 上海·世博中心

EDU 208

上海交大获奖作品

大模型虚拟教师项目分享

陈欣怡

上海交通大学
巴黎卓越工程师学院
2024级

涂瑞峰

上海交通大学
信息与电子工程学院|集成电路学院
2024级

房郁超

上海交通大学
计算机学院
2024级

Contents

01

项目背景与动机

应用于人工智能教学场景的教师系统

02

项目展示

全方位展示系统的各种功能

03

技术与细节

RAG、多模态、用户友好.....

04

亚马逊云科技赋能

Amazon Bedrock, Amazon Polly, Amazon Transcribe, etc.

项目背景与动机

第十届“新生杯” 大模型虚拟教师赛道 决赛圆满落幕



本次比赛中，各个队伍需利用亚马逊云科技提供的基础模型，通过**提示词工程**、**RAG**等技术来构建自己的**虚拟教师模型**。对于每支队伍，需要首先抽取需要回答的题目，接着在四分钟内展示各自的项目成果，再在三分钟内回答抽取的题目，如有剩余时间，则由评委提出问题，选手解答。

在传统的教育模式中，“一对多”的格局

精力不足、个性化程度低



大模型虚拟教师

项目展示



新聊天

聊天 1



- 蓝色的海水
- 岩石海岸
- 模糊的植被背景

4. 考虑可能的鸟类：
- 根据特征，这可能是一只蓝脚鲹鸟

5. 描述图片的整体感觉：
- 动态的飞行姿势
 - 自然环境的美丽背景

6. 组织信息，准备回答

整理后的回答：
我将详细描述这张图片，包括鸟的特征、飞行姿态和背景环境。

这张图片展示了一只大鸟在海岸线上空飞行的精彩瞬间。这只鸟很可能是一只蓝脚鲹鸟，它有以下特征：

1. 身体：主要是白色的，体型较大。
2. 翅膀：宽大的翅膀完全展开，呈现出令人印象深刻的飞行姿态。
3. 脚：最引人注目的是它鲜艳的蓝色脚，这是蓝脚鲹鸟的标志性特征。
4. 飞行姿势：鸟儿正面朝向镜头，翅膀完全展开，脚部向后伸展，整体呈现出一个"十"字形。

背景环境： 1. 海水：下方可以看到蓝色的海水，有些波浪。 2. 海岸：远处可以看到岩石海岸线。 3. 植被：背景中隐约可见一些绿色，可能是沿海的植被。

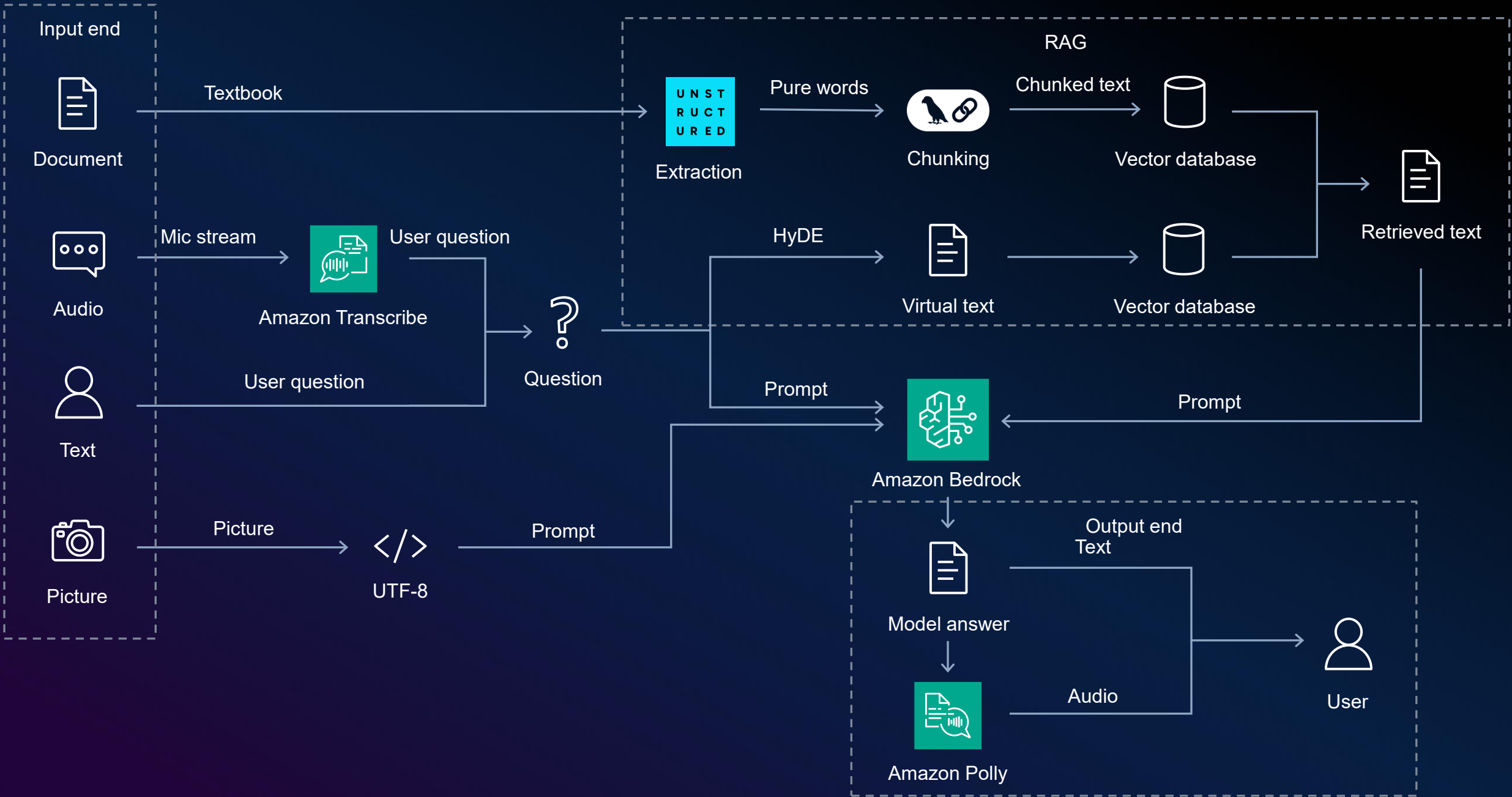
这张照片捕捉到了鸟儿飞行的动态美，展现了自然界的生命力和优雅。蓝脚鲹鸟优雅的飞行姿态，配合着蓝天碧海的背景，构成了一幅生动而富有活力的自然画面。摄影师捕捉到这一瞬间的技巧令人赞叹。 😊



请输入您的消息...



技术与细节



RAG

RAG

Multi-Modal

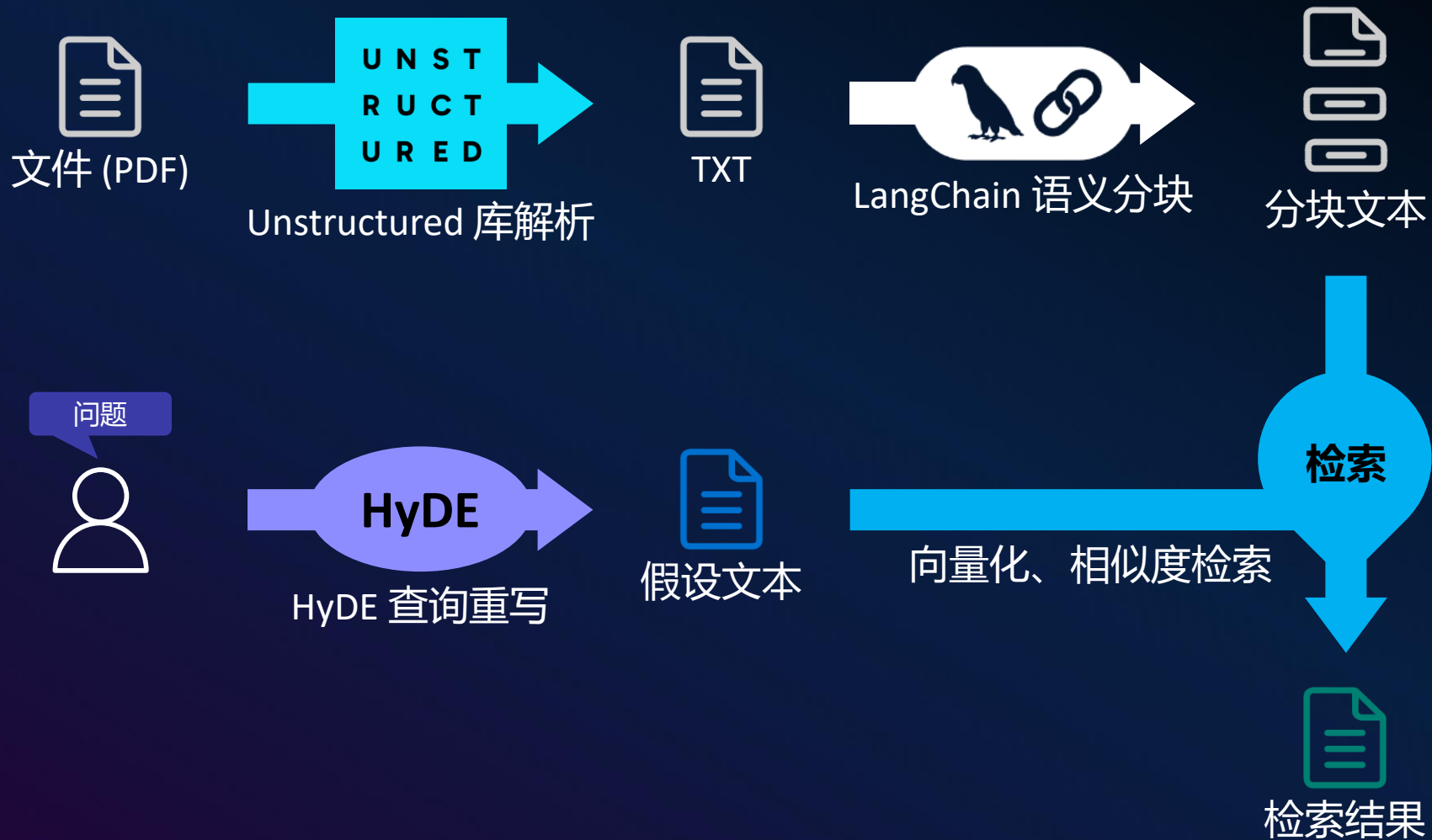
Input end

多模态

Output end

User-Friendly

页面设计



Unstructured 库精准解析

按照不同内容类型对应的标签分
成多个元素 (element)



使用元素坐标分析算法，解决传统
OCR 识别文字时双栏乱序问题



使用 html 标签标注表格类型元素，
最大程度复原图表内容

word based only on its context. Unlike left-to-right language model pre-training, the MLM objective enables the representation to fuse the left and the right context, which allows us to pre-train a deep bidirectional Transformer. In addition to the masked language model, we also use a “next sentence prediction” task that jointly pre-trains text-pair representations. The contributions of our paper are as follows:

- We demonstrate the importance of bidirectional pre-training for language representations. Unlike Radford et al. (2018), which uses unidirectional language models for pre-training, BERT uses masked language models to enable pre-trained deep bidirectional representations. This is also in contrast to Peters et al. (2018a), which uses a shallow concatenation of independently trained left-to-right and right-to-left LMs.
- We show that pre-trained representations reduce the need for many heavily-engineered task-specific architectures. BERT is the first fine-tuning based representation model that achieves state-of-the-art performance on a large suite of sentence-level and token-level tasks, outperforming many task-specific architectures.
- BERT advances the state of the art for eleven NLP tasks. The code and pre-trained models are available at <https://github.com/google-research/bert>.

2 Related Work

There is a long history of pre-training general language representations, and we briefly review the most widely-used approaches in this section.

2.1 Unsupervised Feature-based Approaches

Learning widely applicable representations of words has been an active area of research for decades, including non-neural (Brown et al., 1992; Ando and Zhang, 2005; Blitzer et al., 2006) and neural (Mikolov et al., 2013; Pennington et al., 2014) methods. Pre-trained word embeddings are an integral part of modern NLP systems, offering significant improvements over embeddings learned from scratch (Turian et al., 2010). To pre-train word embedding vectors, left-to-right language modeling objectives have been used (Mnih and Hinton, 2009), as well as objectives to discriminate correct from incorrect words in left and right context (Mikolov et al., 2013).

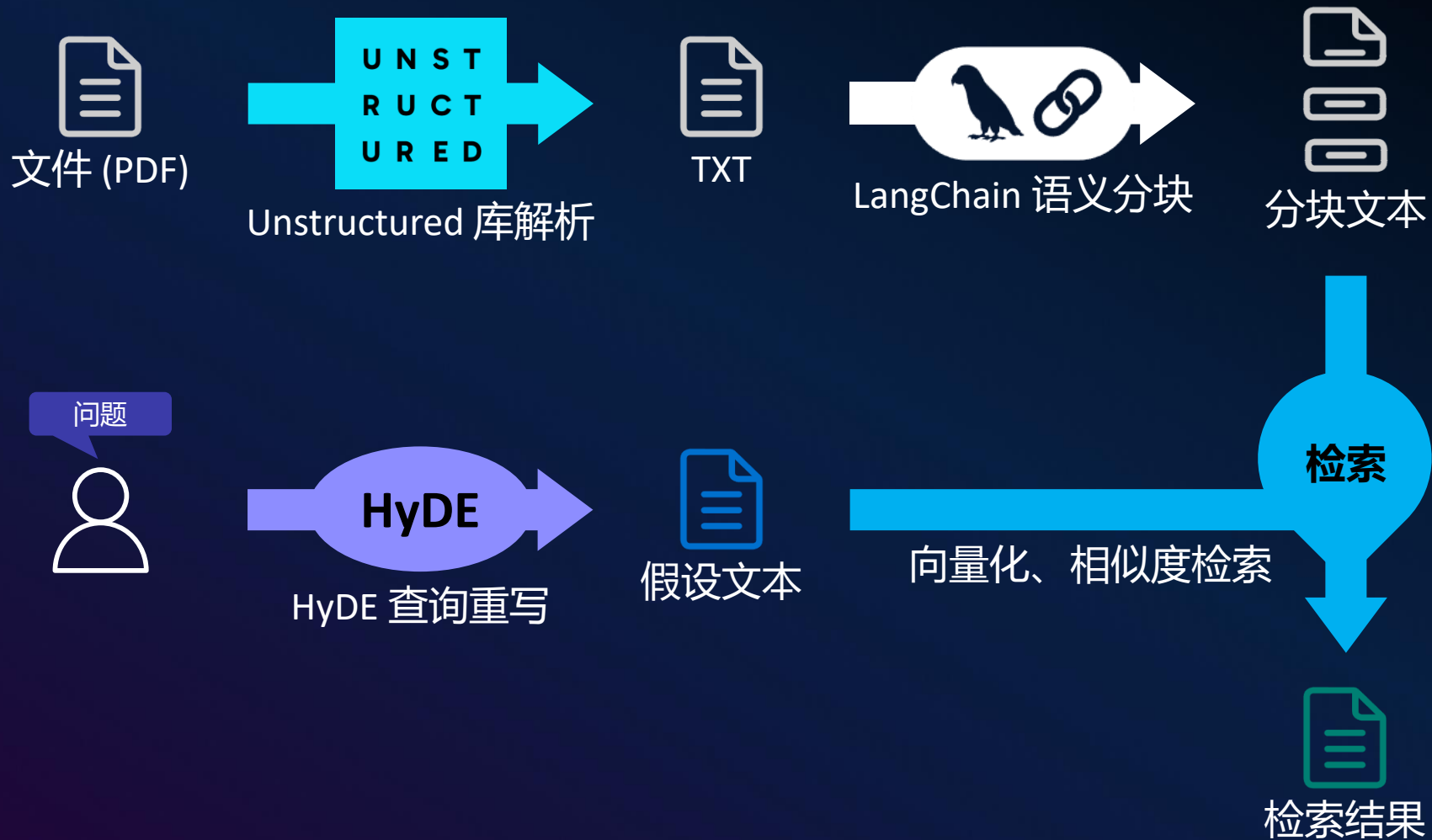
These approaches have been generalized to coarser granularities, such as sentence embeddings (Kiros et al., 2015; Logeswaran and Lee, 2018) or paragraph embeddings (Le and Mikolov, 2014). To train sentence representations, prior work has used objectives to rank candidate next sentences (Jernite et al., 2017; Logeswaran and Lee, 2018), left-to-right generation of next sentence words given a representation of the previous sentence (Liu et al., 2015), or denoising auto-encoder derived objectives (Hill et al., 2016).

ELMo, an ELMo’s predecessor (Peters et al., 2017, 2018a) generalizes traditional word embedding research along a different dimension. They extract *context-sensitive* features from a left-to-right and a right-to-left language model. The contextual representation of each token is the concatenation of the left-to-right and right-to-left representations. When integrating contextual word embeddings with existing task-specific architectures, ELMo advances the state of the art for several major NLP benchmarks (Peters et al., 2018a) including question answering (Rajpurkar et al., 2016), sentiment analysis (Socher et al., 2013), named entity recognition (Tjong Kim Sang and De Meulder, 2003). Melamud et al. (2016) proposed learning contextual representations through a task to predict a single word from both left and right context using LSTMs. Similar to ELMo, their model is feature-based and not deeply bidirectional. Fedus et al. (2018) shows that the cloze task can be used to improve the robustness of text generation models.

2.2 Unsupervised Fine-tuning Approaches

As with the feature-based approaches, the first works in this direction only pre-trained word embedding parameters from unlabeled text (Collobert and Weston, 2008).

More recently, sentence or document encoders which produce contextual token representations have been pre-trained from unlabeled text and fine-tuned for a supervised downstream task (Dai and Le, 2015; Howard and Ruder, 2018; Radford et al., 2018). The advantage of these approaches is that few parameters need to be learned from scratch. At least partly due to this advantage, OpenAI GPT (Radford et al., 2018) achieved previously state-of-the-art results on many sentence-level tasks from the GLUE benchmark (Wang et al., 2018a). Left-to-right language model-



Embedding Model、相似度检索

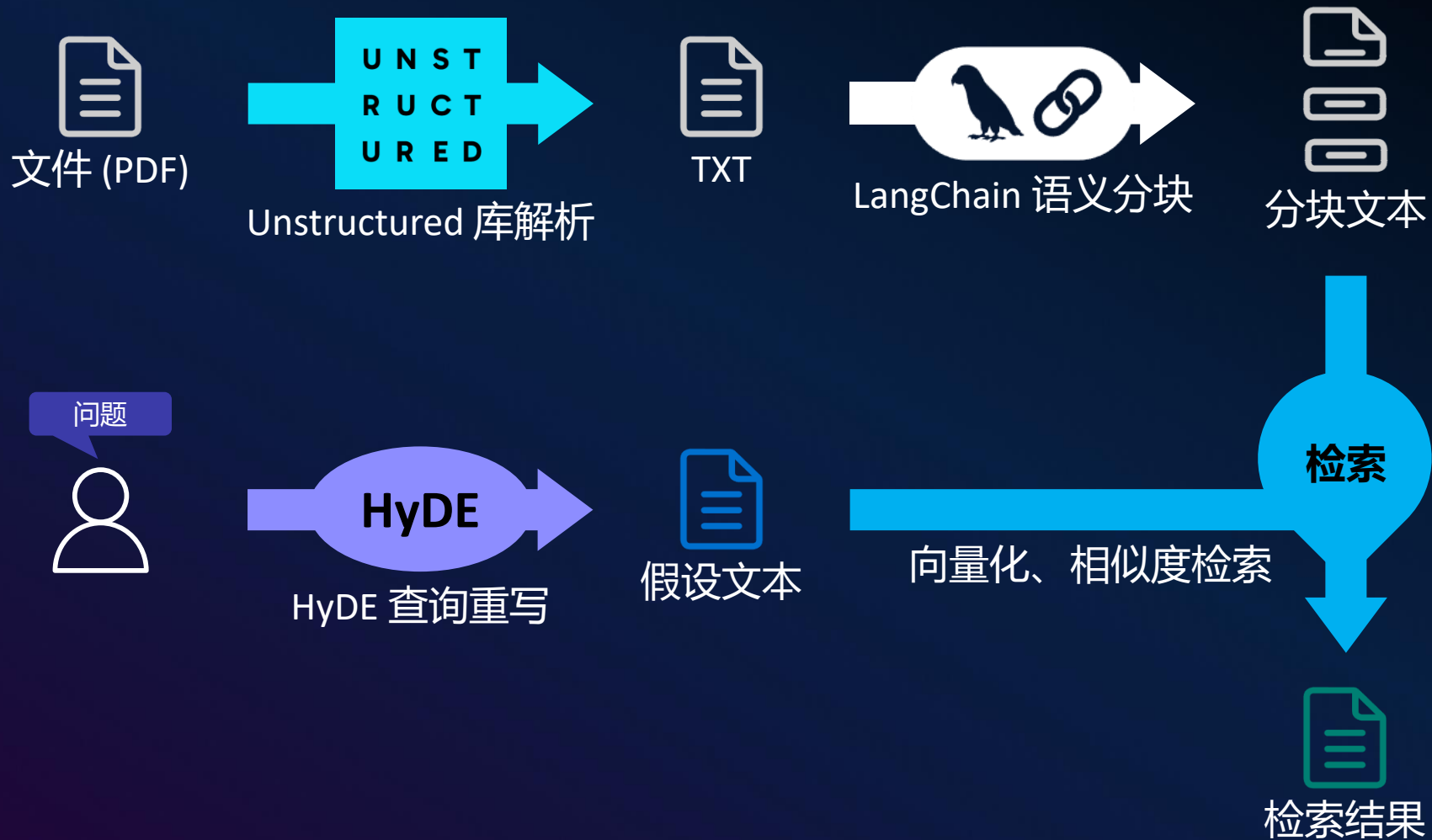
采用 sentence-transformer 加载 embedding model



向量化



相似度排名



RAG

RAG

Multi-Modal

Input end

多模态

Output end

User-Friendly

页面设计

虚拟教师系统输入端

文字



文件



语音



图像





语音输入

呈现出令人印象深刻的飞行姿态。
的蓝色脚，这是蓝脚鲹鸟的标志性特征。
头，翅膀完全展开，脚部向后伸展，整体呈现出一个“十”字形。
到蓝色的海水，有些波浪。2. 海岸：远处可以看到岩石海岸线。3. 植被：背景中隐约可见一些绿
美，展现了自然界的生命力和优雅。蓝脚鲹鸟优雅的飞行姿态，配合着蓝天碧海的背景，构成了一
摄影师捕捉到这一瞬间的技巧令人赞叹。😊

开始语音输入



General Settings

语音语言:

Temperature:

使用音频:

Chinese
Chinese
English
Japanese
French

识别速度快

精确度高

会“变通”

虚拟教师系统输入端

文字



文件



语音



图像





图像输入

通过上传文件通道上传的图片会进行解码。解码后的内容将上传到所选模型请求体中的相应接口。



这张图片展示了一只大鸟在海岸线上空飞行的精彩瞬间。这只鸟很可能是一只蓝脚鲣鸟，它有以下特征：

1. 身体：主要是白色的，体型较大。
2. 翅膀：宽大的翅膀完全展开，呈现出令人印象深刻的飞行姿态。
3. 脚：最引人注目的是它鲜艳的蓝色脚，这是蓝脚鲣鸟的标志性特征。
4. 飞行姿势：鸟儿正面朝向镜头，翅膀完全展开，脚部向后伸展，整体呈现出一个"十"字形。

背景环境： 1. 海水：下方可以看到蓝色的海水，有些波浪。 2. 海岸：远处可以看到岩石海岸线。 3. 植被：背景中隐约可见一些绿色，可能是沿海的植被。

这张照片捕捉到了鸟儿飞行的动态美，展现了自然界的生命力和优雅。蓝脚鲣鸟优雅的飞行姿态，配合着蓝天碧海的背景，构成了一幅生动而富有活力的自然画面。摄影师捕捉到这一瞬间的技巧令人赞叹。 😊

虚拟教师系统输入端

文字



文件



语音



图像





文件输入

通过上传按钮上传 pdf 格式文件，系统会按照之前解析教材的方式，用 unstructured[pdf] 及相应算法进行解析。

设置

- General Settings
- 记忆管理
- 文件管理

已上传文件

[1] 20250509142758_bird_drink.png 删除

[2] 20250509142805_giraffe.png 删除

虚拟教师系统输入端

文字



文件



语音



图像



虚拟教师系统输出端

文字

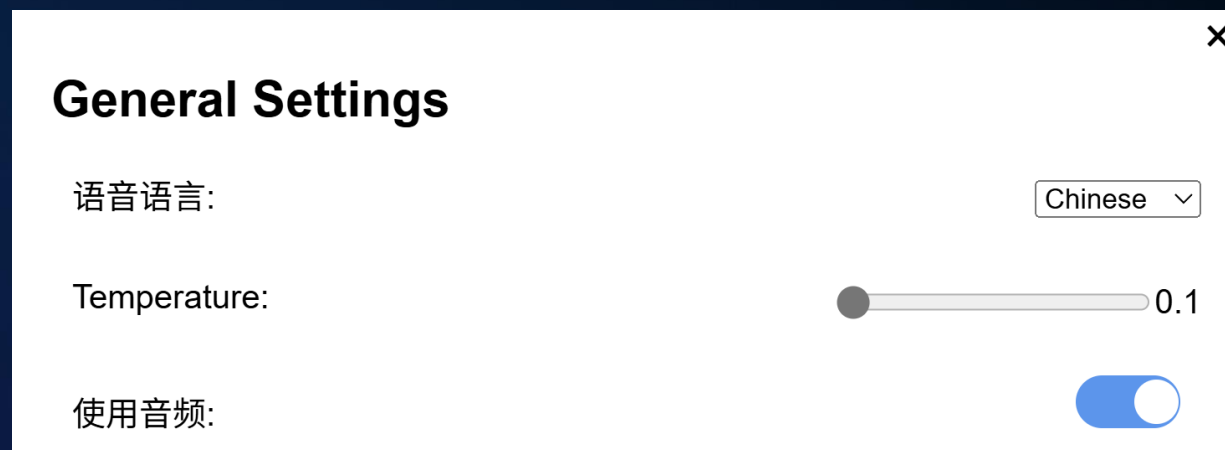


语音



语音输出

```
model.speak(audioUrl);
```

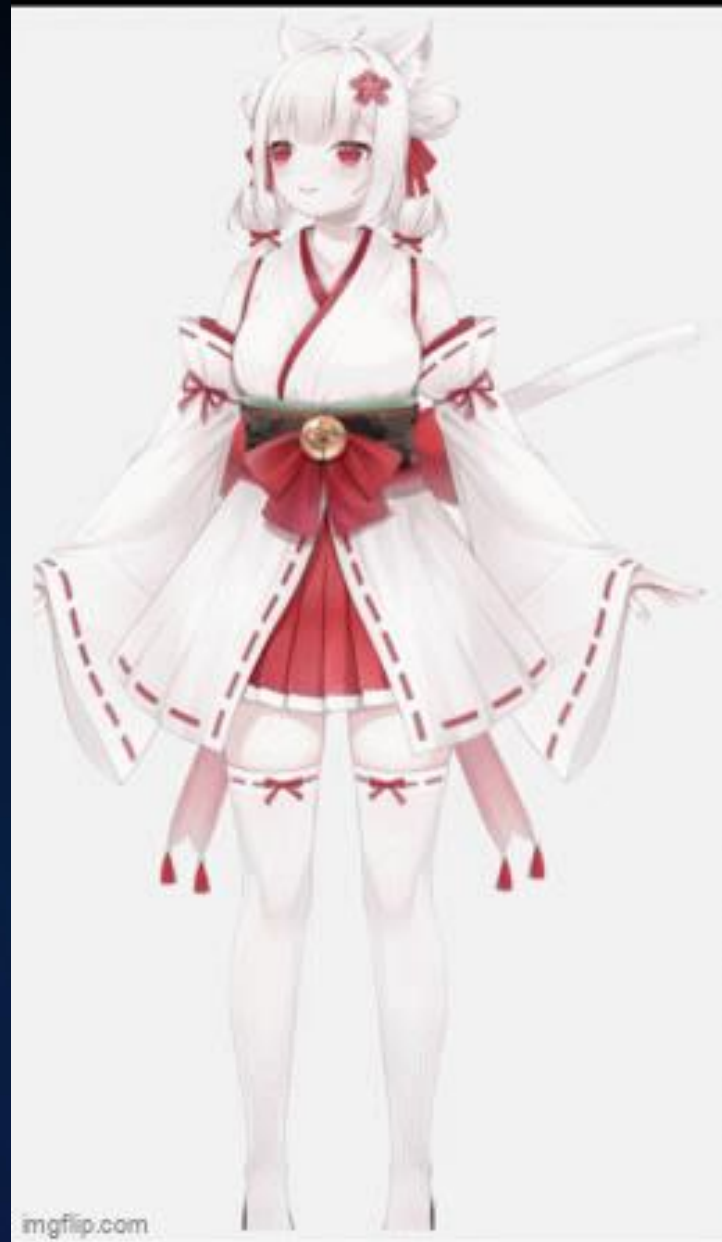




Live2d 虚拟形象

- 使用 model.speak 属性，实现语音和口型同步；
- 实现模型简单交互，模型眼神跟着鼠标移动。

简单易操作、渲染难度低、功能强大



虚拟教师系统输出端

文字



语音

RAG

RAG

Multi-Modal

Input end

多模态

Output end

User-Friendly

页面设计

过程信息显示

可选择模型

管理全局记忆

功能开关

- 查看 API 生成信息，包括 tokens、stop-reason、stop-sequence
- 查看 RAG 检索结果

加性注意力通过引入可学习的参数和



Stop Reason: end_turn
Stop Sequence: none
Output Tokens: 815

RAG检索结果

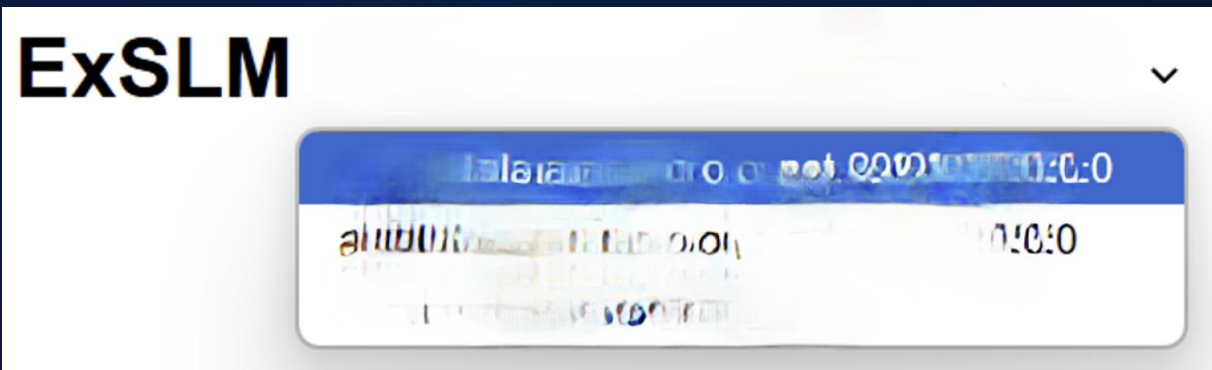
你当前正在编写的代码的哪些部分可以“学习”，即通过学习和自动确定代码中所做的设计选择来改进？你的代码是否包含启发式设计选择？2. 你遇到的哪些问题有许多解决它们的样本，但没有具体的自动化方法？这些可能是使用深度学习的主要候选者。3. 如果把人工智能的发展看作一场新的工业革命，那么算法和数据之间的关系是什么？它类似于蒸汽机和煤吗？根本区别是什么？4. 你还可以在哪里应用端到端的训练方法，比如图1.1.2、物理、工程和计量经济学？Discussions35 35 <https://discuss.d2l.ai/t/1744> 38 1. 引言要学习深度学习，首先需要先掌握一些基本技能。所有机器学习方法都涉及从数据中提取信息。因此，我们先学习一些关于数据的实用技能，包括存储、操作和预处理数据。机器学习通常需要处理大型数据集。我们可以将某些数据集视为一个表，其中表的行对应样本，列对应属性。线性代数为人们提供了一些用来处

过程信息显示

可选择模型

管理全局记忆

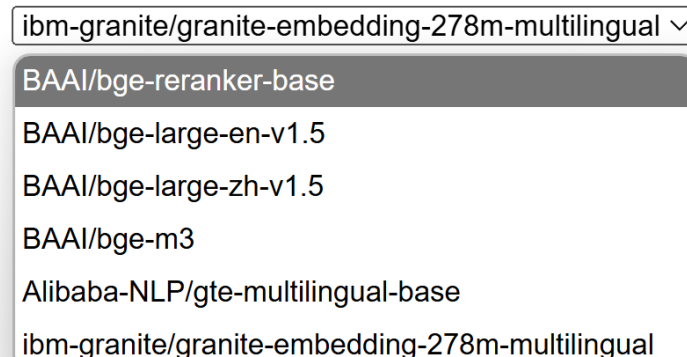
功能开关



**多种语言模型
来自 Amazon Bedrock**

多种 embedding 模型

嵌入模型（每次切换需要加载）：



过程信息显示 可选择模型 **管理全局记忆** 功能开关

打开 设置-记忆管理，可以手动输入用户的偏好、长期状态，提高模型个性化程度。

设置

General Settings

记忆管理

文件管理

记忆管理

添加记忆

[1] 用户希望模型回答的最后都加上表情符号

删除

添加

影响

- AlexNet在ImageNet比赛中的胜利
- 它推动了后续更深层次的神经网络研究
- AlexNet的成功也促进了深度学习在计算机视觉领域的广泛应用

AlexNet的出现不仅提高了图像分类的准确率，还为后续的研究铺平了道路。😊

过程信息显示 可选择模型 管理全局记忆 **功能开关**

设置

General Settings

记忆管理

文件管理

General Settings

语音语言:

Chinese

Temperature:

0.1

使用音频:

使用RAG（需重新加载模型）:

嵌入模型（每次切换需要加载）:

ibm-granite/granite-embedding-278m-multilingual

保存

RAG

RAG

Multi-Modal

Input end

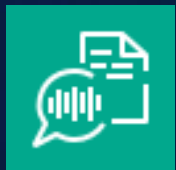
多模态

Output end

User-Friendly

页面设计

亚马逊云科技赋能

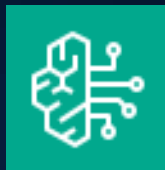


Amazon Transcribe

自动语音识别（ASR）服务

数十亿参数模型

支持 100+ 语言



Amazon Bedrock

一项完全托管的服务

100+ 热门新兴的专业 FM

安全性、隐私性



Amazon Polly

将任何文本转换为音频流

支持 40+ 种语言

提供 100+ 逼真声音



Amazon Transcribe



Amazon Bedrock



Amazon Polly

Thank you!

陈欣怡

上海交通大学
巴黎卓越工程师学院
2024级

涂瑞峰

上海交通大学
信息与电子工程学院|集成电路学院
2024级

房郁超

上海交通大学
计算机学院
2024级