RAG应用与知识图谱

王楠
Jina AI联合创始人兼CTO

Berlin · Beijing · Shenzhen

Jina AI

- 全球化分布式办公
- 开源软件
- 专注AI开发工具

RAG是为数不多的GenAI落地场景

幻觉

- 基于检索结果
- 保证可解释性和可回溯性

回答可以验证追溯

知识更新成本高

- 更新检索知识库
- 支持增删改查

知识可以频繁更新

私有知识注入难

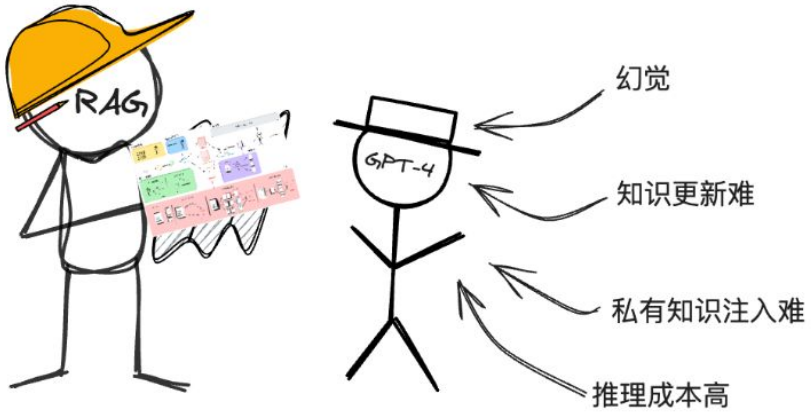
- 支持本地部署
- 本地存储私有数据
- 不需要微调模型

私有数据安全

推理成本高

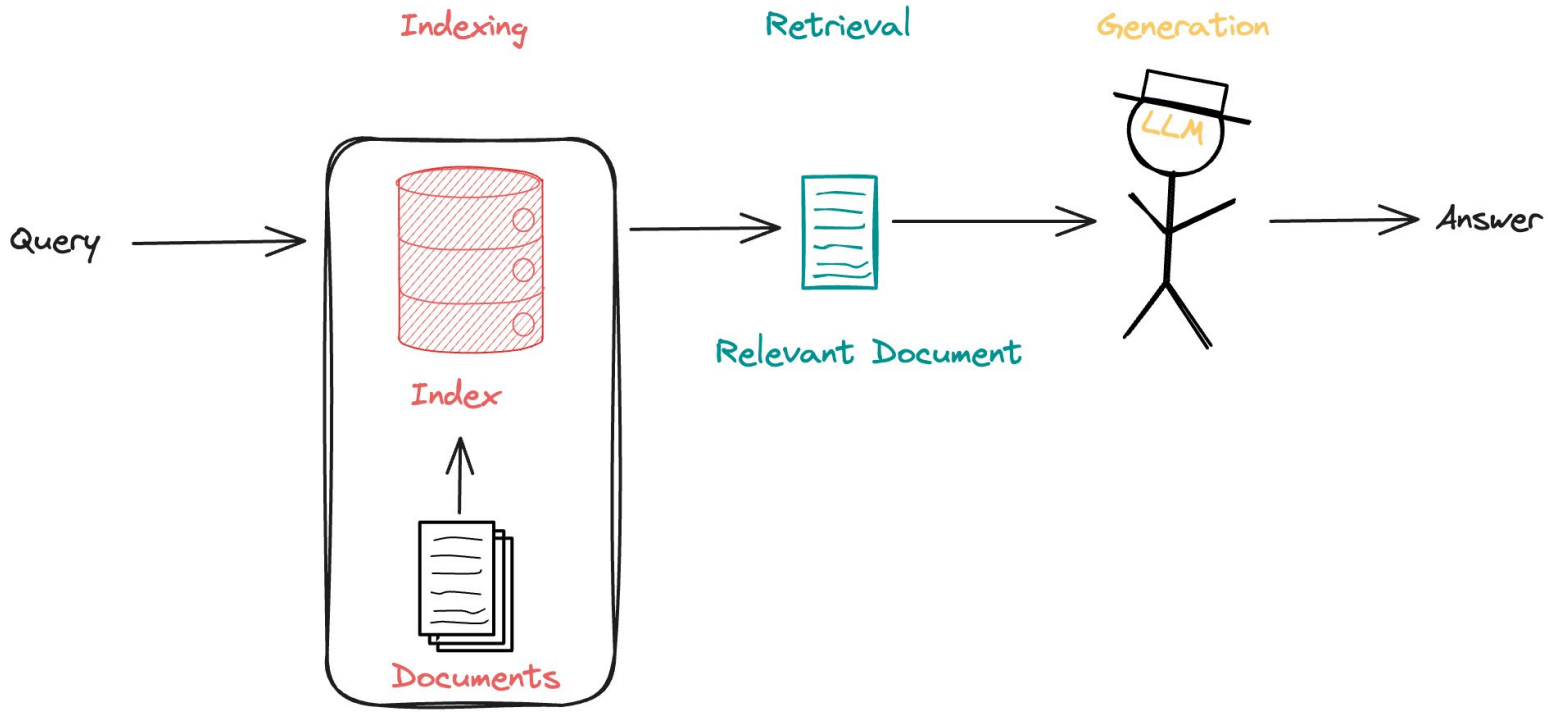
- 推理成本低
- 减少LLM输入，降低LLM推理成本

有效降低LLM成本



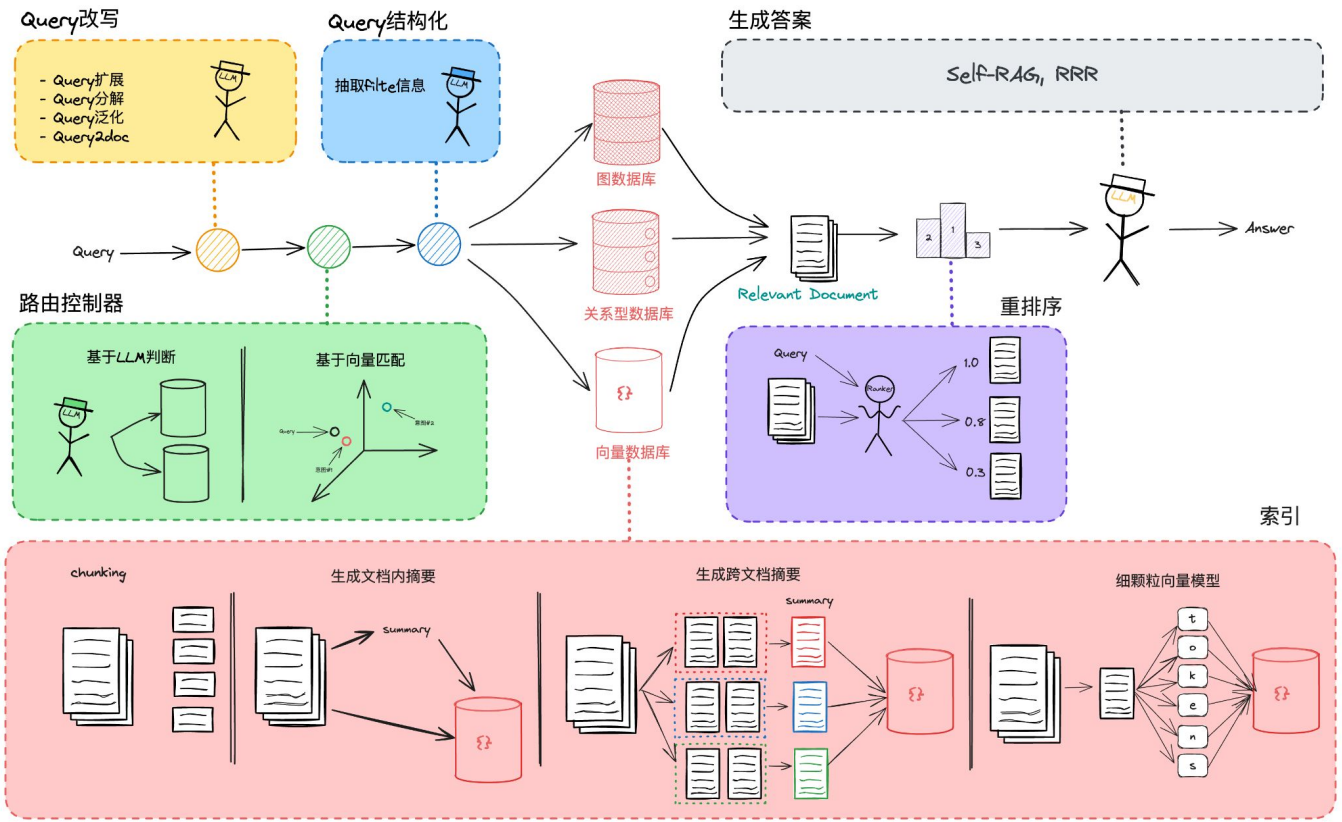
RAG非常复杂

POC的RAG流水线

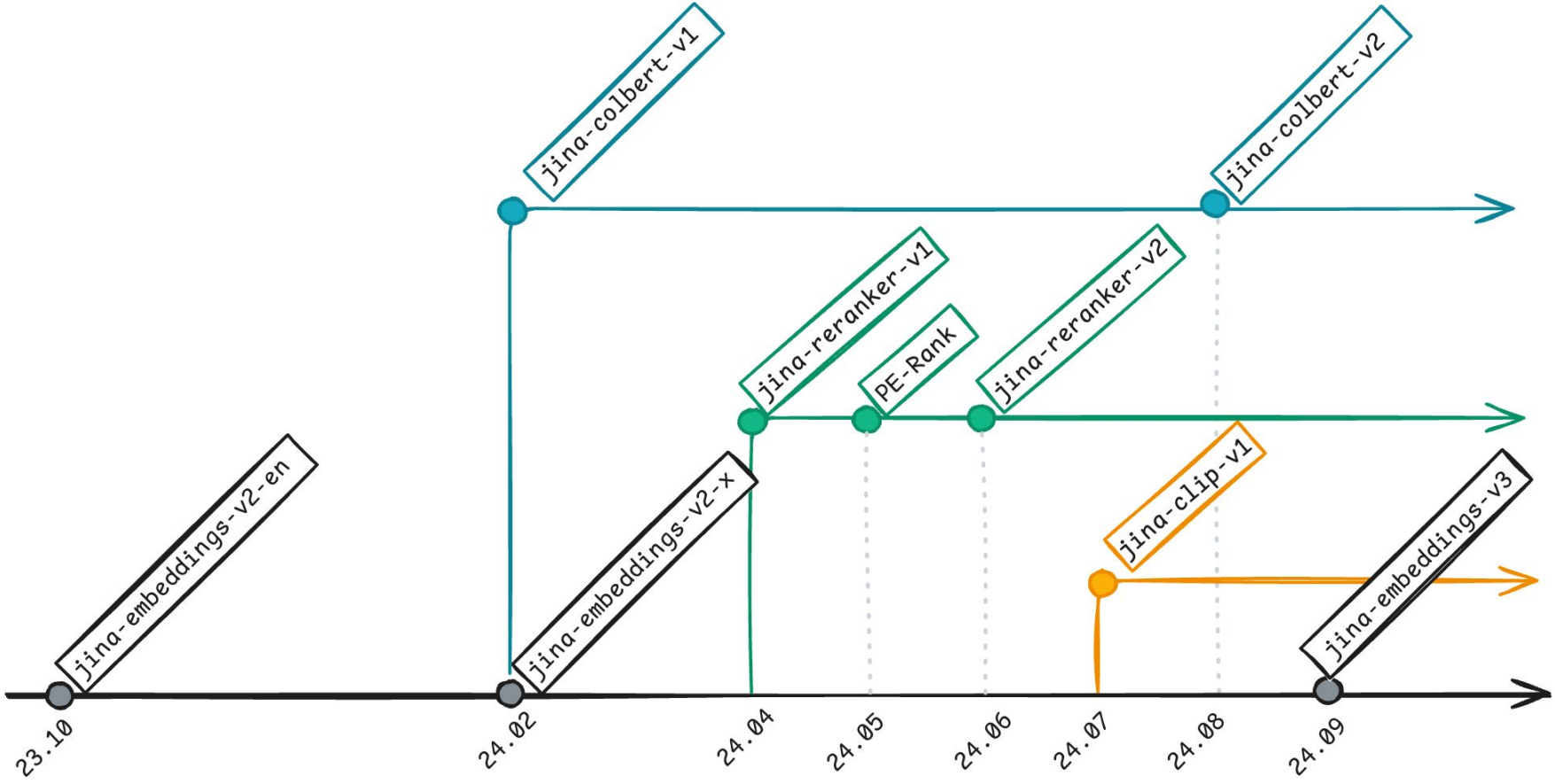

Indexing
Retrieval
Generation
LLM
Query
Answer
Index
Relevant Document
Documents

实际生产的RAG流水线

- Query改写
- 路由控制器
- Query结构化
- 索引优化
- 重排序
- ...

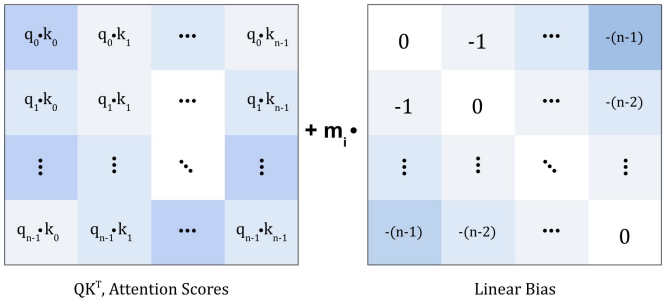


Jina AI的RAG工具


jina-colbert-v1
jina-colbert-v2
jina-reranker-v1
PE-Rank
jina-reranker-v2
jina-clip-v1
jina-embeddings-v3
jina-embeddings-v2-en
jina-embeddings-v2-x
23.10
24.02
24.04
24.05
24.06
24.07
24.08
24.09

jina-embeddings-v2

- 2023年10月发布，全球第一个支持8k输入长度的开源向量模型
- 融合ALiBi，使用 750Gb 语料，预训练 jina-bert-v2

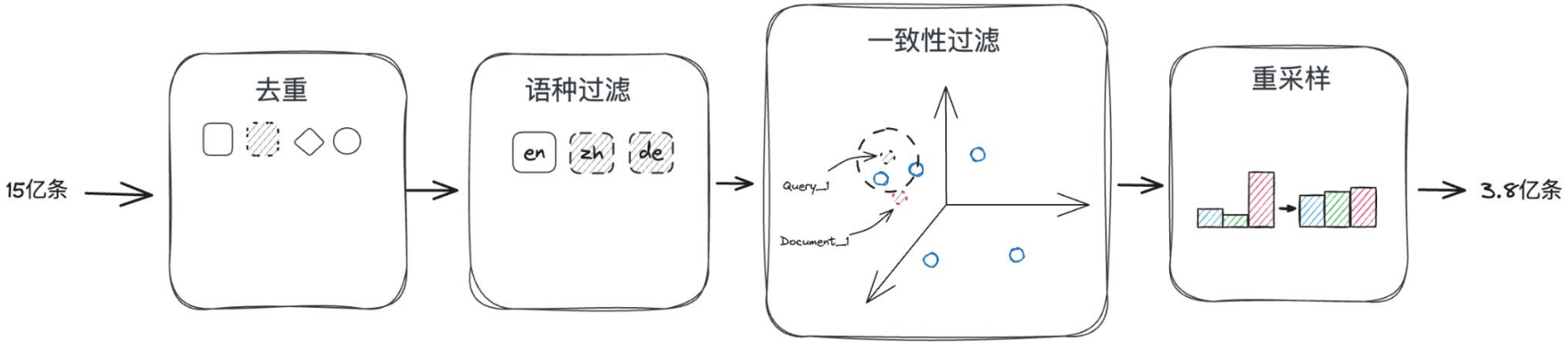


JINA EMBEDDINGS 2: 8192-Token General-Purpose Text Embeddings for Long Documents

Michael Günther, Jackmin Ong, Isabelle Mohr, Alaeddine Abdessalem, Tanguy Abel, Mohammad Kalim Akram, Susana Guzman, Georgios Mastrapas, Saba Sturua, Bo Wang, Maximilian Werk, Nan Wang and **Han Xiao**
Jina AI GmbH, Ohlauer Str. 43, 10999 Berlin, Germany
{michael.guenther, jackmin.ong, isabelle.mohr alaeddine.abdessalem,

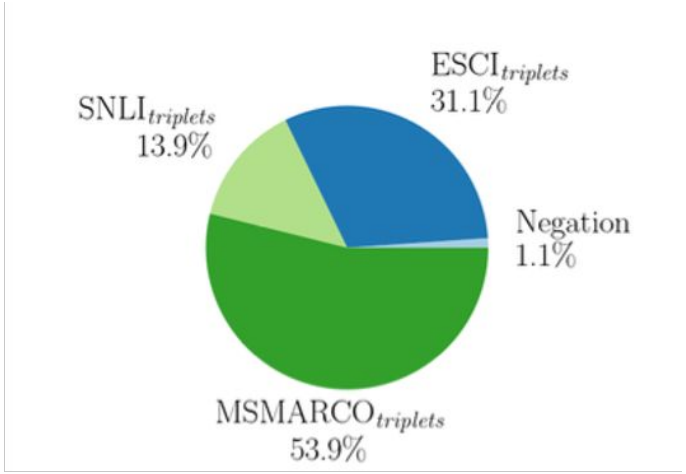
jina-embeddings-v2 弱监督学习

- 收集40+个开源数据源
  - 共15亿条文本对
- 3阶段数据清理
  - 得到3.8亿条高质量文本对，1700亿个token



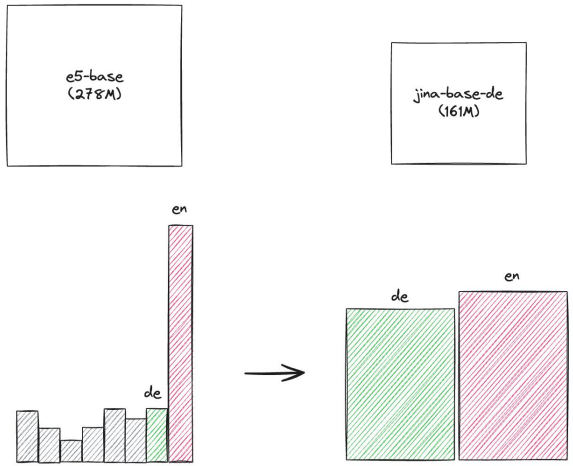
jina-embeddings-v2 有监督学习

- 收集MSMarco, Natural Questions，NLI，fever, ESCI(EN)数据集
- 构造高质量正负样本三元组共300万条
  - 针对检索任务，使用Hard negative mining
  - (anchor, positive, negative_1, ..., negative_15)
- 尽可能增大batch size
  - 使用混合精度
  - 使用activation checkpoint
  - 基于DeepSpeed
  - 使用MiniBatch
  - 使用gradient caching



jina-embeddings-v2双语模型

- 避免英文语料的偏差
- 避免多语言模型过大的词表
- 针对不同任务使用不同的损失函数
- 目标语种效果优于多语言模型



Multi-Task Contrastive Learning for 8192-Token Bilingual Text Embeddings

Isabelle Mohr, Markus Krimmel, Saba Sturua, Mohammad Kalim Akram, Andreas Koukounas, Michael Günther, Georgios Mastrapas, Vinit Ravishankar, Joan Fontanals Martínez, Feng Wang, Qi Liu, Ziniu Yu, Jie Fu, Saahil Ognawala, Susana Guzman, Bo Wang, Maximilian Werk, Nan Wang and **Han Xiao**
Jina AI GmbH, Ohlauer Str. 43, 10999 Berlin, Germany
research@jina.ai

Abstract

We introduce a novel suite of state-of-the-art bilingual text embedding models that are designed to support English and another target language. These models are capable of processing lengthy text inputs with up to 8192 tokens, making them highly versatile for a range of natural language processing tasks such as text retrieval, clustering, and semantic textual similarity (STS) calculations.

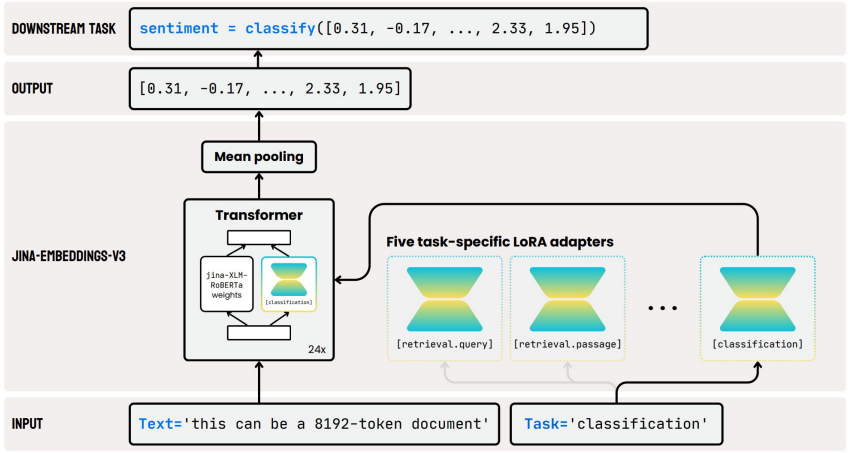
support tens of languages. This is achieved by fine-tuning pre-existing multilingual backbones like XLM-RoBERTa [Conneau et al., 2020] on mainly English data.[1] Alternatively, multilingual knowledge distillation can be applied to align embedding models across various languages using parallel data [Reimers and Gurevych, 2020] to cope with the scarcity of high-quality semantic pairs or triplets in the target languages. Despite these efforts, training data with such an extremely un-

Lang.	Model	CF	CL	PC	RR	RT	STS**	SM
en	jina-de-base	0.688	0.403	0.835	0.549	0.441	0.820	**0.318***
	jina-es-base	0.690	**0.405**	**0.846**	**0.553**	0.464	**0.835**	0.299*
	multilingual-e5-base	**0.730**	0.400	0.836	0.548	**0.489**	0.803	0.301*
de	jina-de-base	0.665	0.299	**0.583***	0.639	**0.387**	**0.714**	–
	multilingual-e5-base	**0.687**	**0.328**	0.541*	**0.648**	0.342	0.674	–
es	jina-es-base	0.671	**0.440**	**0.583***	**0.739**	**0.511**	**0.788**	–
	multilingual-e5-base	**0.685**	0.413	0.541*	0.736	0.478	0.783	–

CF: Classification Accuracy CL: Clustering $\mathcal{V}$ measure PC: Pair Classification Average Precision
RR: Reranking MAP RT: Retrieval nDCG@10 STS: Sentence Similarity Spearman Correlation
SM: Summarization Spearman Correlation

jina-embeddings-v3

- 1b以下参数规模模型SOTA
- 使用LoRA适配不同任务
- 使用MRL自定义向量维度
- 支持8k输入长度
- 支持89种语言



jina-embeddings-v3: Multilingual Embeddings With Task LoRA

Saba Sturua*, Isabelle Mohr*, Mohammad Kalim Akram*
Michael Günther*, Bo Wang*, Markus Krimmel, Feng Wang
Georgios Mastrapas, Andreas Koukounas, Nan Wang and **Han Xiao**
Jina AI GmbH, Prinzessinnenstraße 19–20, 10969 Berlin, Germany
research@jina.ai

jina-colbert-v1

- 第一款支持8k长度的ColBERT模型

长文本上效果优于ColBERTv2

Model	Used context length	Model max context length	Avg. NDCG@10
ColBERTv2	512	512	74.3
Jina-ColBERT-v1 (truncated)	512*	8192	75.5
Jina-ColBERT-v1	8192	8192	83.7
Jina-embeddings-v2-base-en	8192	8192	**85.4**

短文本MSMARCO，与ColBERTv2齐平

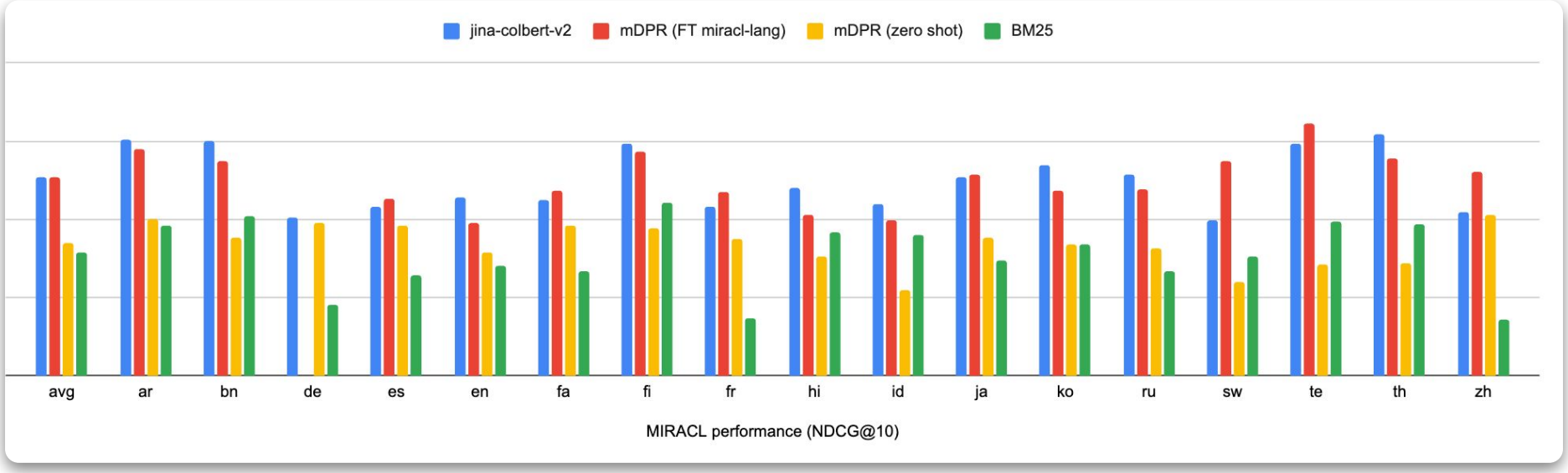
dataset	ColBERTv2	Jina-ColBERT-v1
ArguAna	46.5	49.4
ClimateFEVER	18.1	19.6
DBPedia	45.2	41.3
FEVER	78.8	79.5
FiQA	35.4	36.8
HotPotQA	67.5	65.6
NFCorpus	33.7	33.8
NQ	56.1	54.9
Quora	85.5	82.3
SCIDOCS	15.4	16.9
SciFact	68.9	70.1
TREC-COVID	72.6	75.0
Webis-touché2020	26.0	27.0
Average	50.0	50.2

jina-colbert-v2

- 使用MRL支持最低64维向量
- 支持89种语言

Jina-ColBERT-v2: A General-Purpose Multilingual Late Interaction Retriever

Rohan Jha[1*] and **Bo Wang**[2] and **Michael Günther**[2] and **Georgios Mastrapas**[2] and **Saba Sturua**[2] and **Isabelle Mohr**[2] and **Andreas Koukounas**[2] and **Mohammad Kalim Akram**[2] and **Nan Wang**[2] and **Han Xiao**[2]
[1]The University of Texas at Austin, Austin, Texas, USA
[2]Jina AI GmbH, Prinzessinnenstr. 19-20, 10969 Berlin, Germany
research@jina.ai



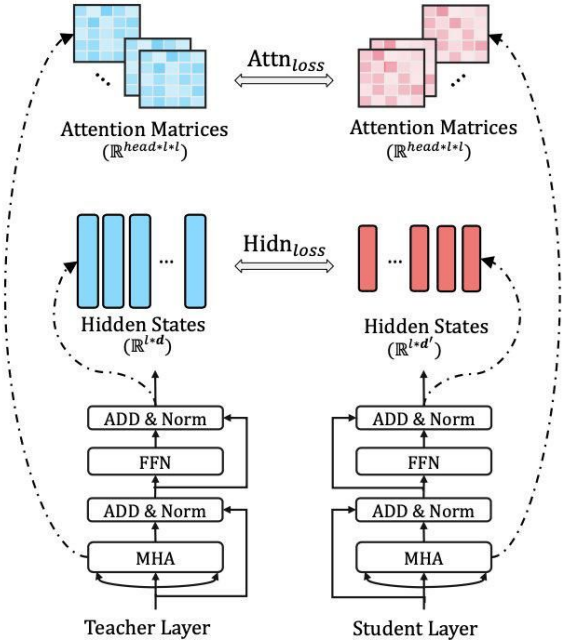
jina-reranker-v1

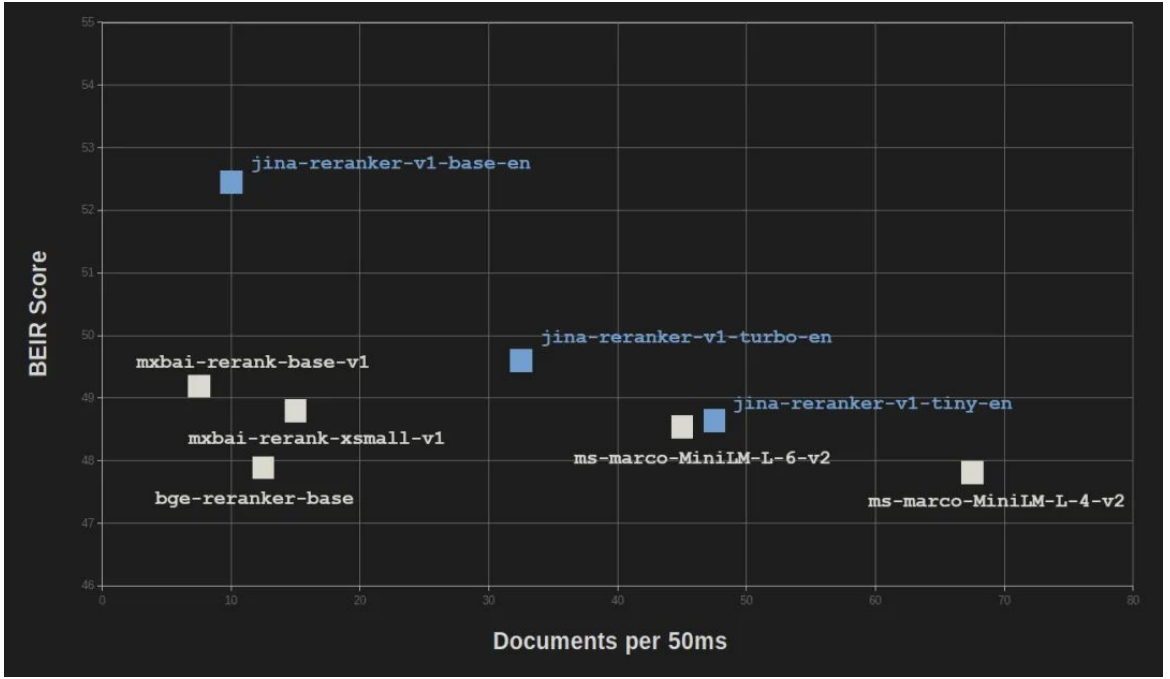
- 预训练：基于Jina BERT v2，支持 8K 上下文输入；
- 分阶段训练：逐步提升模型排序能力；
- 迁移学习：将Embedding模型学习到的知识迁移到Reranker模型；
- 训练数据：使用和Embedding模型相同来源的训练数据；

	No Reranker		jina-reranker		bge-reranker-base		bce-reranker-base_v1		cohere-reranker	
Embedding model	Hit Rate	MRR	Hit Rate	MRR	Hit Rate	MRR	Hit Rate	MRR	Hit Rate	MRR
jina-embeddings-v2-base-en	0.8053	0.5156	0.8737	**0.7229**	0.8368	0.6568	0.8737	0.7007	**0.8842**	0.7008
bge-base-en-v1.5	0.7842	0.5183	**0.8368**	**0.6895**	0.8158	0.6586	0.8316	0.6843	0.8368	0.6739
bce-embedding-base_v1	0.8526	0.5988	0.8895	0.7346	0.8684	0.6927	0.9157	**0.7379**	**0.9158**	0.7296
CohereV3-en	0.7211	0.4900	0.8211	**0.6894**	0.8000	0.6285	0.8263	0.6855	**0.8316**	0.6710
Average	0.7908	0.5307	0.8553	**0.7091**	0.8303	0.6592	0.8618	0.7021	**0.8671**	0.6938

jina-reranker-v1-turbo/tiny

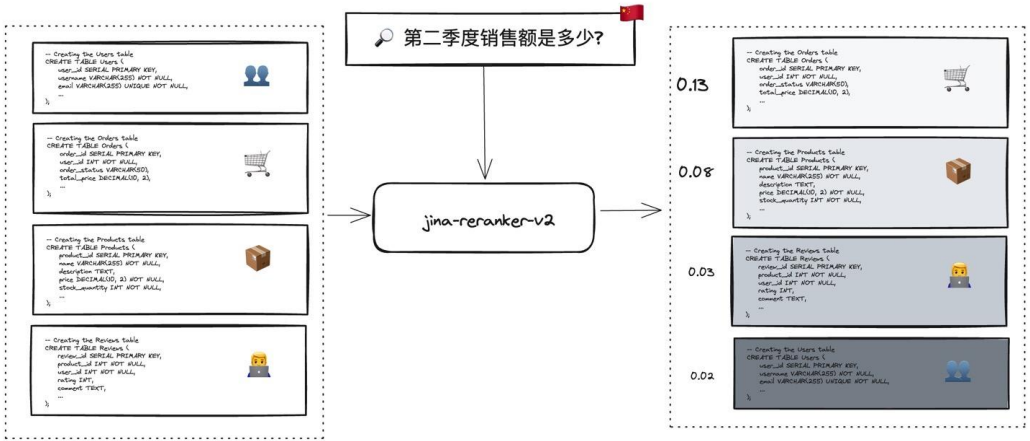
- 使用模型蒸馏技术，平衡准确率和推理速度

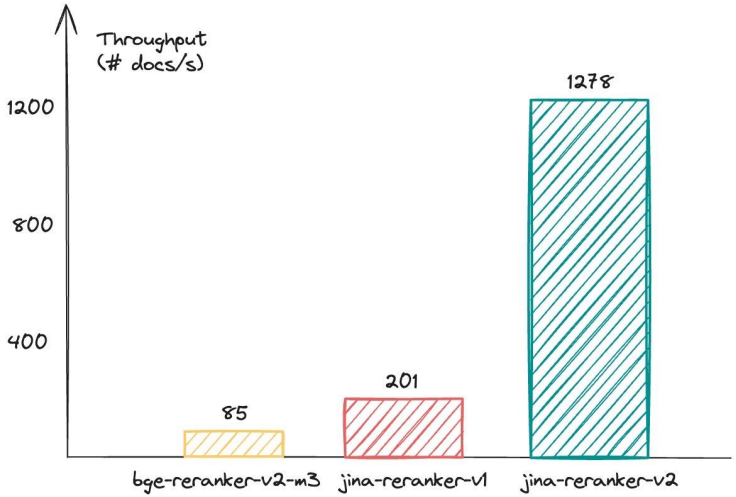




jina-reranker-v2

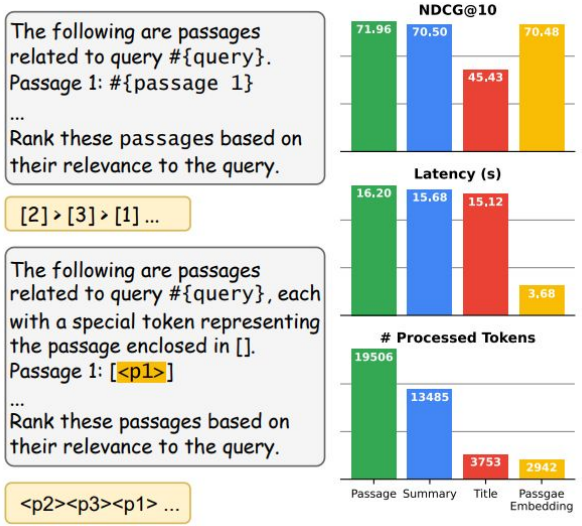
- 支持89种语言
- 针对结构化数据和代码数据专门优化
- 使用模型蒸馏和flash-attention优化推理速度





PE-Rank

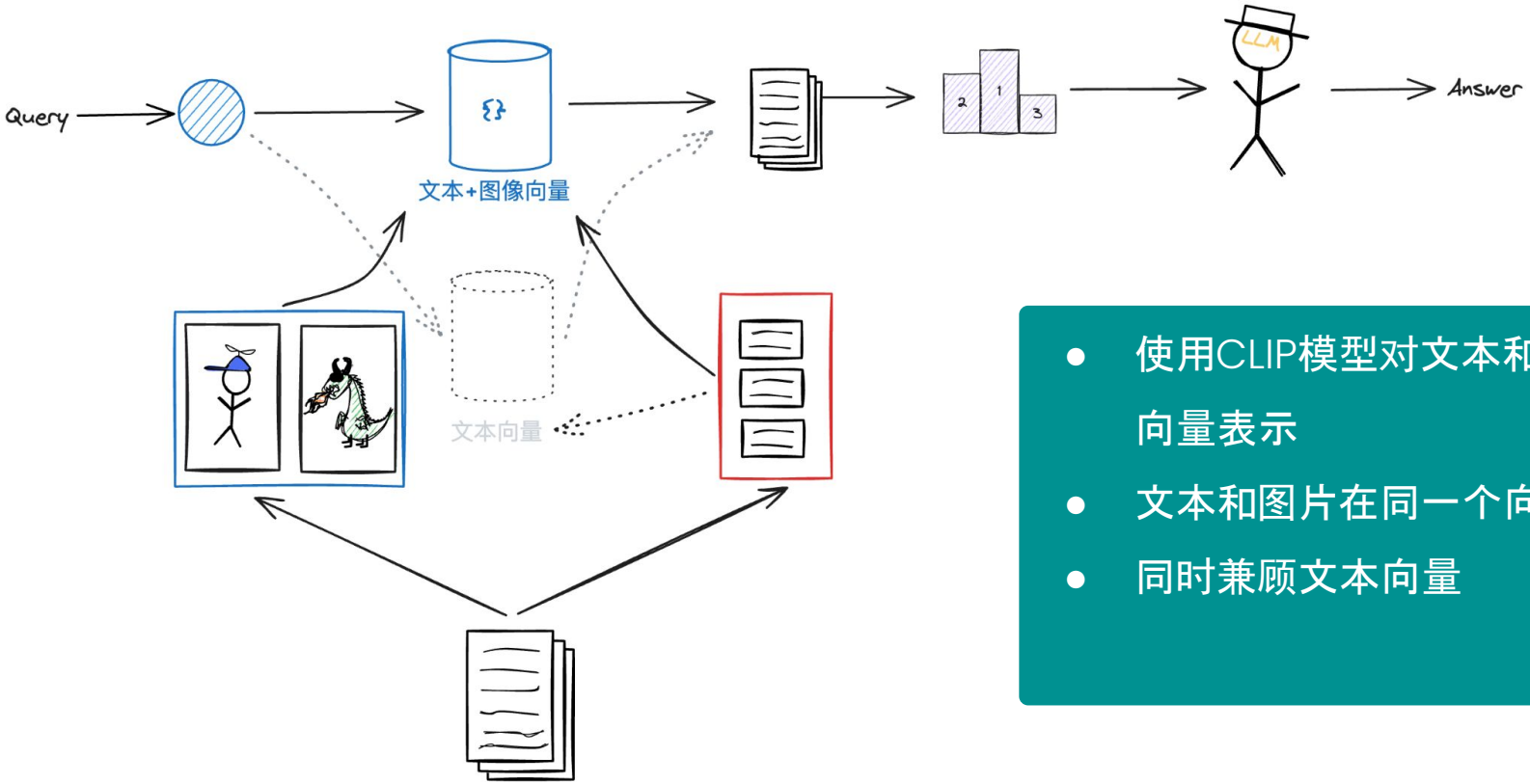
- 使用大模型做排序
- 使用embedding作为模型输入
- 显著提升大模型排序效率



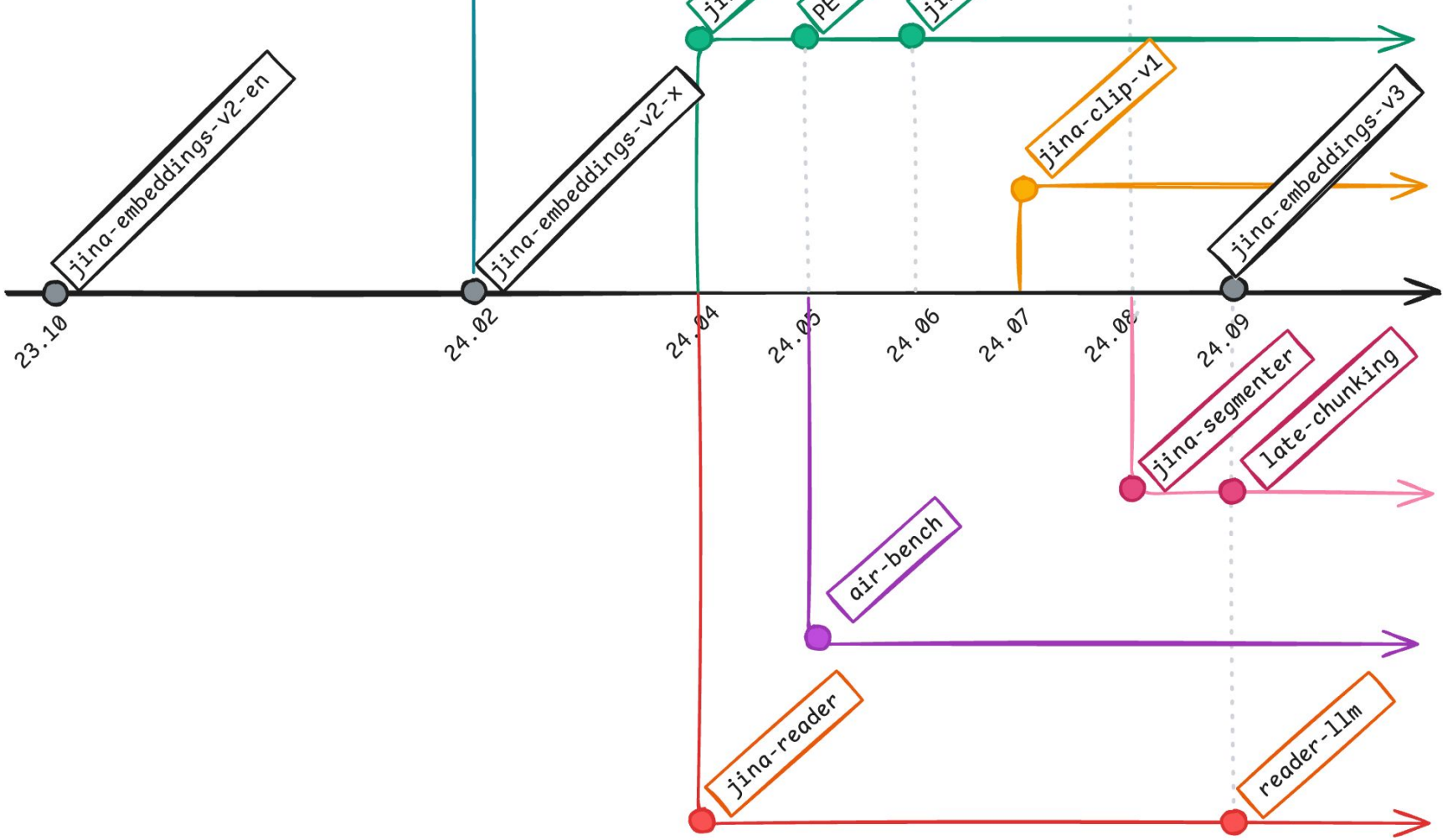
Leveraging Passage Embeddings for Efficient Listwise Reranking with Large Language Models

Qi Liu[1,2], **Bo Wang**[2], **Nan Wang**[2] and **Jiaxin Mao**[1]
[1]Gaoling School of Artificial Intelligence, Renmin University of China, [2]Jina AI
{qiliu6777, maojiaxin}@gmail.com, {bo.wang, nan.wang}@jina.ai

jina-CLIP-v1



- 使用CLIP模型对文本和图片进行向量表示
- 文本和图片在同一个向量空间
- 同时兼顾文本向量


jina-embeddings-v2-en
jina-embeddings-v2-x
jina-clip-v1
jina-embeddings-v3
jina-segmenter
late-chunking
air-bench
jina-reader
reader-llm
23.10
24.02
24.04
24.05
24.06
24.07
24.08
24.09

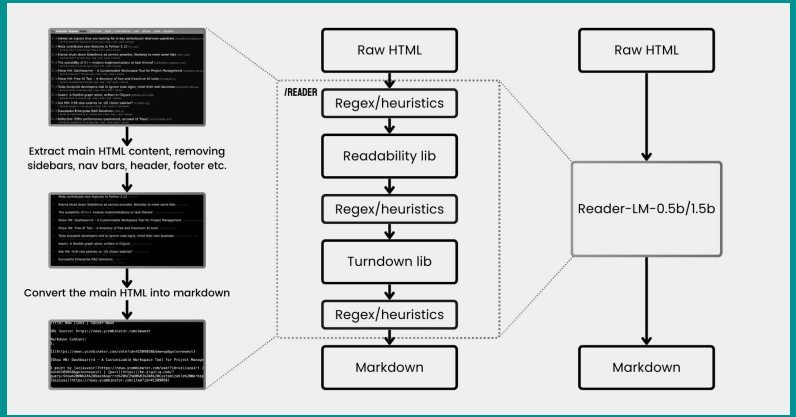
文本解析

jina-reader

- 将html转换为Markdown

reader-llm

- Small Language Model应用市场广泛



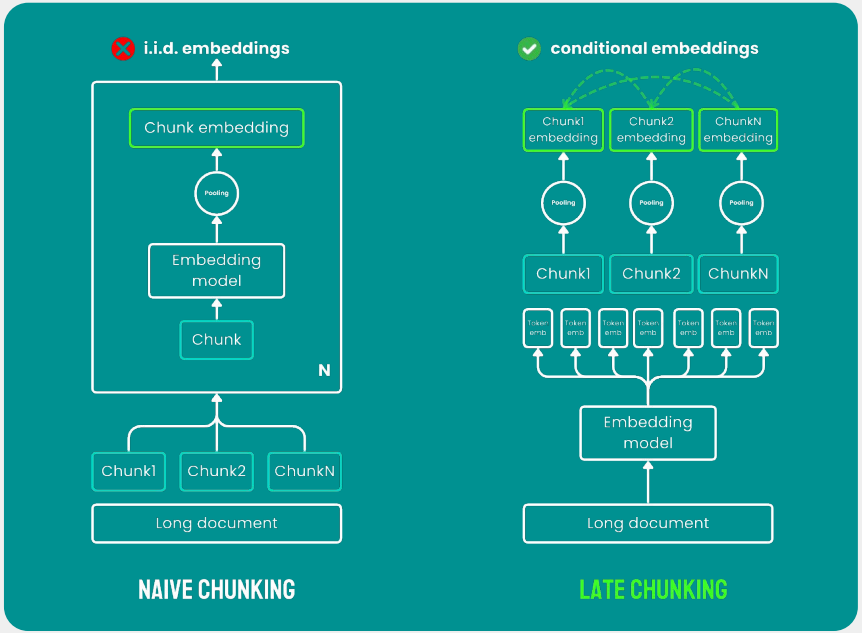
文本块切分（Chunking）

jina-segmenter

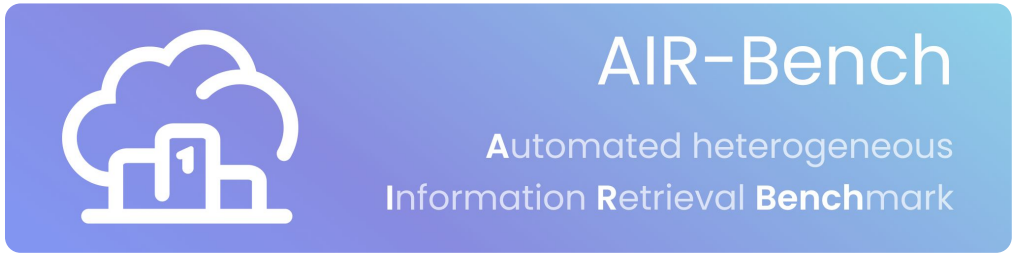
Semantic Chunking Regex v1

```
(?x)  # Used in https://tokenize.jina.ai; Enable verbose mode for readability
(
    # 1. Headings (Setext-style, Markdown, and HTML-style)
    (?:^(?:[#*=-]{1,6}|\w[^\r\n]{0,199}\r?\n[-=]+|<h[1-6][^>]{0,99}>)[^\r\n]{1,200}(?:</h[1-6]>)?(?:\r?\n|$))
    |
    # 2. List items (bulleted, numbered, lettered, or task lists, including nested, up to three levels)
    (?:(?:^|\r?\n)[ \t]{0,3}(?:[-*•·]|\d{1,3}\.|\w\.|\[[ xX]\])[ \t]+[^\r\n]{1,200}
    (?:(?:\r?\n[ \t]{2,5}(?:[-*•·]|\d{1,3}\.|\w\.|\[[ xX]\])[ \t]+[^\r\n]{1,200}){0,5}
    (?:\r?\n[ \t]{4,7}(?:[-*•·]|\d{1,3}\.|\w\.|\[[ xX]\])[ \t]+[^\r\n]{1,200}){0,5})?)
    |
    # 3. Block quotes (including nested quotes and citations, up to three levels)
    (?:(?:^>(?:>|\s{2,}){0,2}[^\r\n]{0,200}\r?\n?){1,10})
    |
    # 4. Code blocks (fenced, indented, or HTML pre/code tags)
    (?:(?:^|\r?\n)(?:```|~~~)(?:\w{0,20})?\r?\n[\s\S]{0,1000}?(?:```|~~~)\r?\n?
    |(?:(?:^|\r?\n)(?: {4}|\t)[^\r\n]{0,200}(?:\r?\n(?: {4}|\t)[^\r\n]{0,200}){0,20}\r?\n?)
    |(?:<pre>(?:<code>)?[\s\S]{0,1000}?(?:</code>)?</pre>))
    |
    # 5. Tables (Markdown, grid tables, and HTML tables)
    (?:(?:^|\r?\n)(?:\|[^\r\n]{0,200}\|(?:\r?\n\|[-:]{1,200}\|){0,1}(?:\r?\n\|[^\r\n]{0,200}\|){0,20}
    |<table>[\s\S]{0,2000}?</table>))
    |
    # 6. Horizontal rules
    (?:^(?:[-*_]){3,}\s*$|<hr\s*/?>)
    |
    # 7. Sentences or phrases ending with punctuation
    (?:[^\r\n]{1,300}(?:[.!?…]|\.{3}|[\u2026\u2047-\u2049]|[\p{Emoji_Presentation}\p{Extended_Pictographic}])
  (?:\s+|\r?\n|\Z))
    |
    # 8. Quoted text, parenthetical phrases, or bracketed content
    (?:"""[^\r\n]{0,300}"""
    |['"`'"][^\r\n]{0,300}['"`'"]
    |\([^\r\n()]{0,200}(?:\([^\r\n()]{0,200}\)[^\r\n()]{0,200}){0,5}\)
```

late chunking



AIR-Bench：自动化的多样性信息检索评测基准



https://huggingface.co/spaces/AIR-Bench/leaderboard

标注成本高

- 使用LLMs生成测试数据
- 通过向量模型、排序模型、LLMs组合进行质量把控
- 覆盖多领域和多语言

自动生成评估数据

面向多种任 务

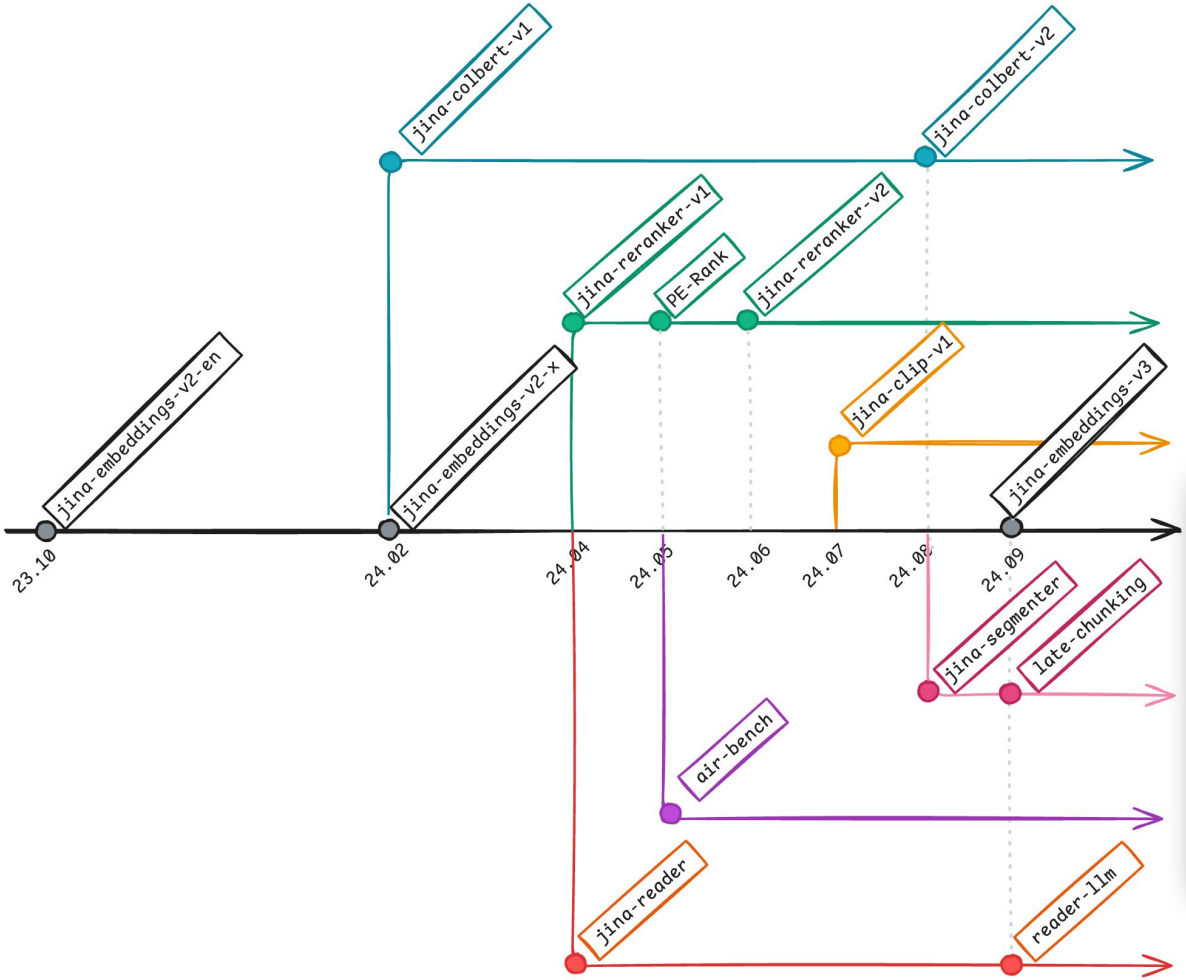
- 首次提出长文档内检索
- 专注问答任务

针对RAG场景设计

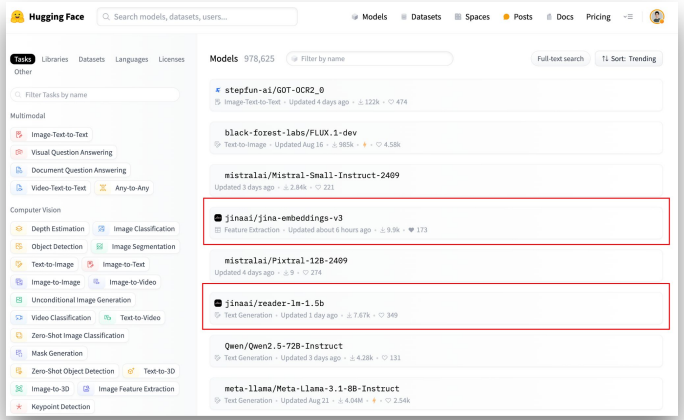
数据泄露

- 基于真实世界的语料库进行生产
- 定期进行更新，满足社区不断变化的评测需求

便于更新和扩展



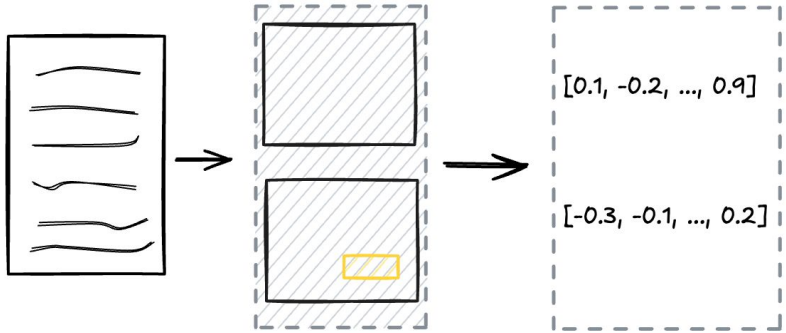
- 累计下载量2000万
- API每日消耗100亿token
- 多次登顶HuggingFace Trending排行榜



最优的文本块的大小是多少？

向量模型在RAG应用中的困境

- 语义溢出
  - 文本块过大，细颗粒度语义无法表示



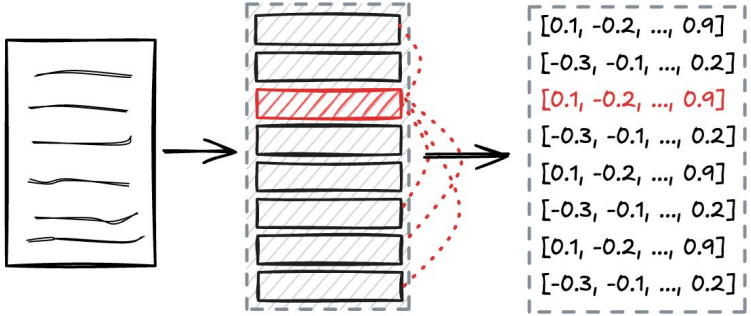
query：中国在奥运会上有哪些重要历史时刻？

2006年冬季奥林匹克运 动会.txt:焦点

=== 第六天 - 2月16日 ===
单板滑雪 - 美国选手塞思·韦斯科特（Seth Wescott）获得男子技巧 赛第一名，在首度成 为冬季奥运 竞赛之一的项目上称王。
冬季两项 - 法国女将弗洛 伦斯·巴维雷尔-罗贝尔（Baverel-Robert）于7.5公里小 项中获得金牌。银牌花落瑞典的安娜·卡伦·奥洛夫松（Anna CarinOlofsson），铜牌则由乌克兰的莉莉娅·叶夫列莫娃捧走。上届世界杯 优胜者德国籍奥运 卫冕冠军凯蒂威廉表 现失常，只 获得第七名。
北欧两项 - 在北欧两 项团体赛项目中，奥地利 选手在该项目夺冠。
速度滑冰 - 东道主意大利在男子速度滑冰比 赛中战胜加拿大 选手，以2.82 秒之差拿下金牌。
越野滑雪 - 爱沙尼亚选手安德 鲁斯·维尔帕鲁在越野滑雪女子 传统技术10公里的比 赛中获得金牌，挪威 选手包揽了二到四位。中国选手王春 丽获得18名，这是中国运 动员在此项目比赛上的最好成 绩。
花样滑冰 - 在花样滑冰的男 单的比赛中俄罗斯"冰王子"普鲁申科，以90.66分的个人 历史最高分 获得冠军。
...

向量模型在 RAG应用中的困境

- 背景截断
  - 文本块过小，上下文背景信息 丢失



query：巴黎奥运会的会徽设计有什么含义？

2024年夏季奥林匹克运 动会.txt:会徽

本届奥运及帕运首次共享同个会徽，会徽由金牌、火焰与法国人民和革命象征的　玛丽安娜三元素构成。

query：总结奥运会上的兴奋剂丑闻

2016年夏季奥林匹克運動會 .txt: 赛前资格问题

同时，该国的参 赛运动员必须在赛前进行「特别药检」，才可以参加奥运 。

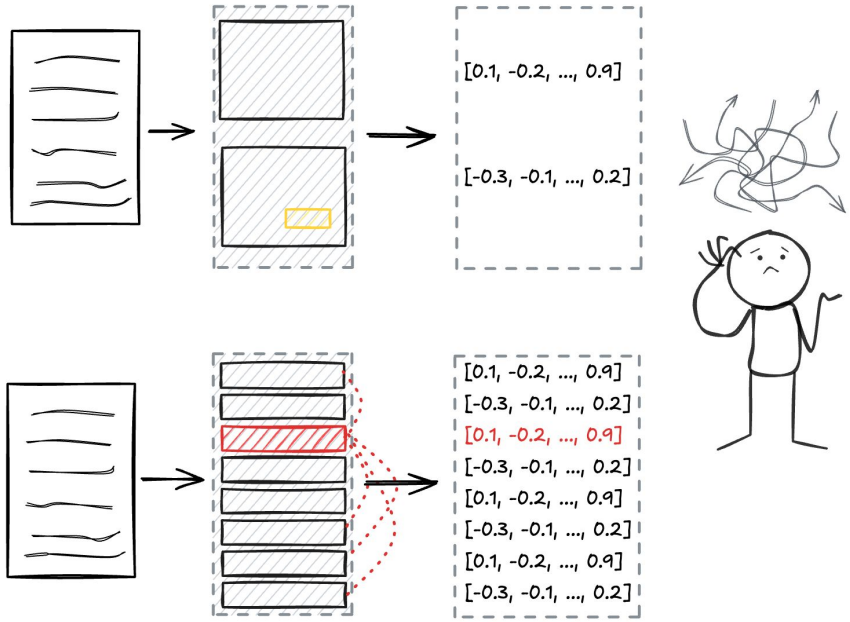
2006年冬季奥林匹克运 动会.txt:焦点

雪车 - 巴西选手桑托斯（Armando dos Santos）因被检验出服用禁 药诺龙（nandrolone），已 经将他送回巴西，他原本 2月24日参加四人雪 车赛。

向量化过程是信息的有 损压缩

- 文本是半结构化信息
  - 无法直接进行高效检索
- 向量是结构化信息
  - 一个固定长度的浮点数数组
- 向量模型是把半结构化信息转化为结构化信息的函数
- 计算向量的过程是有损压缩
  - 输入信息过多，需要压缩
  - 输入信息不足，缺少上下文背景信息

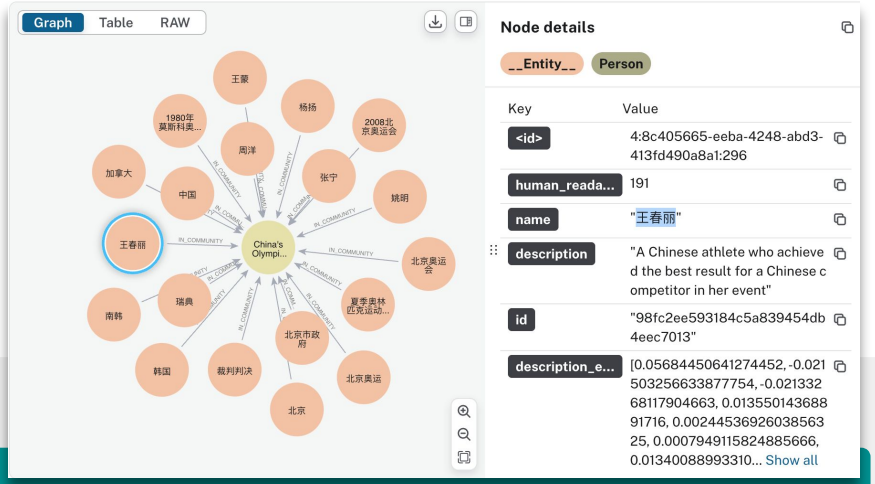
💡 增加向量的维度，可以有效提升向量模型性能



知识图谱是高效的知识结构化存储形式

知识图谱可以解决文本 块切分的困境

- 语义溢出
  - 用实体表示不同的语义颗粒度
  - 通过检索实体，准确匹配



query：中国在奥运会上有哪些重要历史时刻？

2006年冬季奥林匹克运 动会.txt:焦点

=== 第六天 - 2月16日 ===
单板滑雪 - 美国选手塞思·韦斯科特（Seth Wescott）获得男子技巧 赛第一名，在首度成 为冬季奥运 竞赛之一的项目上称王。
冬季两项 - 法国女将弗洛 伦斯·巴维雷尔-罗贝尔（Baverel-Robert）于7.5公里小 项中获得金牌。银牌花落瑞典的安娜·卡伦·奥洛夫松（Anna CarinOlofsson），铜牌则由乌克兰的莉莉娅·叶夫列莫娃捧走。上届世界杯 优胜者德国籍奥运 卫冕冠军凯蒂威廉表 现失常，只 获得第七名。
北欧两 项 - 在北欧两 项团体赛项目中，奥地利 选手在该项目夺冠。
速度滑冰 - 东道主意大利在男子速度滑冰比 赛中战胜加拿大 选手，以2.82 秒之差拿下金牌。
越野滑雪 - 爱沙尼亚选手安德 鲁斯·维尔帕鲁在越野滑雪女子 传统技术10公里的比 赛中获得金牌，挪威 选手包揽了二到四位。中国选手王春 丽获得18名，这是中国运动员在此项目比赛上的最好成 绩。
花样滑冰 - 在花样滑冰的男 单的比赛中俄罗斯"冰王子"普鲁申科，以90.66分的个人 历史最高分 获得冠军。
...

知识图谱可以解决文本块切分的困境

- 背景截断
  - 通过实体检索子图，提供充分的背景信息

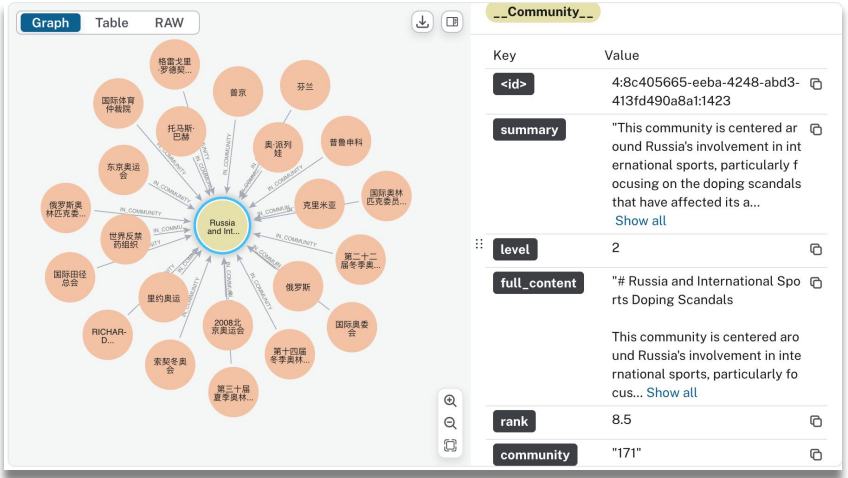
query：总结奥运会上的兴奋剂丑闻

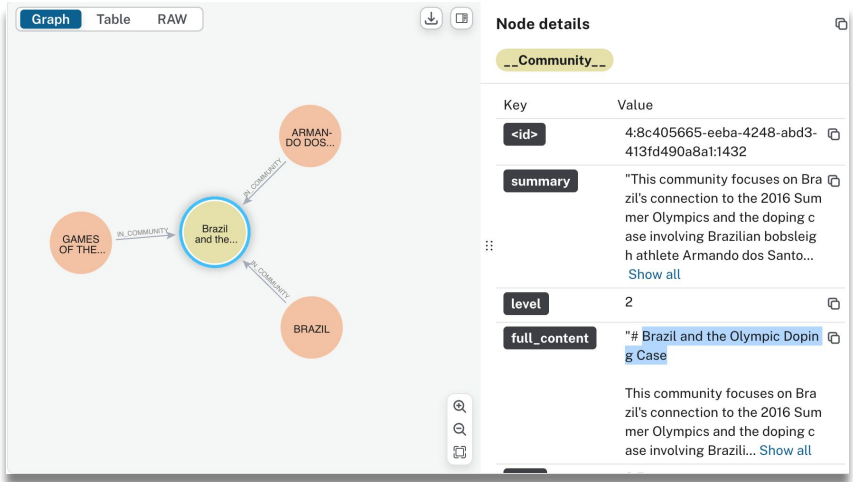
2016年夏季奥林匹克運動會 .txt: 赛前资格问题

同时，该国的参赛运动员必须在赛前进行「特别药检」，才可以参加奥运。

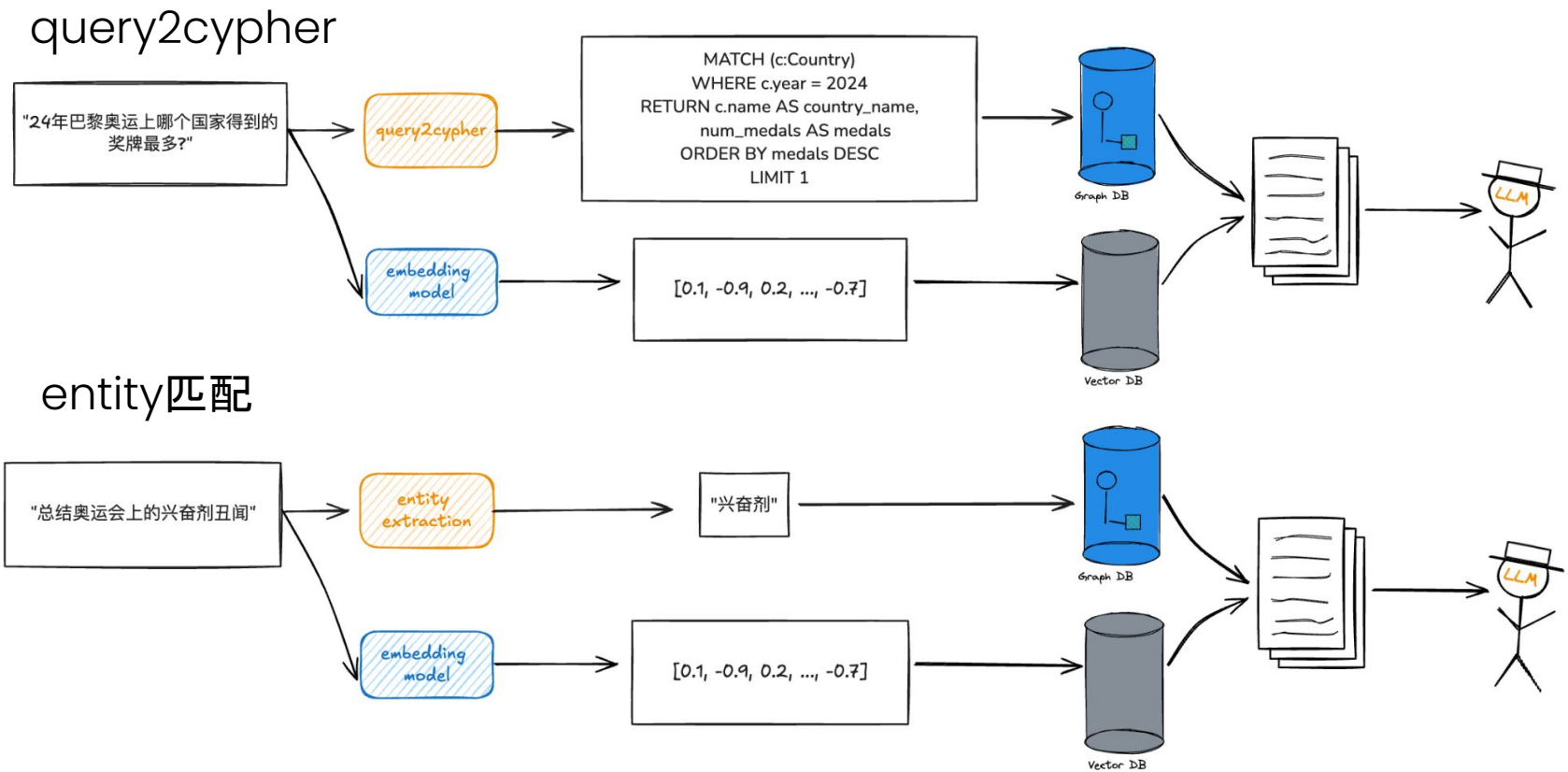
2006年冬季奥林匹克运 动会.txt:焦点

雪车 - 巴西选手桑托斯（Armando dos Santos）因被检验出服用禁 药诺龙（nandrolone），已经将他送回巴西，他原本2月24日参加四人雪 车赛。



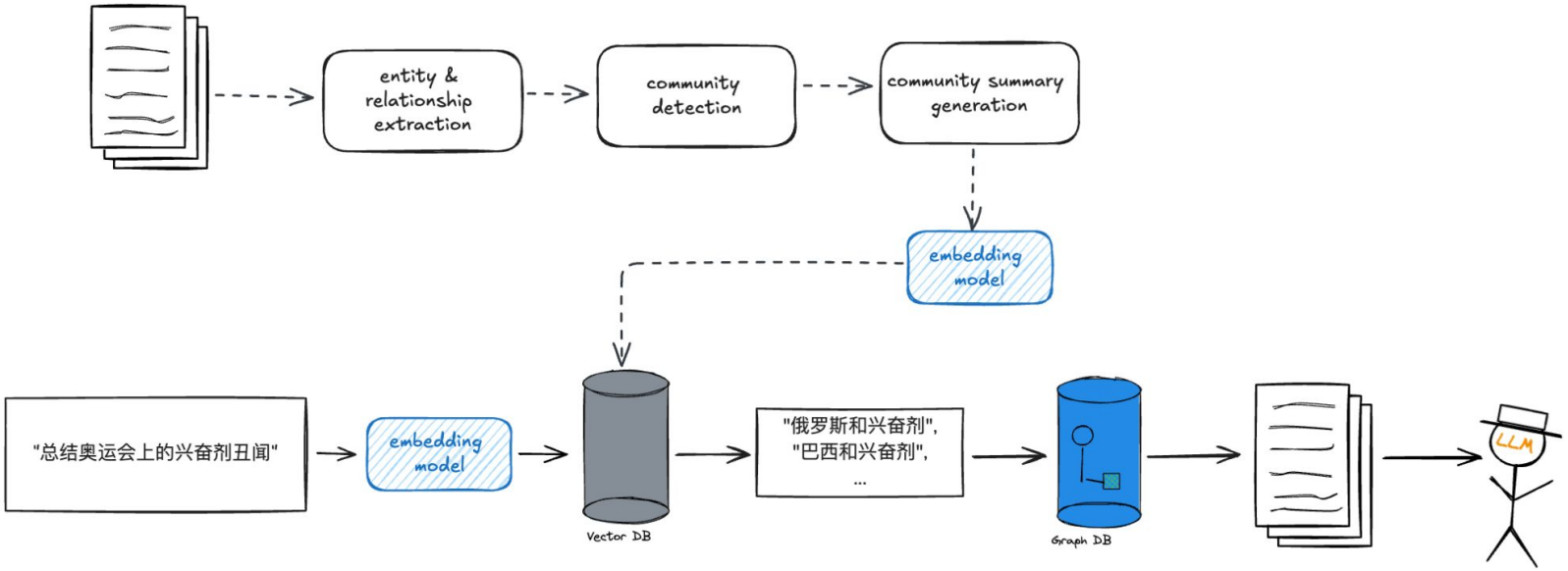


知识图谱在RAG中应用方式-1



知识图谱在RAG中应用方式-2

使用LLM构建知识图谱，通过构建的知识图谱增强结构化信息: GraphRAG



知识图谱在RAG应用中的挑战

查询效率低+成本高

20倍的token消耗用于索引

每次查询消耗60k-120k tokens

每次查询平均耗时15-40s

构建知识图谱成本高

准确率不高，短文本平均准确率90%，长文本准确率60-70%

依赖于人工定义schema，定义schema需要专业知识

无法高效的更新

需要保证知识图谱的时效性

每次更新需要重新构建graph和community

.jina

不要高估6个月后的变化，
不要低估18个月后的变化。
在AI浪潮中，找到自己不 变的价值。

Q & A
Thanks!

X: https://x.com/nanwang_t

GitHub: https://github.com/nan-wang