智能新篇章：有道“子曰”大模型的创新与开源探索

孙艳庆　网易有道

科技生态圈峰会 + 深度研习

——1000 + 技术团队的共同选择

KEYLINK ing

K+峰会

K+峰会 上海站
K+全球软件研发行业创新峰会
时间：2024.06.21-22

K+峰会 敦煌站
K+思考周®研习社
时间：2024.10.17-19

K+峰会 香港站
K+思考周®研习社
时间：2024.11.10-12

K+峰会详情

AiDD峰会

AiDD峰会 上海站
AI+研发数字峰会
时间：2024.05.17-18

AiDD峰会 北京站
AI+研发数字峰会
时间：2024.08.16-17

AiDD峰会 深圳站
AI+研发数字峰会
时间：2024.11.08-09

AiDD峰会详情

演讲嘉宾

孙艳庆

网易技术总监

网易有道AI语音交互团队负责人
2010年获中科院声学所信号与信息处理专业博士学位
毕业后参与并主导了三星S-Voice在线/离线语音方案、打造了国内首发的免触语音拨号/接听、拍照等产品解决方案。19年初加入网易有道、组建语音技术团队，在语音、大模型、AI老师等多方向上结合场景不断打磨、取得突破，支撑联通集团、网易云音乐、网易传媒、长安深蓝汽车、OPPO离线通话翻译、宝宝树、Hi Echo、有道词典笔/听力宝等明星产品！发表学术论文10余篇，授权专利10余件，带领团队在相关国际评测中获得多项佳绩。目前聚焦在教育学习场景打造下一代、更极致好用的AI语音/大模型解决方案！

目录
CONTENTS

PART 01
人工智能技术与有道的布局

团队持续投入Transformer

2023
「子曰」大模型

有道AI+教育布局

虚拟人技术

面向下一代AI+教育的需求
推出了多项虚拟人/数字人技术，包括在线2D真人数字人，
离线数字人技术以及其它相关技术

计算机视觉（OCR）

OCR、手指点读、口算批改
OCR目前支持104种语言，准确率高达99.6%。有道OCR能力支持公式、模糊文字、拼音与手写体等多场景识别，也支持特殊字体、复杂背景、特殊材质等的扫描识别。有道OCR目前已全面应用于有道系产品，智能硬件与第三方合作企业、学校等，大大加强了有道AI应用落地的能力。

大语言模型（子曰）

“场景为先”的教育垂类大模型，作为基座模型支持诸多下游任务，向所有下游场景提供语义理解、知识表达等基础能力，为不同学习场景设计了定制化的模型，以实现模型与场景的高度契合

智能语音（ASR/TTS/CAPT）

基于Transformer的ASR技术，行业领先的多语种识别准确率，支持识别几十种语言。高质量定制语音合成技术，媲美真人的自然度、富有感情。能提供单词和句级别纠音能力的语言学习技术，提升学习效率和兴趣。

自然语言（YNMT）

NMT、对话技术、作文批改，有道神经网络翻译（NMT），支持86种语言互译。在通用的测试标准BLEU上，有道NMT在新闻领域测试集效果远优于国际翻译引擎Google和Microsoft。在此基础上，有道研发了离线NMT，在无需连网的情况下翻译质量依然优质。

高性能计算（HPC）

基于GPU的高性能多机多卡训练集群
基于GPU的云端推理、
基于端侧CPU、NPU、TPU等框架的推理加速库和离线引擎技术
保障了有道AI能力在不同云平台、设备端高效、快速的落地效果

PART 02
“子曰”大模型的进展

通用大模型概览

- 大力出奇迹：
  - 大规模、高质量的数据集（可购买）
  - 算法和模型的优化（卷各参数、人才、经验）
  - 算力资源（GPU、存储、网络带宽，可购买）

- 重金砸出的大模型

LLM	Release	Parameters	Context	Pretraining Tokens	Supervised fine-tuning	Human Preferences	MMLU	MATH	GSM8K	HumanEval
GPT4	2023.3.14	1.7T	128K				86.40%	52.90%	92%	67%
Claude 3 Opus	2024.3.4	2T	200K	40T			86.80%	61.00%	95%	85%
Llama2	2023.7.18	70B	4K	2T	100K+	1M+	69.80%		54.10%	31%
Grok-1	2023.3.17	314B	8K				73%	24%	62.90%	63.20%
DBRX	2023.3.27	132B/36B	32K	12T			73.70%		66.90%	70.10%
Grok-1.5	2023.3.28		128K				81.30%	50.60%	90%	74.10%

用有限的资源研发领域LLM

- 从 拿着锤子找钉子
- 到 对着钉子找锤子
- 资源用在刀刃上
- 从场景出发，聚焦在一两个核心功能
- 不追求通用能力，但要在目标场景做到最好
- 设计一套完整的系统，而不全依赖大模型本
- 大模型只是核心能力，而不能解决所有问题

- 选择适合的模型尺寸
  - 7B、14B、70B？Qwen全家桶
  - 不同阶段看是动态的，长期看还是要往大参数发展
- 持续加强领域数据建设
  - 通用的数据，可以快速获取
  - 领域的高质量数据，需要长期、持续的投入
- 算力资源
  - 短期紧张，长期看，有各种办法能够缓解
  - 技术要不断突破

LLM研发的各类技术

- DPO（Direct Preference Optimization）
  - 一种基于人类偏好优化语言模型的方法
  - 与RLHF不同，DPO不依赖于明确的奖励建模或强化学习过程
  - 它直接优化模型输出，使其更符合人类的偏好
  - DPO通过比较好的和不好的响应，然后调整模型以增加好的响应的概率
  - 这种方法简化了训练过程，减少了计算成本，并且能够提高模型输出的质量，特别是在情绪控制方面表现出色

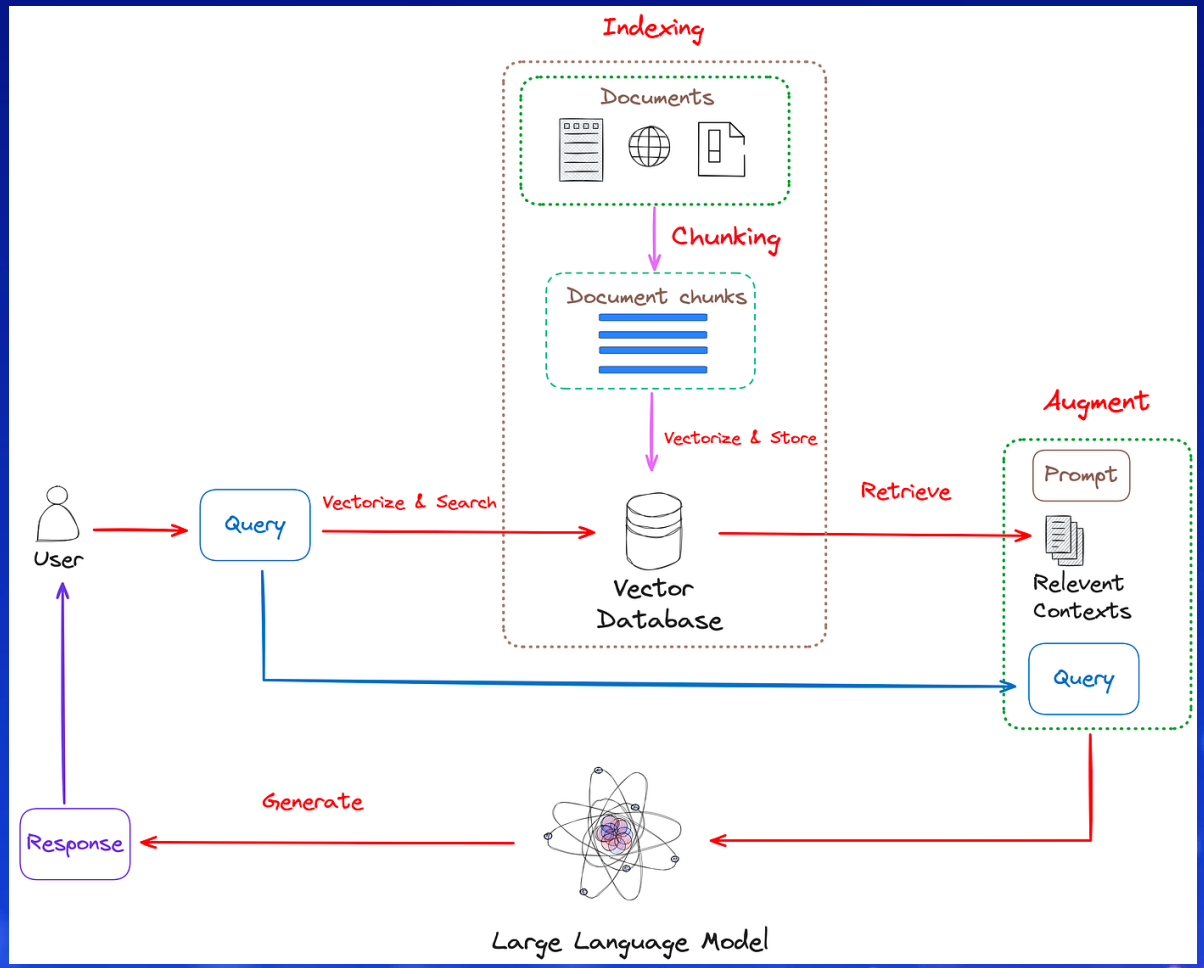
- Agent
  - 赋予LLM一种策略性思维结构，模拟人类处理问题的方法
  - Agent可以是“角色框架”，它让模型能够根据特定的角色或情境来生成响应
  - 这种方法使得LLM能够更好地理解和响应复杂的用户指令，提供更加个性化和情境化的交互体验

- RAG（Retrieval-Augmented Generation）
  - RAG结合了信息检索（IR）和生成模型的优势，通过从大型文档数据库中检索相关信息来增强模型的生成能力
  - RAG技术首先提出了Naive RAG，然后发展到Advanced RAG，再到Modular RAG
  - 这些进展使得RAG能够更有效地处理特定知识，提高生成内容的准确性和相关性
  - RAG通过迭代搜索和生成过程，使得模型能够生成更加准确和可靠的响应，特别是在需要最新信息或专业知识的场景中

- FT（Fine-tuning）
  - FT是LLM开发中的一个关键步骤，它通过在特定任务的数据集上进一步训练预训练模型来提高模型的性能
  - FT允许模型学习特定任务的特征和要求，从而在特定领域或任务中表现得更好
  - FT可以增强模型的知识，调整输出以符合特定的结构、风格或格式，并教授模型执行复杂指令

RAG流程的哲学

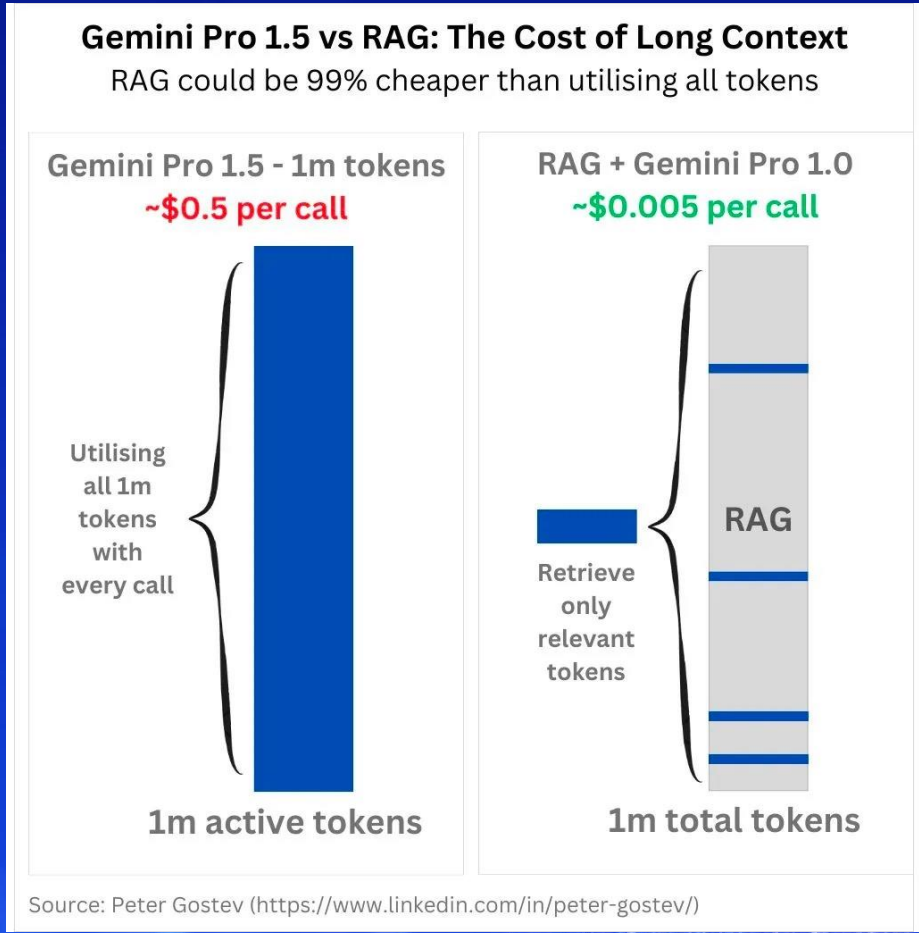
- Retrieval (Augmented) Generation
- 问错了问题
  - Chunk size多大?
- RAG = chunk + vectorDB + LLM?



RAG流程的哲学

- RAG vs finetune?
- RAG vs Long context LLM?

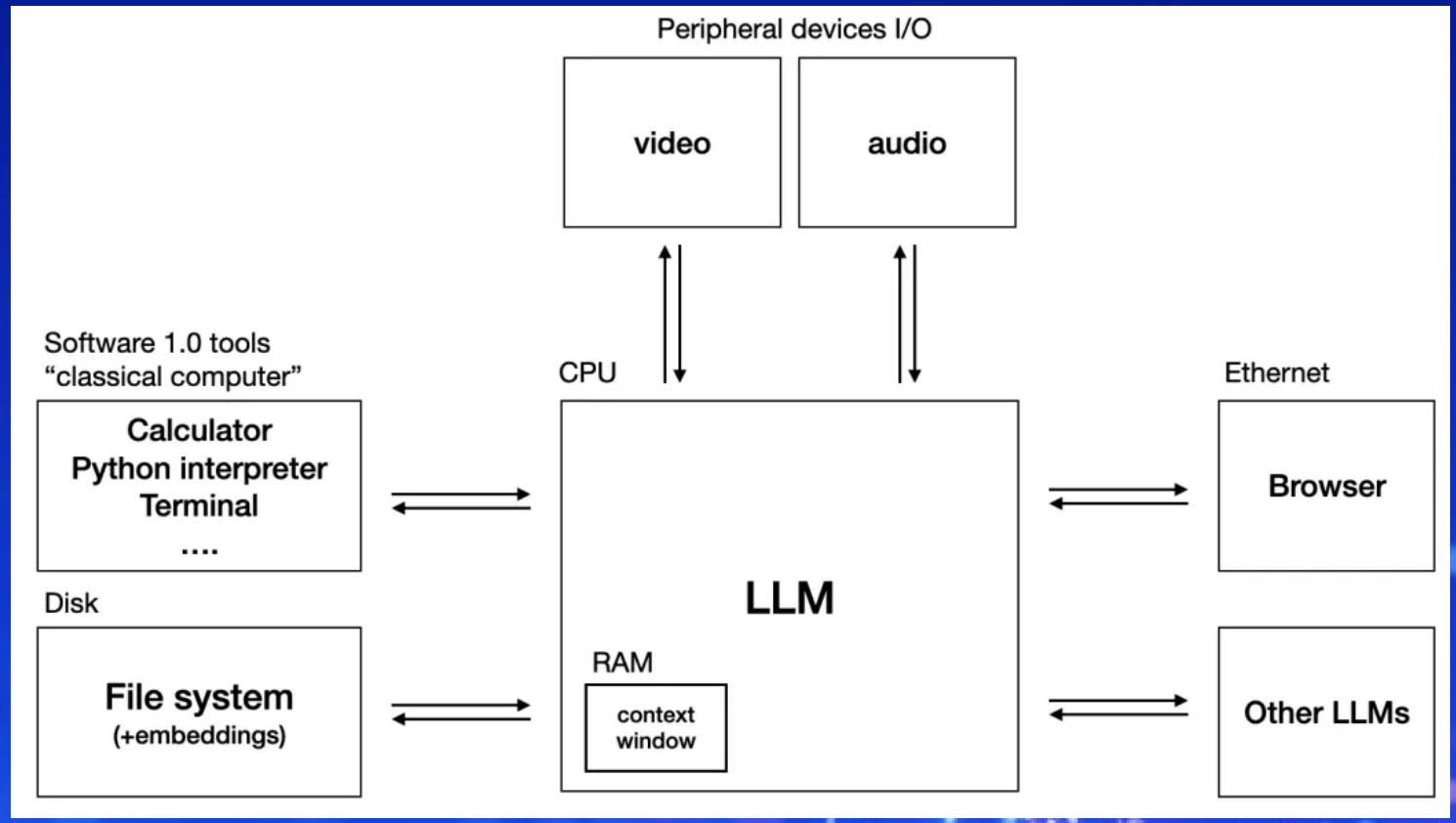
- RAG 满满的求生欲
  - Velocity （速度）
  - Value (价值/成本)
  - Volume (数据量)
  - Variaty（数据多样性）
  - 溯源



RAG流程的哲学

- RAG VS LLM context
- 硬盘 vs 内存

- 什么是RAG的关键?
- 形式多样的数据
  - 高质量的输入处理
  - 灵活的查询排序
- LLM的理解力与可靠性
  - 摘要
  - 翻译
  - 可控性



Embedding/Rerank

- 训练的好处

- 训练的关键点
- 问题的定义
  - 符合RAG需求
- 任务的安排
  - 难易要合适
- 数据的生成
  - 句子/GPT4
  - 真实问题

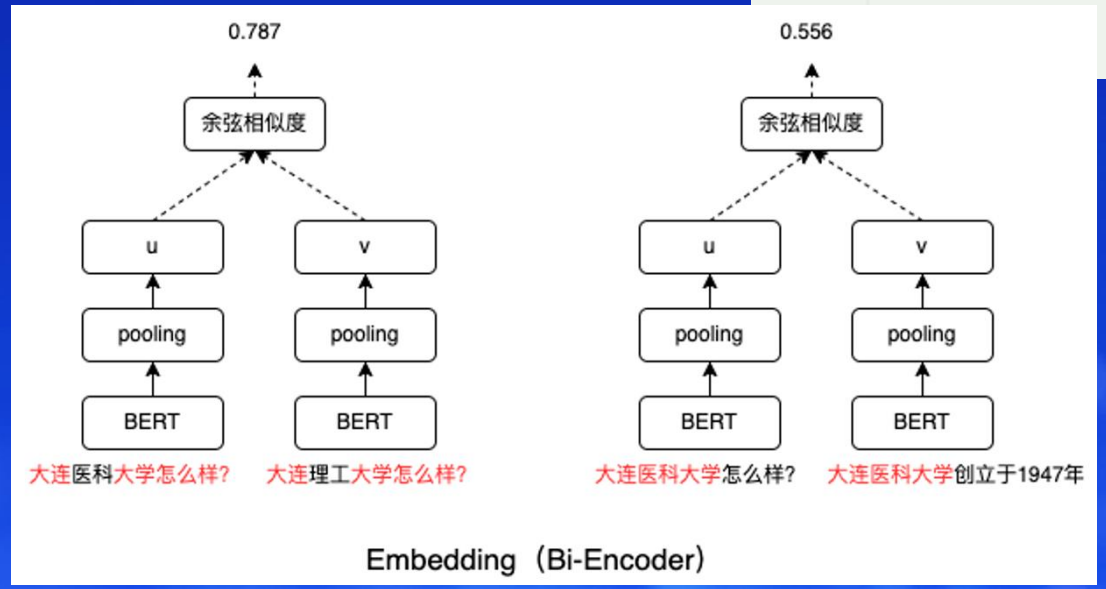
- Rerank score
- 可比较的分数

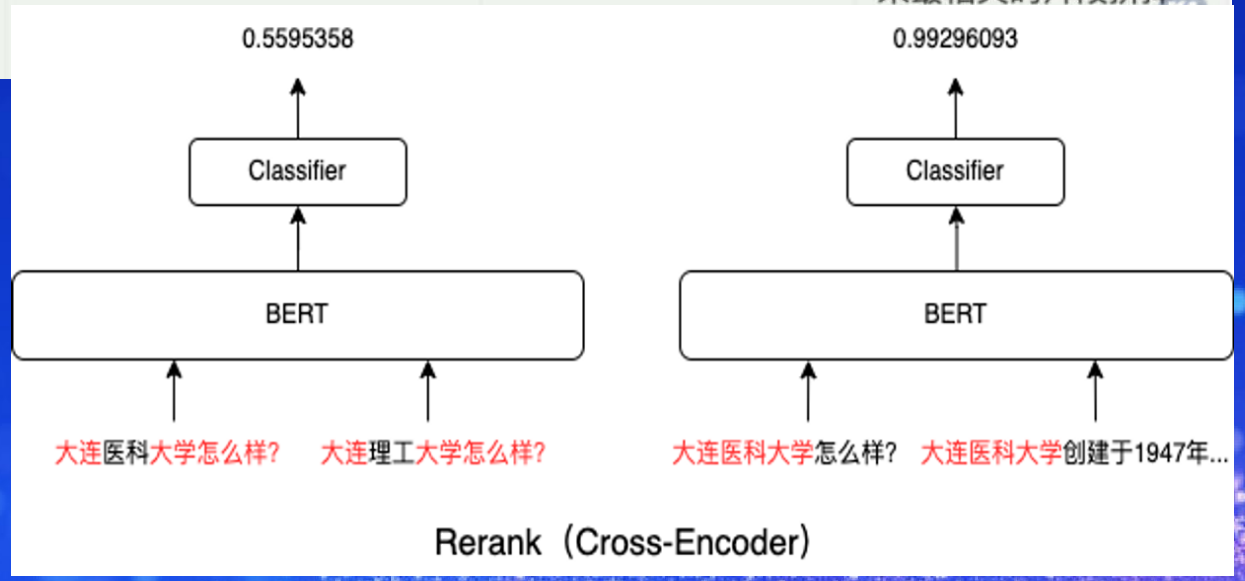
领域评测集	paper-v3	20230524	baike	books	law	paper	金融研报	速读日志 10w
模型版本	**top1/top3 acc(%) 耗时**	**top1/top3 acc(%) 耗时**	**top1/top3 acc(%) 耗时**	**top1/top3 acc(%) 耗时**	**top1/top3 acc(%) 耗时**	**top1/top3 acc(%) 耗时**	**top1/top3 acc(%) 耗时**	**top1/top3 acc(%) 耗时**
Embedding								
m3e_base	77.4/86.4	55.2/71.0	52.2/63.8	30.5/42.5	26.0/36.2	67.8/76.1	31.5/39.3	61.4/68.6
openai-text-embedding-ada2	en: 86.9/94.5	ch: 66.0/80.8	ch: 61.8/72.8	ch: 41.2/53.9	ch: 33.2/44.6	ch: 88.2/93.2	ch: 48.9/58.9	--
bge-base-zh-v1.5	79.9/88.8	54.7/70.5	51.4/63.5	29.3/40.5	22.6/31.9	65.7/74.1	31.7/40.3	62.5/68.6
llm-embedder	62.0/71.1	31.7/40.6	30.3/36.7	16.2/22.0	16.2/22.0	43.6/48.4	22.6/27.7	31.5/37.3
v7_0610_online	96.0/99.4	82.8/95.3	**82.7/93.1**	66.1/82.5	**66.3/82.4**	96.6/98.8	77.9/87.5	**74.1/81.7**
v9_新版	96.5/99.5	81.6/94.3	**77.7/88.5**	64.8/80.3	**62.0/78.0**	96.2/98.7	75.8/85.4	**84.6/90.4**
Rerank								
bge-reranker-base	93.0/97.1	82.2/90.4	71.2/78.9	69.9/79.4	50.2/62.4	86.0/90.4	72.1/79.8	87.0/90.7
v7_0610_online	96.0/99.4	83.2/96.0	84.4/93.7	65.8/83.0	65.6/81.1	96.9/99.0	79.2/88.0	83.2/87.5
rerank_新版	**97.8/99.3**	**94.8/98.6**	**91.4/96.0**	**89.8/94.6**	**75.7/85.6**	**97.8/99.2**	**87.8/91.4**	**90.4/98.6**

为什么要Rerank?

- 缺数据?
- 数据越多越好吗?
- 精度与速度的tradeoff

版本	数据	正确率	数据说明	备注
v1	+第一批数据	75/176=42.6%	第一批数据，已有数据：FAQ，书本，其他整理资料	先加一批数据感受一下
v2	+第一批数据 +第二批数据	106/176=60.2%	第二批数据，互联网数据：（1w+文件）	灌入大量数据后，效果提升非常明显
v3	+ 第一批数据 + 第二批数据 + 第三批数据	93/176=52.8%（-7.4%）	第三批数据，有道领世志愿院校信息+985院校信息总结	这个版本加了更多数据，但是效果变差了，原因是数据多了之后检索到的相关片段也多了，把之前本来最相关的片段挤掉





子曰大模型进展

- 2023年7月26日，网易有道正式发布“子曰”大模型
- 2023年11月4日，“子曰”教育大模型正式通过相关备案
  - 国内首个教育领域的垂直大模型
- 2024年1月3日，正式推出子曰教育大模型2.0

Show case：文言文翻译

- 一个偏科生的故事
- 词典笔
  - 红海
  - 用户痛点：英语、古文
  - 有道的标签：英语
- 突破口
  - 大模型：古文翻译

- 一个top1的核心能力
- 查词翻译
  - 首屏
  - 流量入口
  - 场景分类
- 持续迭代的场景和需求
  - 语文精讲

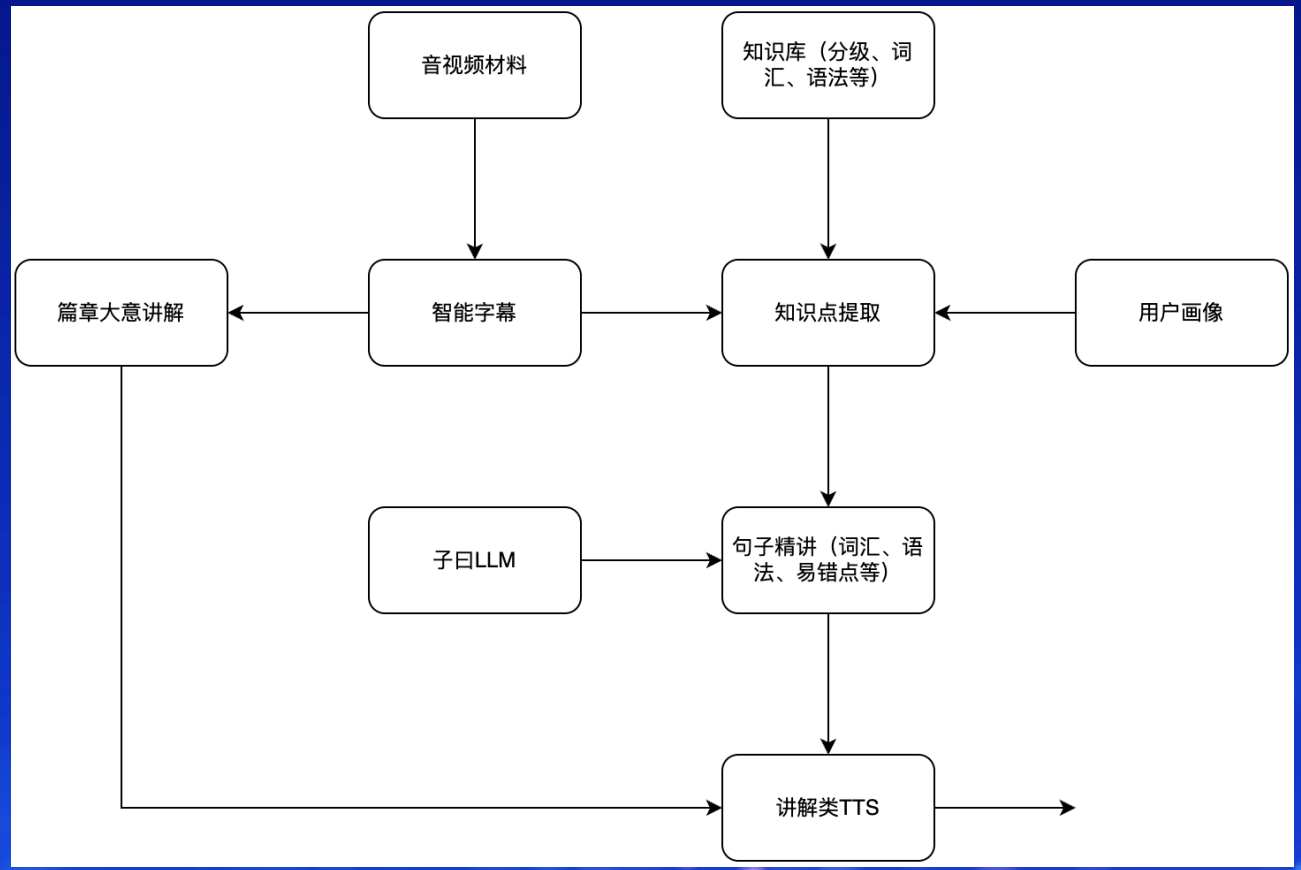
不止是LLM

- 打造更加极致的交互体验
  - 多模态大模型（语音、图像）
  - 传统ASR、TTS、OCR技术的结合
- 不止是WER
  - 断句、标点
  - 影响很多交互的基础
- 专有名词
  - 难，却非常重要

- 技术选型
  - Transformer、Conformer、Paraformer
  - OpenAI Whisper
  - Meta Massively Multilingual Speech
  - AudioPaLM、SpeechGPT
- 业务场景驱动
  - 教材
  - 文言文、古诗
  - 方言、多语种

打造极致的精讲能力

- 丰富的教育场景知识库
- 精准的教研、知识点的提取
- 智能的字幕技术（支持音视频的输入）
- 智能的子曰大模型讲解能力



亲切有温度的声音：类真人TTS技术

- 在教育场景，发音、听力、口语，相比文字更有感染力
- 从能发音，到动听、好听，标准、地道，甚至还要有口音
- 口语教练、AI老师，则需要AI更优人格魅力、亲和力和感染力
- 对TTS提出了越来越高的要求！

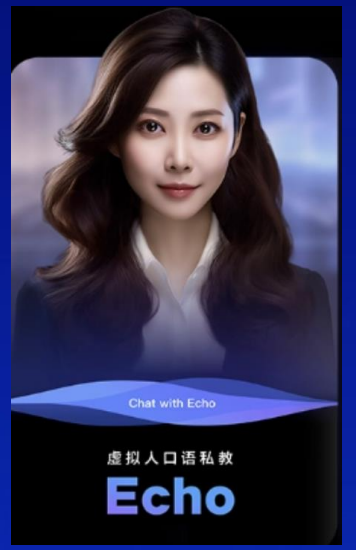
- 一些不同场景的样例
- AI爸妈讲故事
- 不同风格的网红老师、主播
- 来一段动画吧
-

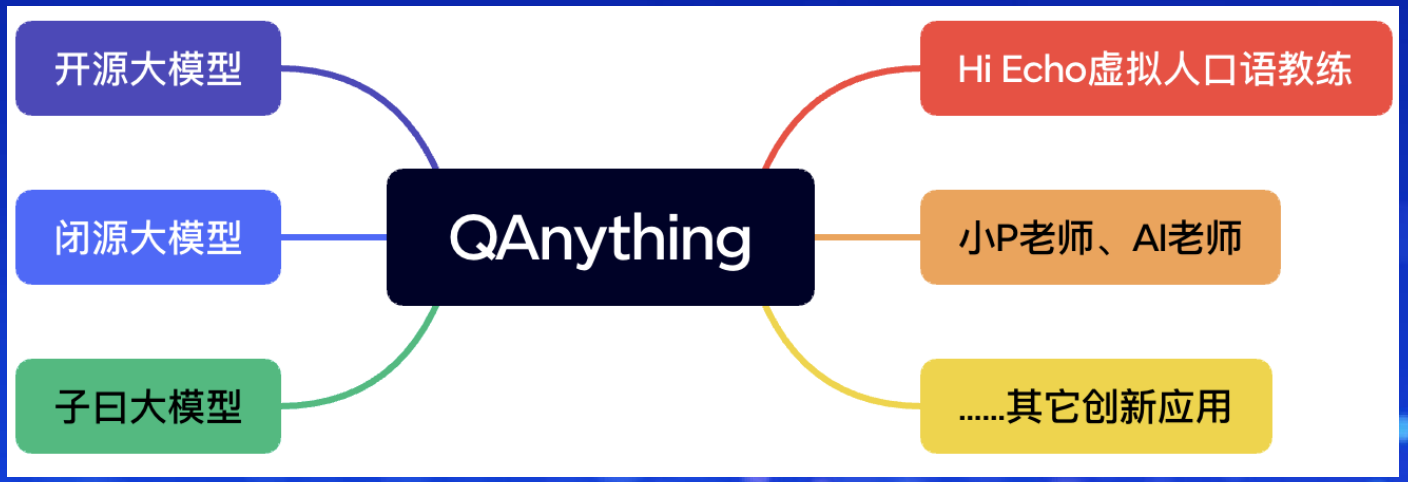
PART 03

创新应用：基于“子曰”的实践案例

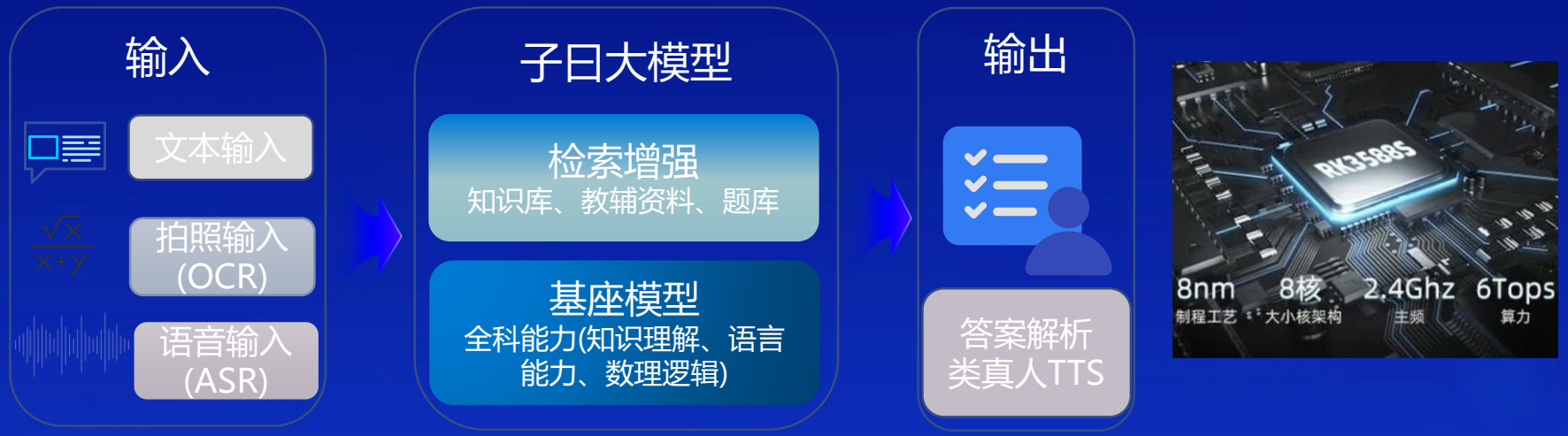
从LLM到应用的灵活架构

- LLM作为整个应用的大脑和核心
- 以自研RAG——QAnything为关键，扩展LLM的能力
- 基于LLM+QAnything搭建一系列关键应用
  - 不局限于某个单一大模型
  - 根据应用场景灵活组合





有道学习机X20 小P老师方案



小模型
端侧NPU处理

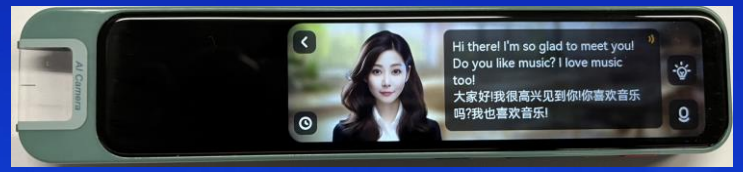
大模型
云端GPU处理

小模型
端侧NPU处理

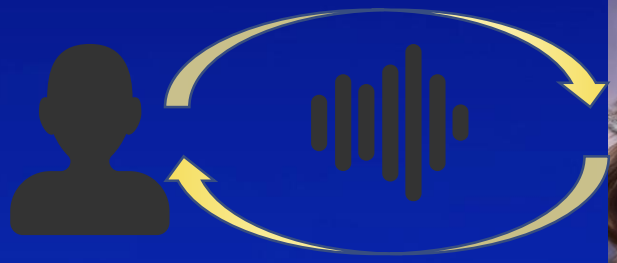
- 多模态识别，支持文本、图片、音频等多种形式输入
- 支持全学段、全学科（10科科目）讲解
- AI智能解答，支持答案解析、知识点、视频讲解和举一反三，支持多轮交互，苏格拉底式的教学

Hi Echo

- 全球首个虚拟人口语私教
  - 1v1口语专属教练
  - 海量对话场景和话题
  - 从小学到职场，分级
  - 对话评价，发音、语法、单词全面提升
- 2023年8月 词典笔X6 pro上线
- 2023年8月 听力宝Pro上线
- 2023年10月 APP上线



词典笔X6 Pro Hi Echo方案



语音识别ASR
语音合成TTS

虚拟人形象

子曰大模型

本地计算（RK3562 NPU）

云处理

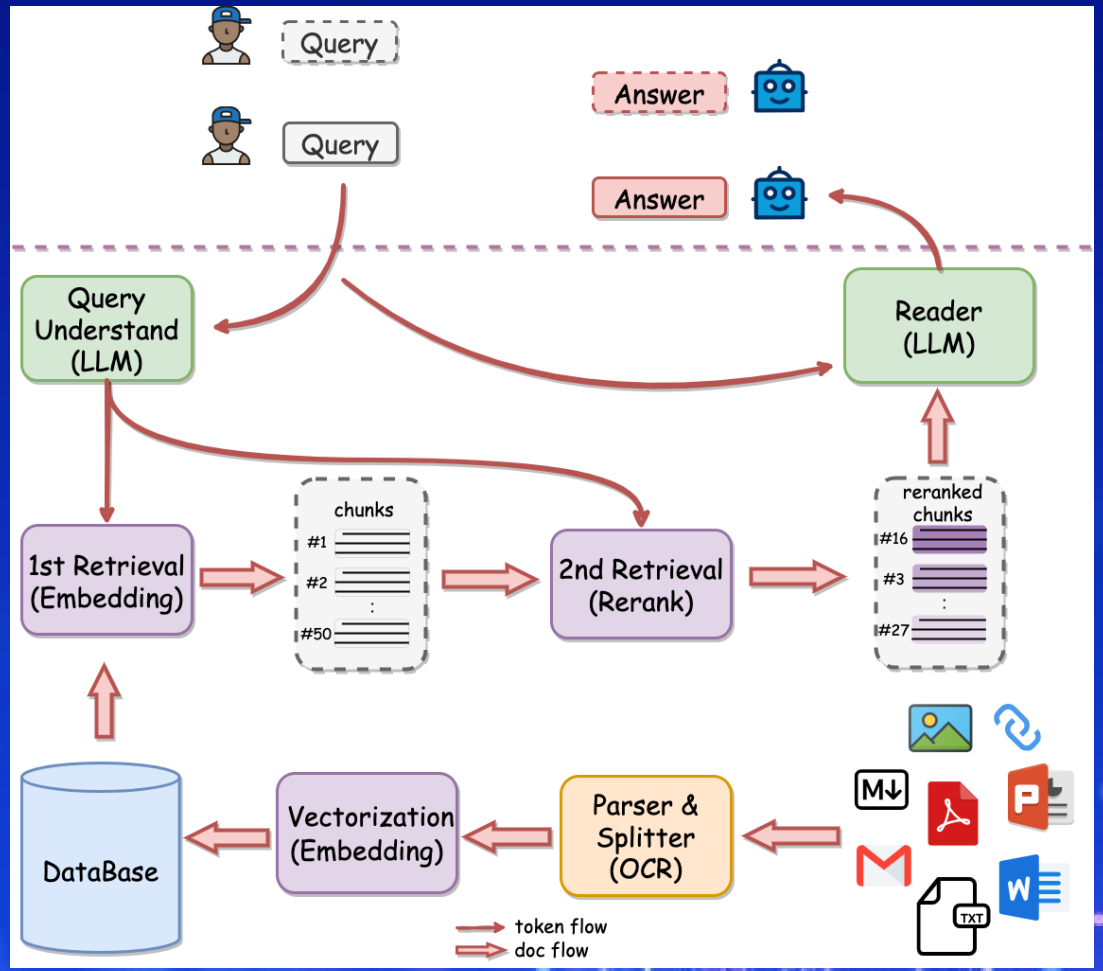
低延迟、低成本

PART 04

开源精神：有道的承诺与实践

QAnything——有道开源RAG引擎

- 关键模块
  - 文档解析
  - Embedding/rerank
  - LLM
  - vectorDB
- 主要流程
  - Query理解
  - 搜索
  - 相关性排序
  - LLM生成答案



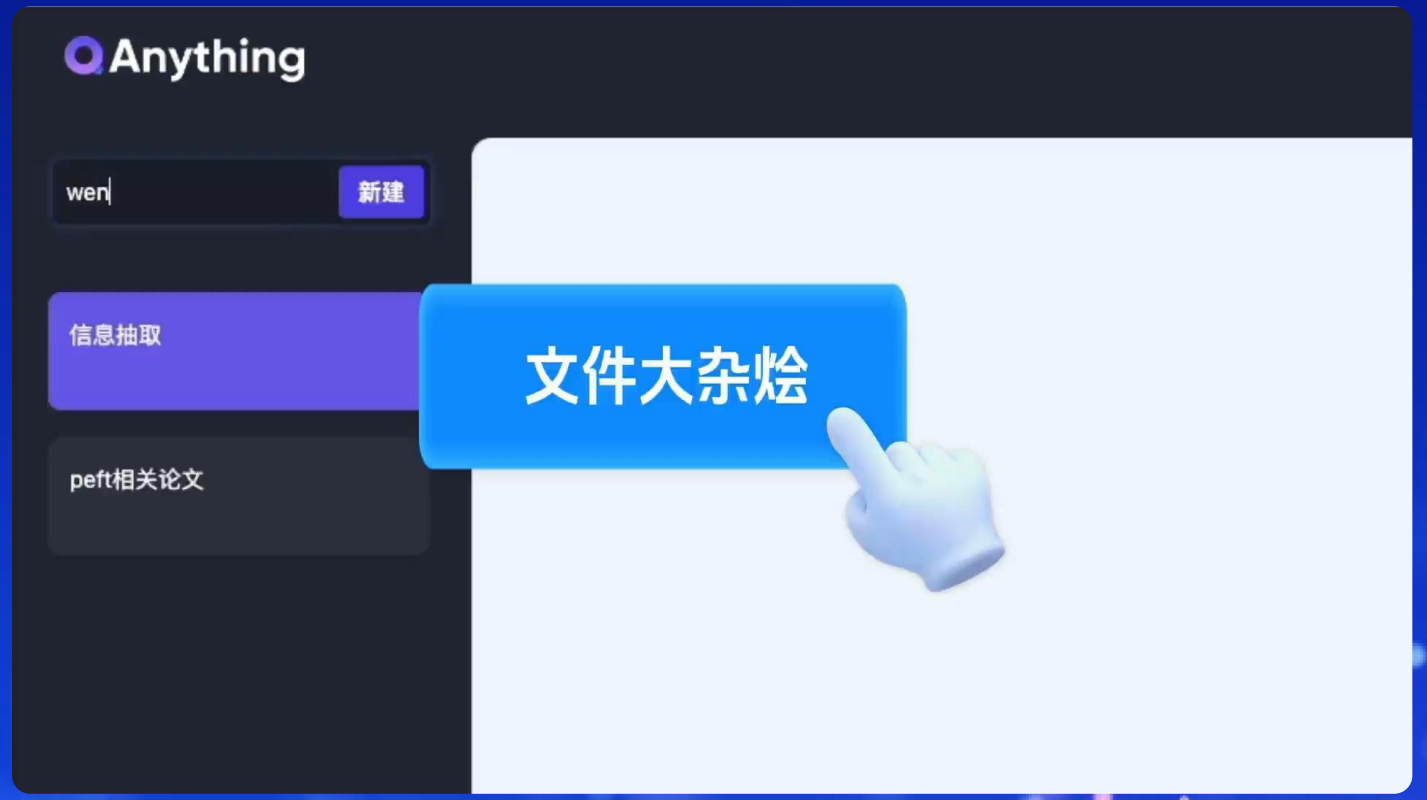
EmotiVoice易魔声——有道开源情感TTS引擎

- 为什么开源?
  - 行业发展趋势和潮流
  - 有道AI有2B的业务场景
  - 开源期待共建
- 开源短期目标
  - Star
  - 形成一定口碑&影响力
  - 更了解用户需求

- https://github.com/netease-youdao/EmotiVoice
- 刚上线仅一周时间就暴涨4200颗星，问鼎当周GitHub trending流行榜第一
  - 如今已达到 6.2k Star
- 上线后，完成多个版本迭代：
  - 提升易用性：demo page、docker、类openai tts的api
  - 支持任意中英文混合文本的合成，解决了一系列崩溃问题
  - 新增voice list、roadmap等一系列文档
  - 在智云上线 The EmotiVoice HTTP API
  - 发布Voice Cloning with your personal data
  - MAC版独立APP，12月底发布

QAnything——有道开源RAG引擎

- 202401-
- Retrieval-Augmented Generation
- 万物皆可问
  - doc,ppt,excel,pdf,图片等
  - 网页链接
  - 视频/音频
- 一键安装，快速使用
  - https://qanything.ai
  - https://github.com/netease-youdao/QAnything.git
  - 支持纯本地部署
- Github 8900 stars



QAnything——有道开源RAG引擎

- 跨语言问答
  - 中英日韩
- 多领域
  - 教育、医疗、法律
  - 金融、科研、客服
  - ...
- 竖排对比
  - Embedding
- 横排对比
  - Rerank
- 整体组合最优

Low → *High*

Embedding Models	WithoutReranker [*hit_rate/mrr*]	bge-reranker-large [*hit_rate/mrr*]	bge-reranker-v2-m3 [*hit_rate/mrr*]	CohereRerankMultilingualV3 [*hit_rate/mrr*]	***bce-reranker-base_v1*** [*hit_rate/mrr*]
OpenAI-ada-2	81.04/57.35	88.89/69.64	88.39/70.10	89.16/72.49	90.71/75.46
OpenAI-embed-3-small	83.01/58.10	89.20/69.86	89.20/70.22	90.02/72.78	90.91/75.49
OpenAI-embed-3-large	83.78/59.65	89.59/70.20	90.05/70.96	90.60/73.34	91.37/75.82
bge-large-en-v1.5	52.67/34.69	64.71/52.05	63.97/51.89	64.51/54.12	65.36/55.50
bge-large-zh-v1.5	69.81/47.38	80.11/63.95	79.33/64.64	79.95/66.23	81.19/68.50
bge-m3-large	84.67/61.25	89.94/70.17	89.40/70.57	90.13/72.93	91.72/76.14
llm-embedder	50.85/33.26	63.54/51.32	68.11/54.89	68.81/56.78	64.47/54.98
CohereV3-en	53.10/35.39	66.29/53.31	65.02/52.96	66.02/55.07	66.91/56.93
CohereV3-multilingual	79.80/57.22	86.76/68.56	86.15/68.72	87.11/71.04	88.35/73.73
JinaAI-v2-Base-zh	71.63/49.62	81.77/64.89	80.96/65.16	81.39/67.03	83.13/69.64
gte-large-en	53.17/34.71	65.02/52.04	64.05/51.86	64.55/53.68	65.67/55.87
gte-large-zh	59.48/39.38	71.56/58.27	70.67/58.89	71.01/60.22	72.37/62.71
e5-large-v2-en	61.03/40.67	71.52/56.61	70.90/57.01	71.28/58.62	72.37/60.91
e5-large-multilingual	79.14/55.54	87.35/68.50	86.92/68.77	87.73/71.15	88.97/73.81
bce-embedding-base_v1	85.91/62.36	91.80/71.13	91.02/71.43	91.87/73.34	*93.46/77.02* *96.36/78.93*(★)

(★): Hybrid search for top10 recall of bce-embedding and top10 recall of bm25, which are reranked together by bce-reranker subsequently.

https://github.com/netease-youdao/BCEmbedding

QAnything---开源RAG引擎

- AI to B
- 30000+用户
- 十几个行业
- 近百个进行中的订单
- 包括大型央国企等

科技生态圈峰会 + 深度研习

——1000 + 技术团队的共同选择

KEYLINK ing

K+峰会

K+峰会 上海站
K+全球软件研发行业创新峰会
时间：2024.06.21-22

K+峰会 敦煌站
K+思考周®研习社
时间：2024.10.17-19

K+峰会 香港站
K+思考周®研习社
时间：2024.11.10-12

K+峰会详情

AiDD峰会

AiDD峰会 上海站
AI+研发数字峰会
时间：2024.05.17-18

AiDD峰会 北京站
AI+研发数字峰会
时间：2024.08.16-17

AiDD峰会 深圳站
AI+研发数字峰会
时间：2024.11.08-09

AiDD峰会详情

THANKS