



上海交通大學
SHANGHAI JIAO TONG UNIVERSITY

ExSLM

虚拟教师专题赛项目展示

By 示例队 (队伍号: 2)

Contents



上海交通大学
SHANGHAI JIAO TONG UNIVERSITY

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

RAG

利用多种技术与算法，提高检索能力

03

虚拟形象

使用Live2D模型，PixiJS+Three.js驱动

04

多模态

文字+语音+图像+文件

05

基础功能

调节temperature、新对话

06

我们的特色

手动添加全局记忆、丰富的模型选择.....



项目背景与动机

- AI 的迅猛发展为教育领域带来了前所未有的变革机遇。
- 我们的项目旨在探索大模型在 **人工智能教育领域** 的应用潜力，提供更加个性化、互动性强的学习体验，弥补传统在线教育的不足。



上海交通大學
SHANGHAI JIAO TONG UNIVERSITY

RAG

我们利用unstructured解析pdf，结合chunking、HyDE等技术，提高检索能力。

例题解答

在 注意力评分函数 这一节中，作者介绍了两种注意力评分函数，请问第二种注意力评分函数是什么？它的数学表达式是什么？请直接给出答案



ExSLM

softmax运 算得到的: $\alpha(q, k_i) = \text{softmax}$
聚操作。本节将介绍两个流行的评分函 数，稍后将用
1.1 掩蔽softmax操作
正如上面提到的，softmax
由集，某些文本序列被填充了没有意义的 特殊词元。为
然的
1. masked_softmax函数实现了这样的掩蔽softmax

基于这些信息，我可以给出缩放点积注意力的数学表达式：

s page)\n10.\n注意力机制\nQueries r0\n\n5\nKeys\n10.3.3 缩放点积注意力\n使用
所有元素都是独立的随机变量，并且都满足零均值和单位方差，那么两个向量的点积的均值为0，方差为
，则缩放 点积注意力 (scaled dot-product attention) 评分函数为: $\alpha(q, k) = q \cdot k / \sqrt{d_k}$
对计算注意力，其中查 询和键的长度为d，值的长度为v。查询 $Q \in R^{n \times d}$ 、键 $K \in R^{m \times d}$ 和值 $V \in R^{n \times v}$
力的实现使用了暂退法进行模型正则化。
#@save\nclass DotProductAttention(nn

$\sqrt{d_k}$

其中， q 是查询向量， k 是键向量， d_k 是键向量的维度。这个缩放因子 $\frac{1}{\sqrt{d_k}}$ 有助于防止在键向量维度较大时，点积的值变得过大，从而使 softmax 函数的梯度变得过小。😊😊😊😊



Unstructured库精准解析PDF

- unstructured库将PDF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
- 用html标签标注表格类型的元素，复原表格内容

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 (Kiros et al., 2015), or denoising auto-encoder derived objectives (Hill et al., 2016).

ELMo advances its 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). Melanud 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-

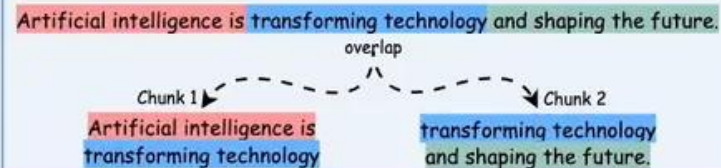
LangChain语义分块

- 利用 BAAI 的模型，按语义相关性动态调整chunk大小

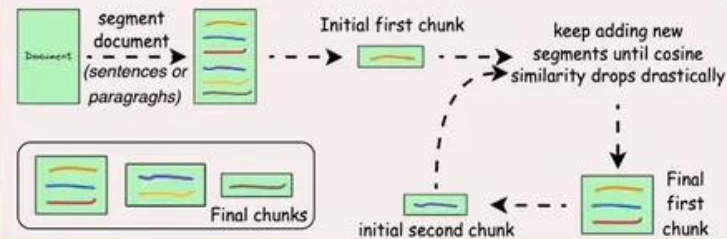
5 Chunking Strategies for RAG

blog.DailyDoseofDS.com

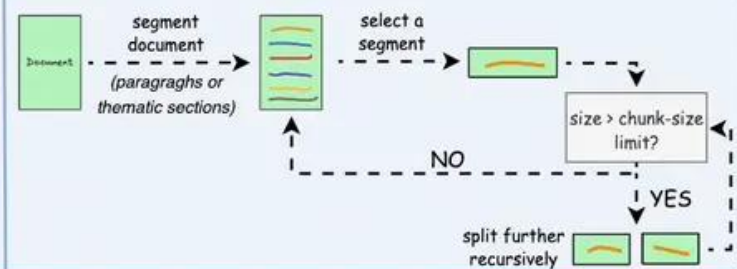
1) Fixed-size chunking



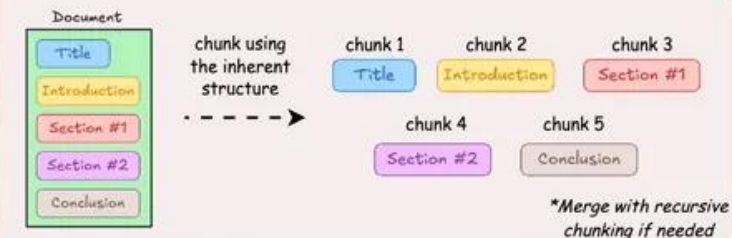
2) Semantic chunking



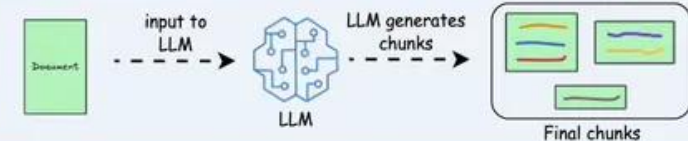
3) Recursive chunking



4) Document structure-based chunking

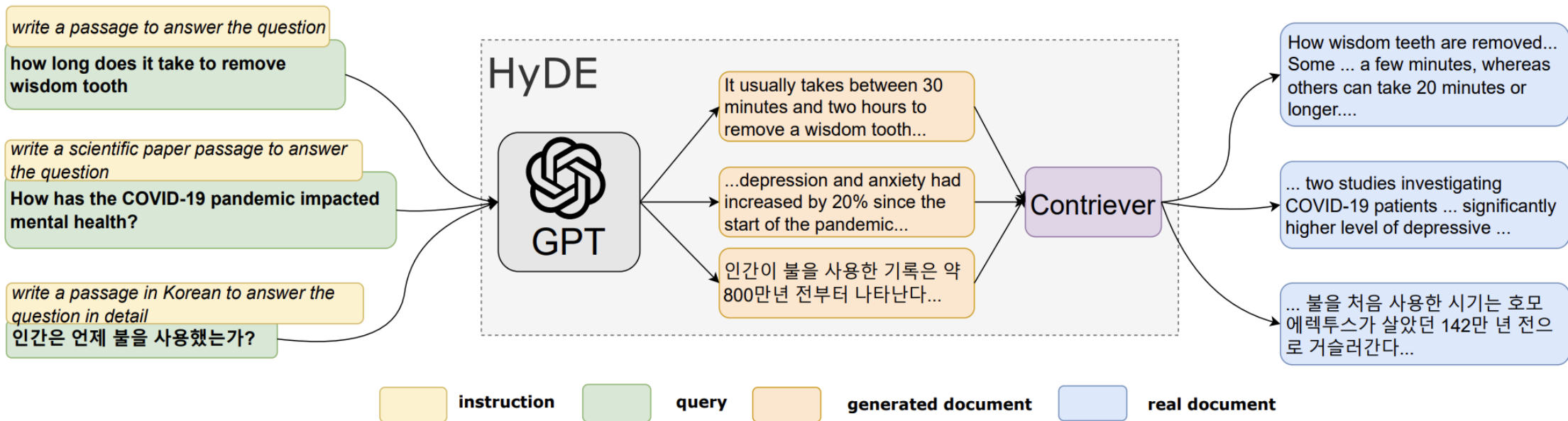


5) LLM-based chunking



HyDE查询重写

- 使用 HyDE (Hypothetical Document Embeddings) 技巧, 让AI先“想象”理想答案的样子, 再用这个想象结果去检索, 从而提高召回率。



多种Embedding模型选择

- 我们提供以下模型：
 - BAAI/bge-reranker-base （支持中英双语）
 - dunzhang/stella_en_400M_v5 （仅支持英文）
 - BAAI/bge-large-zh-v1.5 （仅支持中文）
 - BAAI/bge-m3 （Multilingual）
 - Alibaba-NLP/gte-multilingual-base （Multilingual）
 - ibm-granite/granite-embedding-278m-multilingual （Multilingual）

嵌入模型（每次切换需要加载）：

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

- BAAI/bge-reranker-base
- dunzhang/stella_en_400M_v5
- BAAI/bge-large-zh-v1.5
- BAAI/bge-m3
- Alibaba-NLP/gte-multilingual-base
- ibm-granite/granite-embedding-278m-multilingual

RAG状态提示

- RAG 由独立线程管理，根据用户设置，有以下三种状态，页面右上角会提示当前 RAG 加载状态。
 - RAG 加载中...
 - RAG 已就绪
 - RAG 未开启

RAG 正在加载中...



上海交通大學
SHANGHAI JIAO TONG UNIVERSITY

虚拟形象

我们使用Live2D虚拟形象技术，实现语音口型同步。

Live2D 模型

- 将模型文件接入前端，实现交互式虚拟形象
- **口型同步 (Lip Sync)**，Live2D 内置该技术，使用 `model.speak` 属性调用
- 开启音频后模型开始“说话”，支持语音打断





上海交通大學
SHANGHAI JIAO TONG UNIVERSITY

多模态交互

我们支持 语音+文本+视觉+文件 多通道输入，上传的文件可在设置中轻松管理。

多模态

我们的模型支持:

文字



文件



语音



图像



文字/语音模式、多语言交互、音频输出

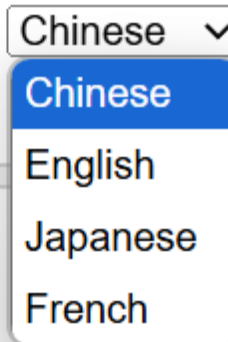
- 文字模式和语音模式自由切换，不影响对话内容和对话记录
- 两种模式均支持多语言交互（在 设置-语音语言 中选择）
- 在 设置- 使用音频 处开启语音输出

General Settings

语音语言:

Temperature:

使用音频:



文件管理

- 打开 设置-文件管理，查看已上传文件（文件名中添加了时间戳防止重名）。
- 一键删除

时间戳



图像识别 与 PDF 解析

- 我们调用 Claude-3.5 / Claude-3 图像识别API，可以读取图像内容信息，以及图片中的文本信息。
- 我们使用 unstructured 解析 PDF，与 RAG 处理 PDF 接口相同。



上海交通大學
SHANGHAI JIAO TONG UNIVERSITY

基础功能

调节temperature、新对话.....

调整 temperature

- 使用 设置-temperature 的滑块调整 temperature 值 (0.1~1)

三个臭皮匠，接下半句（没有标准答案）

temp = 0.1

temp = 1

可能的回答：

1. 三个臭皮匠，胜过一个CEO
2. 三个臭皮匠，抵得上一个AI
3. 三个臭皮匠，创造新奇迹
4. 三个臭皮匠，智慧无极限

最终选择第二个答案，因为它既现代又有趣，同时暗示了人类集体智慧的重要性。

三个臭皮匠，抵得上一个AI

1. 三个臭皮匠，胜过诸葛亮。
2. 三个臭皮匠，顶个诸葛亮。

这两个版本都强调了集体智慧的力量，意味着即使是普通人，只要团结起来，也能产生超越个人天才的智慧和能力。

新聊天、跨对话记忆机制

- 用户可以创建新聊天。
- 聊天记录将被保存在本地，每次打开都能重新渲染。
- 模型能在同一/跨对话中记住用户信息。





上海交通大學
SHANGHAI JIAO TONG UNIVERSITY

我们的特色

手动添加全局记忆、丰富的模型选择.....

一键复制、查看info

- 支持一键复制模型回复Markdown源码
- 查看 API 生成信息，包括 tokens、stop-reason、stop-sequence



ExSLM

思考:

理解问题:

用户要求我完成“三个臭皮匠”这个成语的下半句。这是一个中国的成语，通常用来表示集体智慧的力量。

思考过程:

1. 这个成语的标准下半句是“顶个诸葛亮”。
2. 但用户特别指出没有标准答案，这意味着我可以发挥创意。
3. 我需要考虑一些创新的、有趣的、或者现代化的方式来完成这个成语。
4. 同时，新的下半句应该保持原成语的精神，即强调团队合作的重要性。
5. 考虑到用户可能喜欢幽默或者新颖的表达。
6. 记得在回答的最后加上表情符号，以符合用户的偏好。

可能的回答:

1. 三个臭皮匠，胜过一个CEO
2. 三个臭皮匠，抵得上一个AI
3. 三个臭皮匠，创造新奇迹
4. 三个臭皮匠，智慧无极限

最终选择第二个答案，因为它既现代又有趣，同时暗示了人类集体智慧的重要性。

三个臭皮匠，抵得上一个AI

这个现代版的成语延续了原有的精神，强调了团队合作的力量。它暗示即使是普通人，只要团结一致，也能够产生令人惊叹的智慧和创造力，甚至可以与先进的人工智能相媲美。这个说法既幽默又发人深省，提醒我们不要低估人类集体智慧的潜力。😊😊😊😊😊😊



A 6x6 grid of dots. A 4x4 square is highlighted in the center, with a white border. The dots are arranged in a regular grid pattern.

- 

[illegible]

```
class BinaryTree {
private:
    TreeNode* root;
```

全局记忆管理

- 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

请输入要添加的记忆内容...

添加

接入多个模型 API

- 根据用户需求选择不同特点的大语言模型，为此我们接入了以下模型：
 - anthropic.claude-3-5-sonnet-20240620-v1:0
 - anthropic.claude-3-sonnet-20240229-v1:0
 - meta.llama3-70b-instruct-v1'

ExSLM

anthropic.claude-3-5-sonnet-20240620-v1:0 ▼

anthropic.claude-3-5-sonnet-20240620-v1:0

anthropic.claude-3-sonnet-20240229-v1:0

meta.llama3-70b-instruct-v1



上海交通大學
SHANGHAI JIAO TONG UNIVERSITY

谢谢!
Thanks