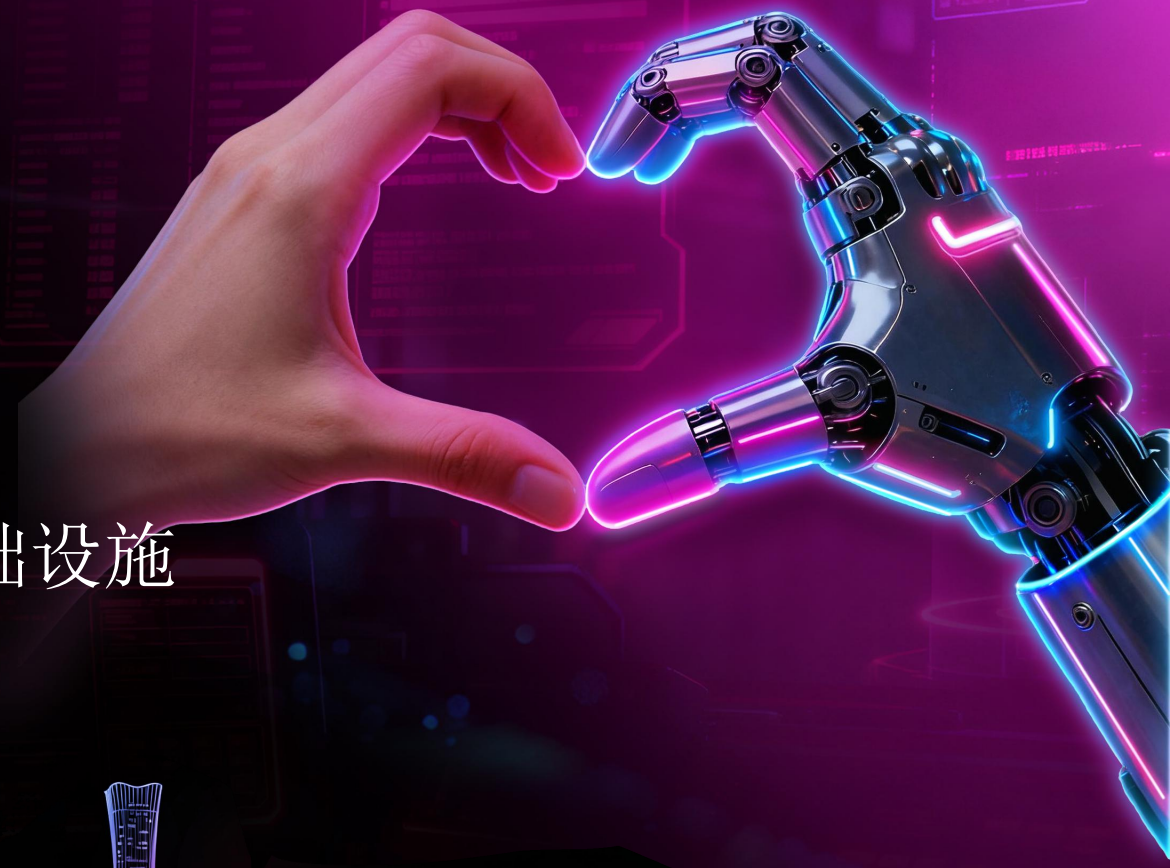
开源社 kaiyuanshe

AI Infra
CCF开源发展技术委员会

COSCon'25
第十届中国开源年会

众智开源 | Open Source, Open Intelligence



回顾与展望
——构建大模型时代的核心数据基础设施

张颖峰@RAGFlow

Data Infra 总览

开源社 kaiyuanshe

Application Middle Platform

ERP/SCM | CRM | Analytics

Applications

Micro Services/Service Mesh

Operational Databases

Analytics

- OCEANBASE
- SAP HANA HTAP
- ORACLE
- TiDB
- SingleStore
- PolarDB
- Microsoft SQL Server

TSDB

- influxdata

Search

- Elasticsearch

Operational Analytics

- Elasticsearch
- Druid
- ROCKSET

Serving

OLTP

- MySQL
- mongoDB
- aws
- Microsoft SQL Server
- ORACLE
- PolarDB
- OCEANBASE
- TiDB

NoSQL/Multi Model

- neo4j
- redis
- Lindorm

Data Middle Platform

Ingestion/Transport | Storage | Query Processing | Transformation | Analysis/Output

- CONFLUENT
- StreamNative
- Fivetran

OLAP

- aws
- snowflake
- GREENPLUM DATABASE

Lakehouse

- Apache hudi
- ICEBERG
- databricks
- presto

Real time OLAP

- Druid
- ROCKSET
- Clickhouse
- pinot
- StarRocks

Reverse ETL

- Census
- hightouch

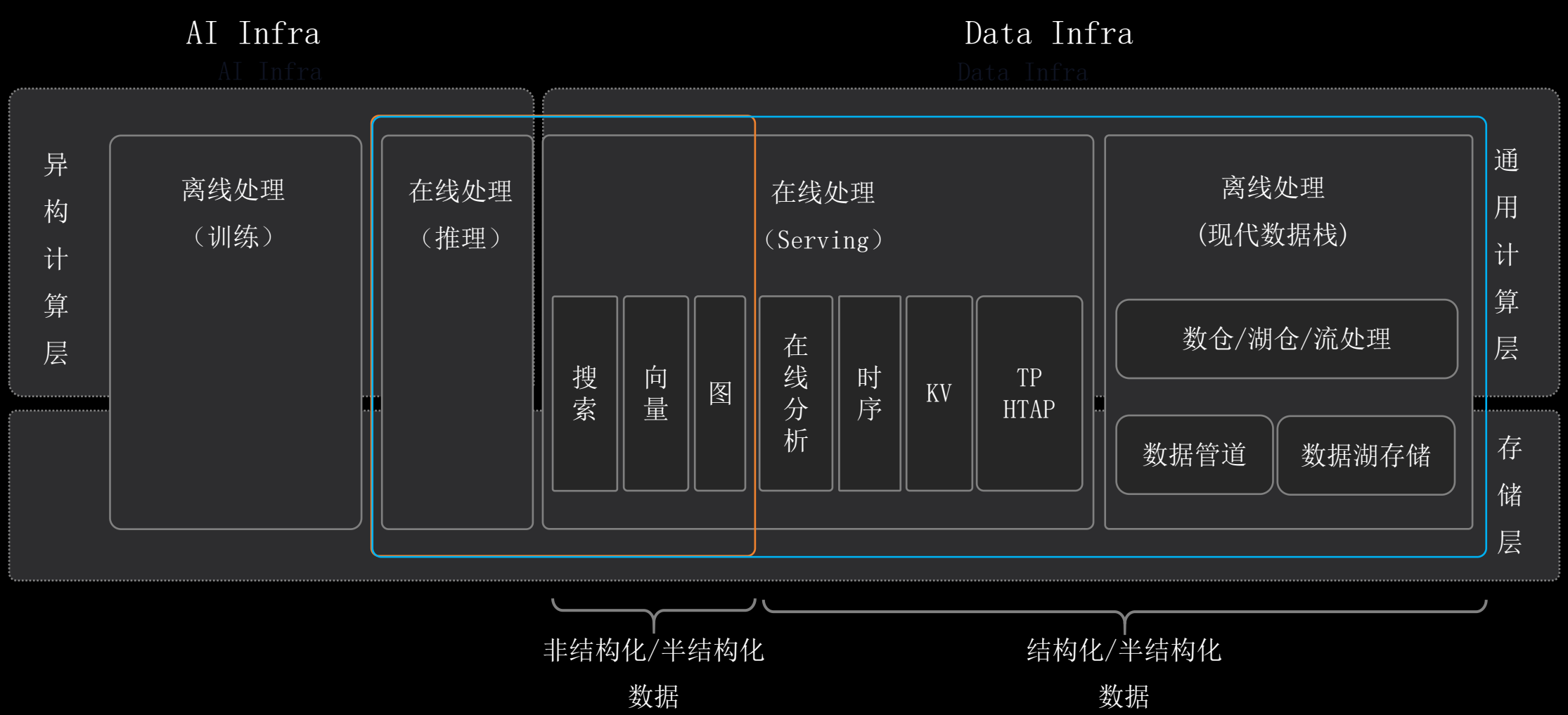
Stream Processing

- Flink
- databricks
- CONFLUENT
- Materialize

- dbt
- Looker PART OF Google Cloud

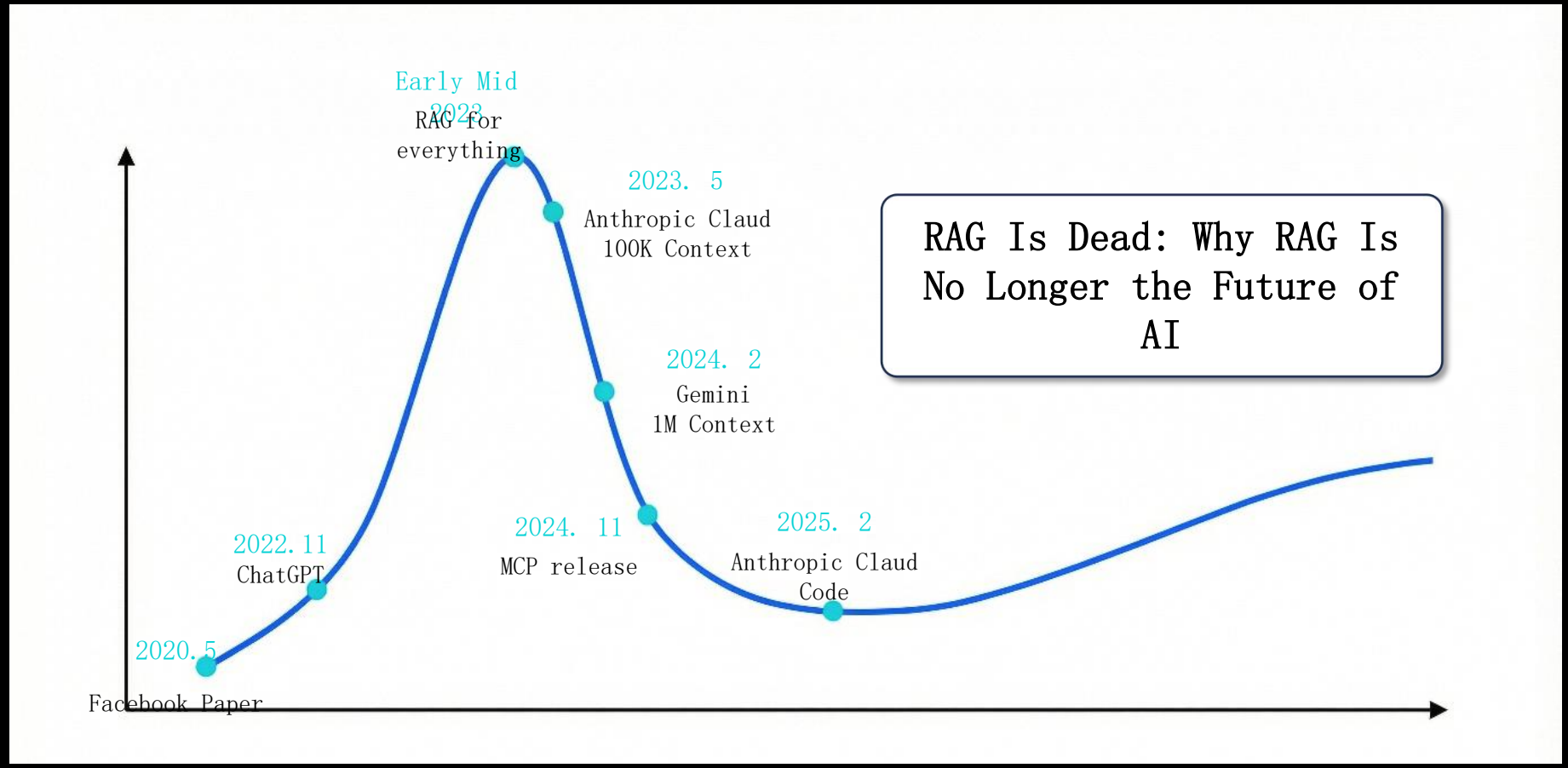
- Dashboards
- Embedded Analytics
- Augmented Analytics
- Data Workspace
- ML Tooling
- Custom Applications

AI Infra 和 Data Infra



RAG Is Dead

- RAG 效果不好
- Long Context 足够了
- 有 Agents 不需要 RAG
- Memory 比 RAG 效果好



向 LLM 提供输入的四种模式

- 只依靠 LLM 的 Context　　82 x 💰
- 借助于 KV Cache　　12 x
- 无需索引的搜索 => Grep　　💰 7 x 💰
- 基于搜索引擎 => RAG

Context Engineering 1.0

- 过长上下文导致成本激增和理解力下降

数据类型	Token消耗
单张技术图表	~1,000 tokens
典型代码库	数千万 tokens
企业级仓库	数亿 tokens
1,000页文档	600,000 tokens

- 为什么 Grep 不能满足要求

- RAG 就是 Context Engineering 1.0

过去一年 RAG 发展的共识

开源社
kaiyuanshe

切片很重要

用 VLM 改造的 OCR Parser 广泛使用

混合搜索被大多数用户接纳

但这些远远不够

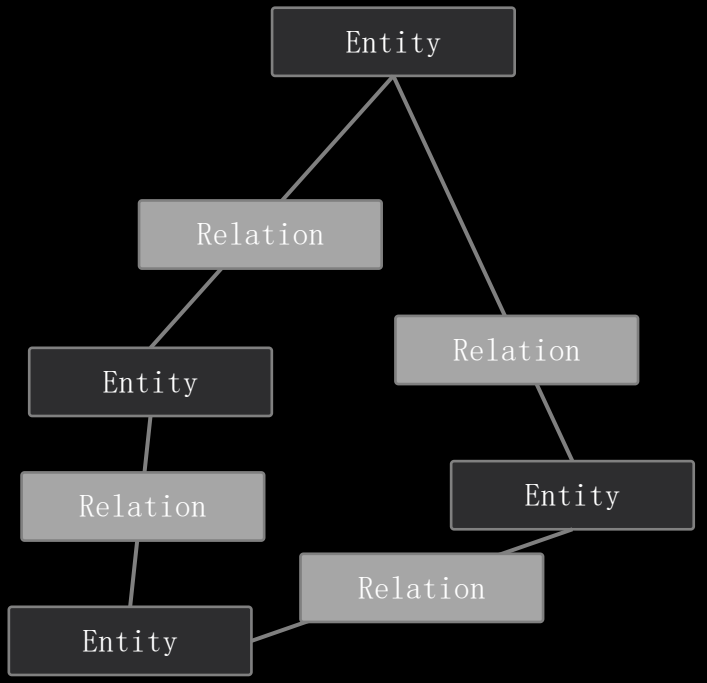
RAG 的痛点还有解么？

开源社
kaiyuanshe

各种方案都是

- Chunk 很难切得准
- Search 和 Retrieval 的矛盾
  - Search：扫描和定位，需要小块以提高语义匹配精度
  - Retrieval：阅读和理解，需要大块以提供完整上下文
- 性价比很低的 GraphRAG

RAG 的痛点应该如何解



Long Context RAG

借助于 LLM 弥补语义鸿沟

RAW Text

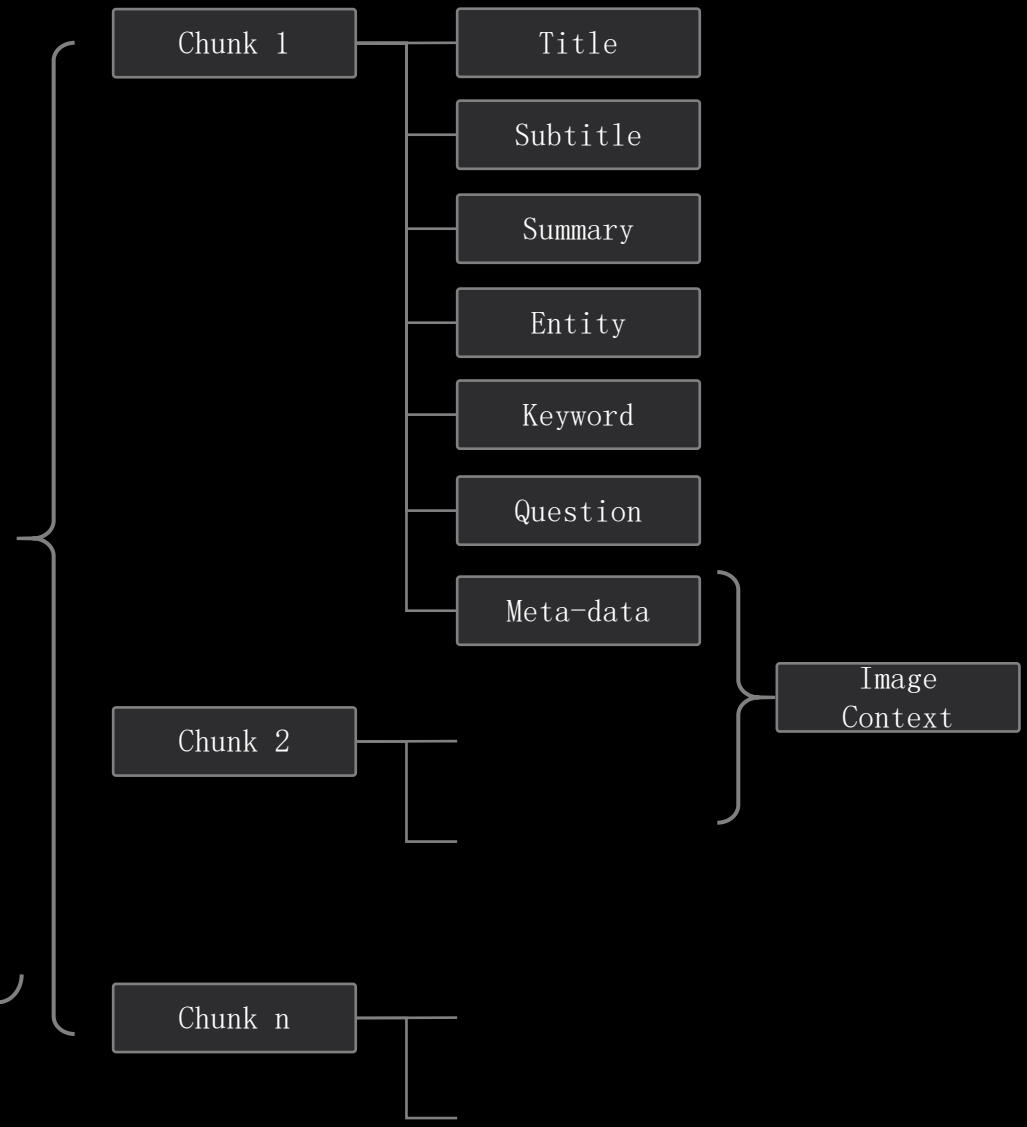
Indexing

Graph Structure

Tree Structure

Retrieval

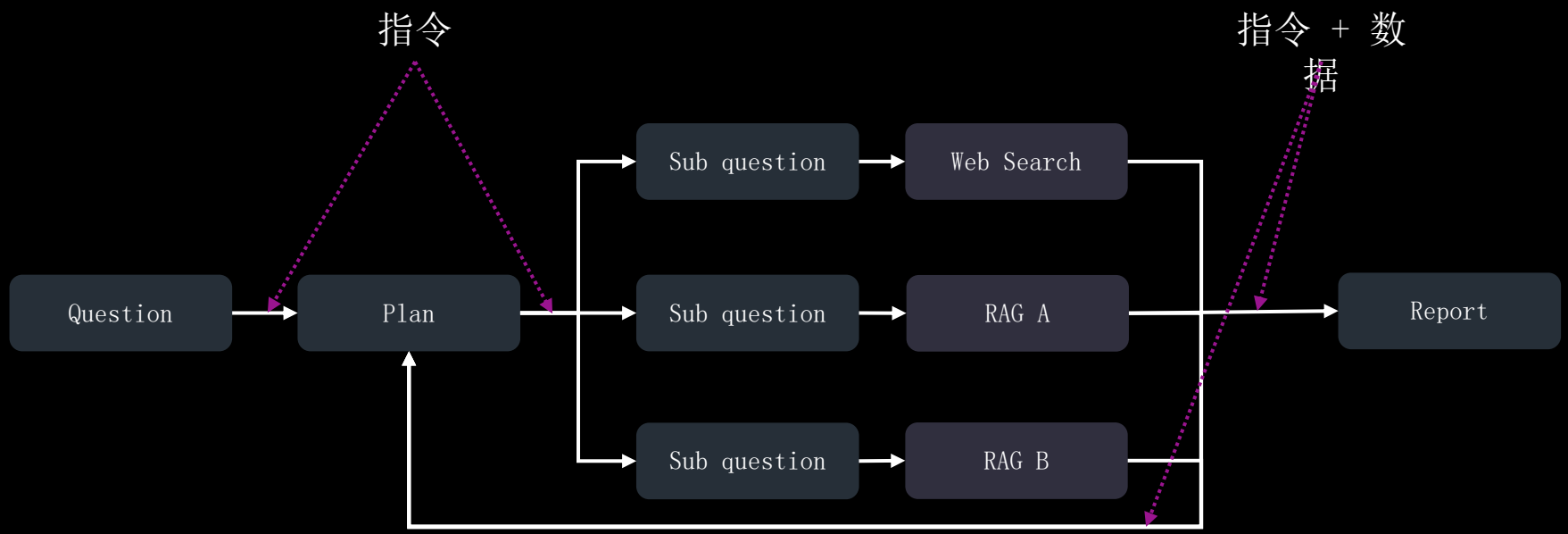
树/图结合



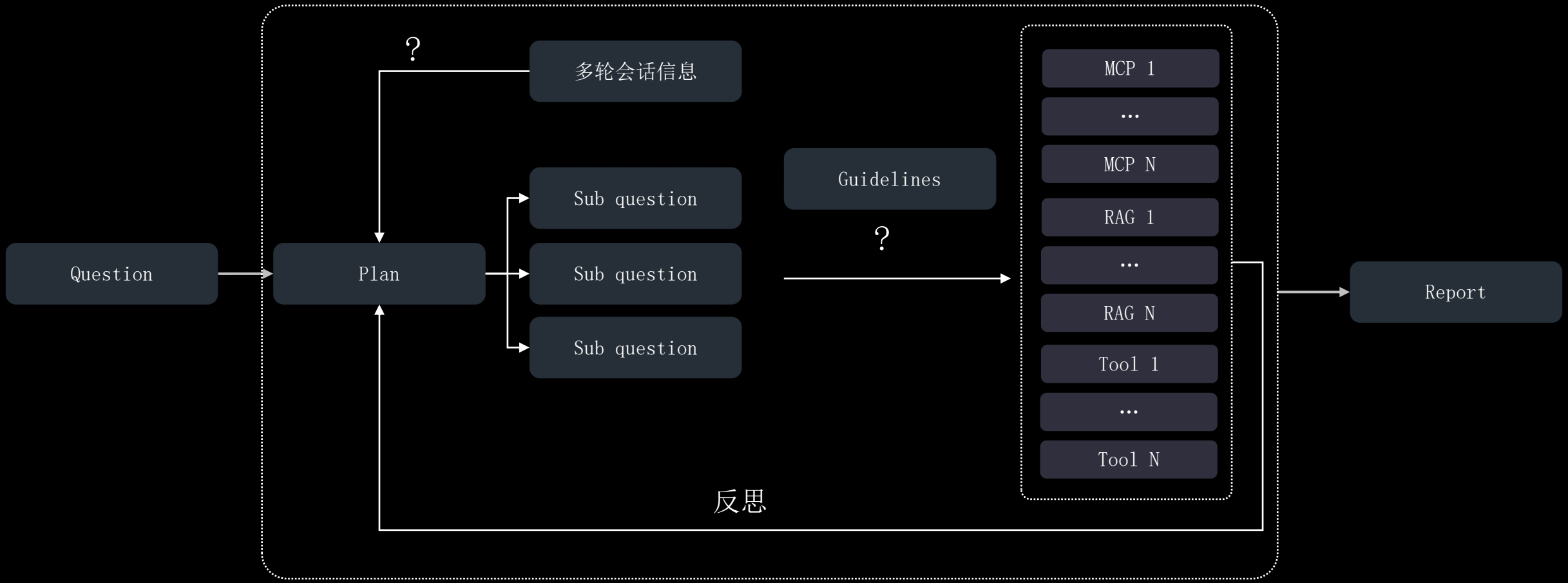
从 RAG 到 Agent

Context = 指令 + 数据

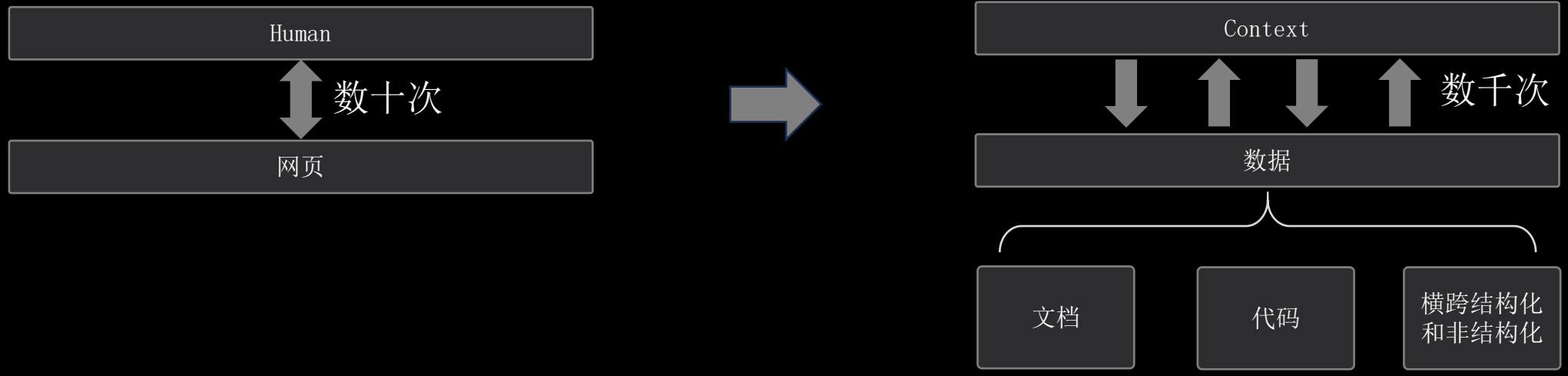
- 为什么需要 Agent
  因为 LLM 一次只能干一件事
- 为什么 RAG/Agent 密不可分
  反思无处不在
- 一切都是为了 Context



指令也不能无脑填充

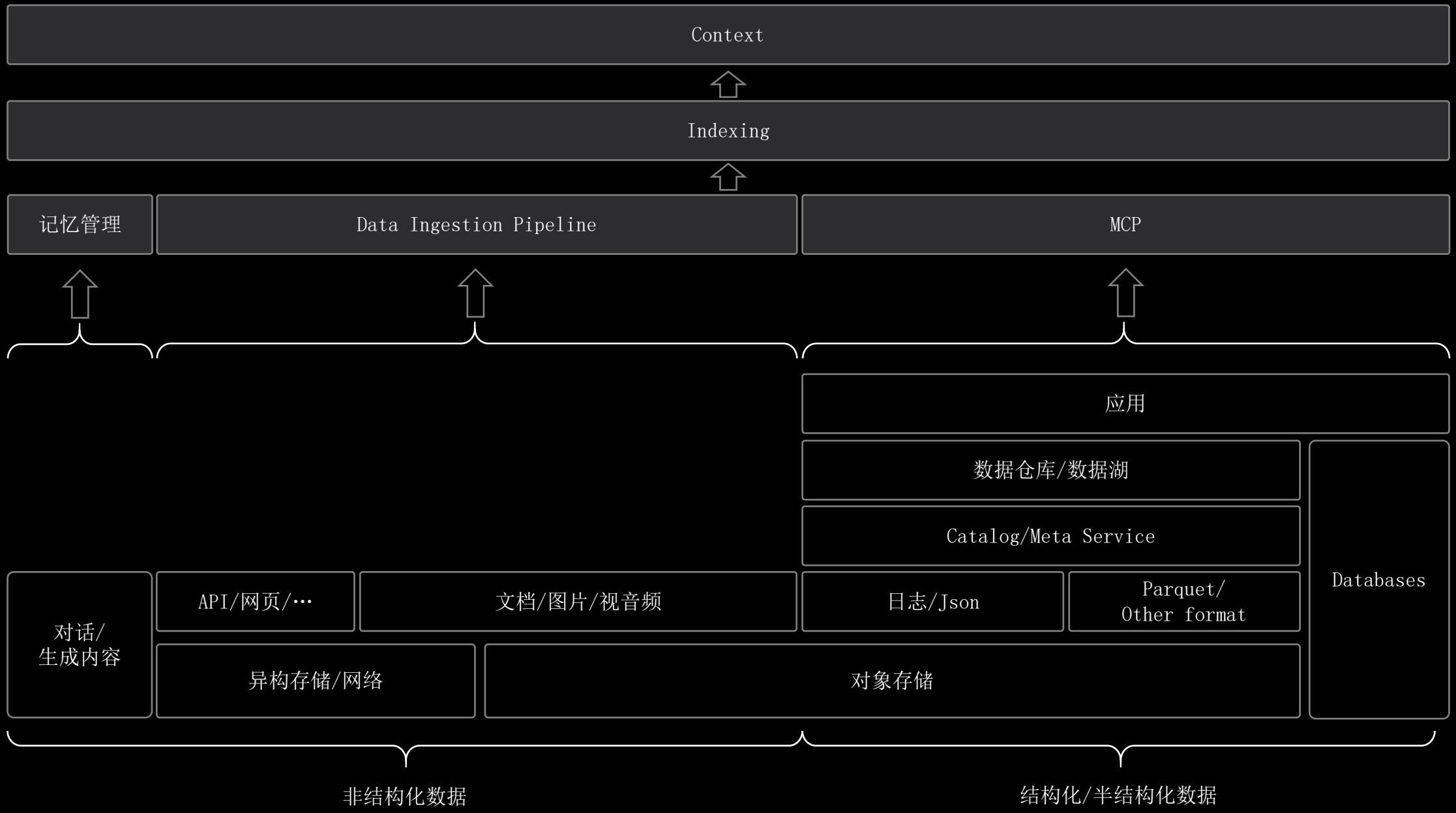

?
多轮会话信息
Question
Plan
Sub question
Sub question
Sub question
Guidelines
?
MCP 1
...
MCP N
RAG 1
...
RAG N
Tool 1
...
Tool N
Report
反思

为什么 Retrieval 对 Agents 如此重要

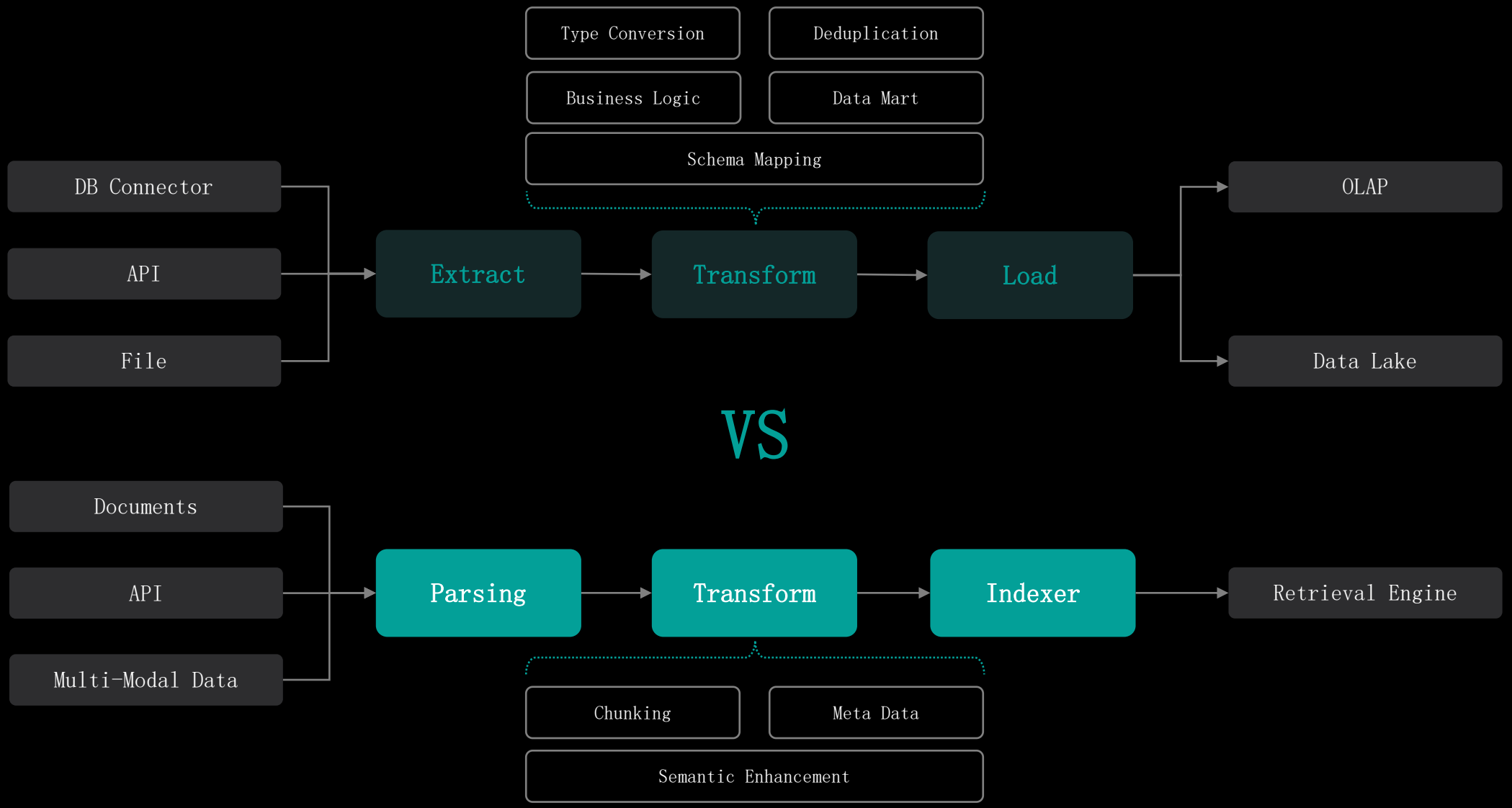


Agent Context 需要什么样的数据底座

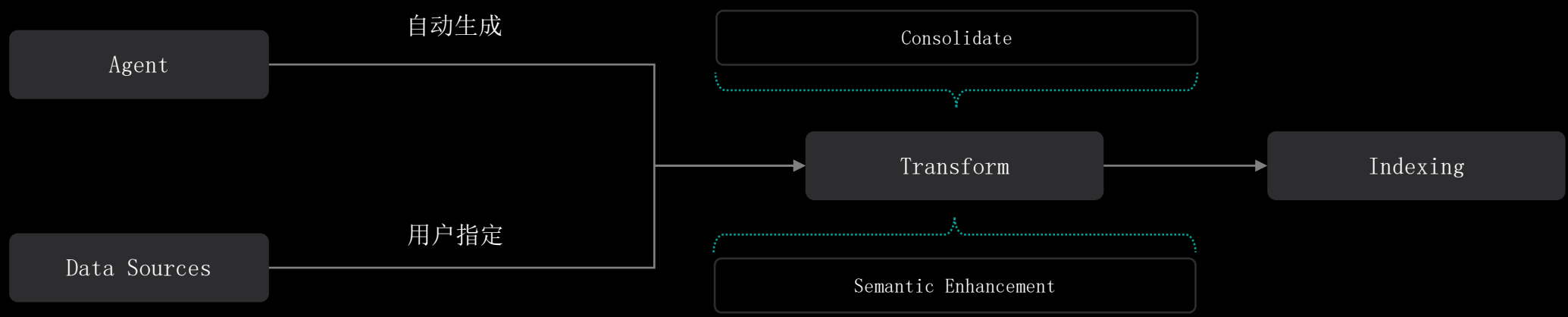
开源社 kaiyuanshe

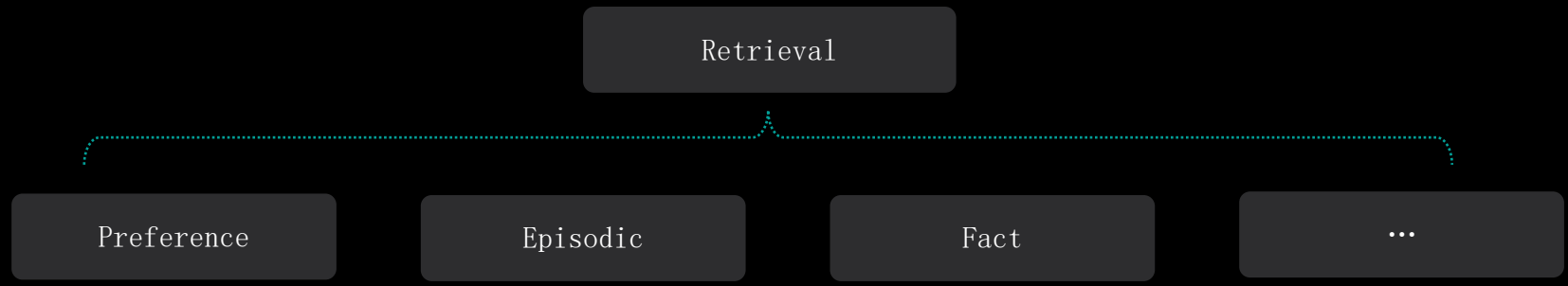


从 ETL 到 PTI


Type Conversion
Deduplication
Business Logic
Data Mart
Schema Mapping
DB Connector
API
File
Extract
Transform
Load
OLAP
Data Lake
VS
Documents
API
Multi-Modal Data
Parsing
Transform
Indexer
Retrieval Engine
Chunking
Meta Data
Semantic Enhancement

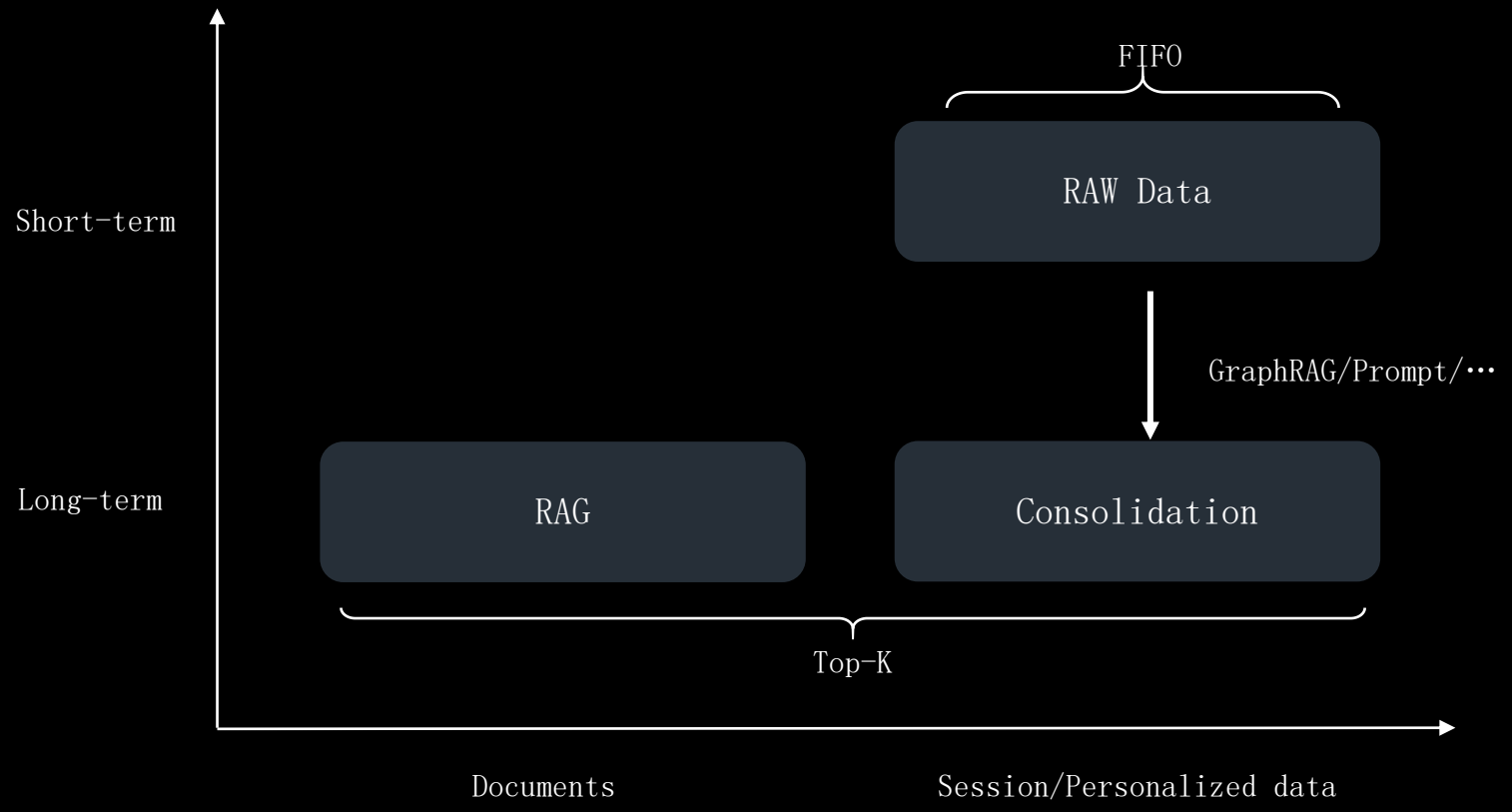
Memory 和 RAG——区别仅在于保存的数据



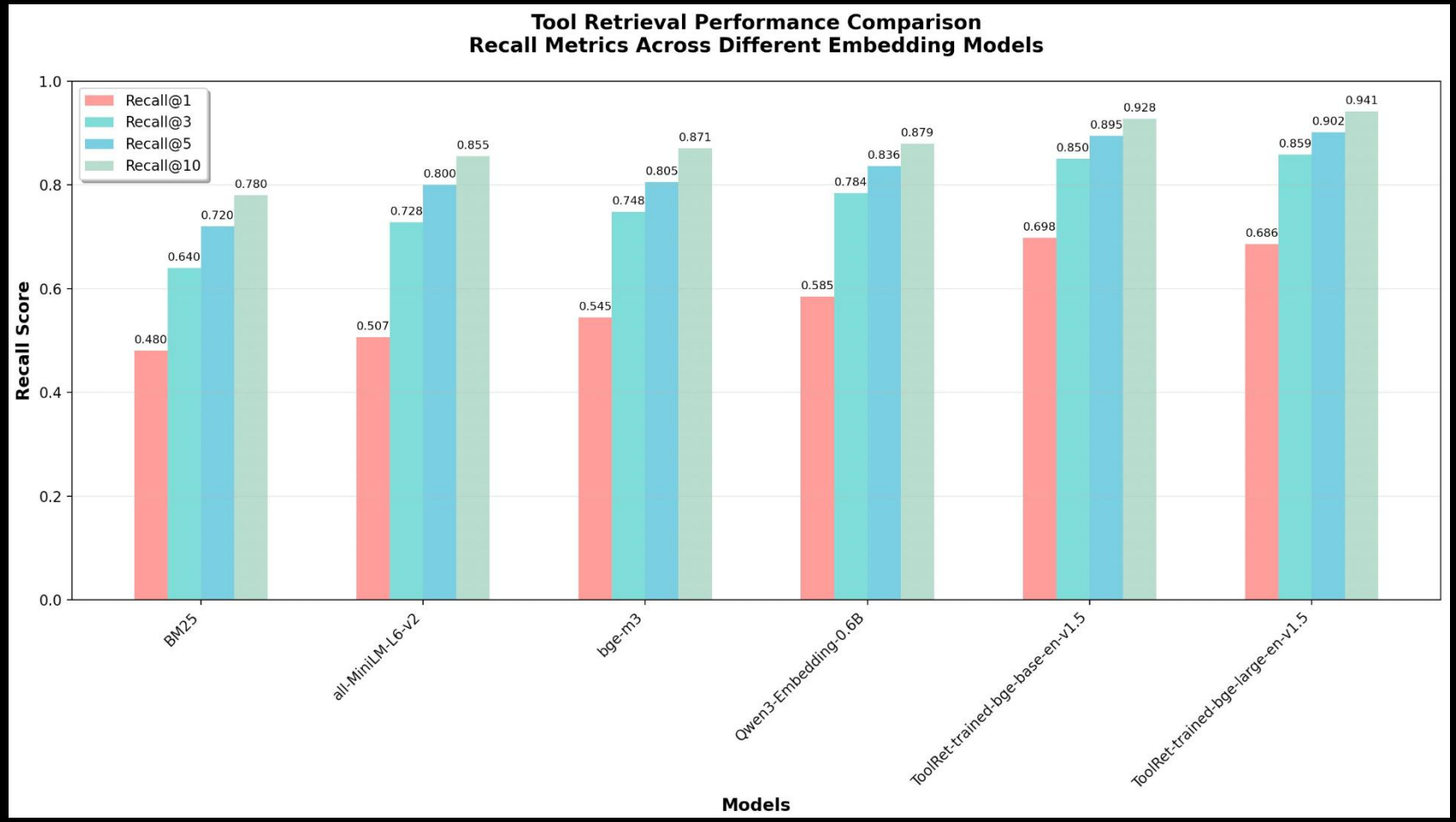


记忆可以成为独立平台么？

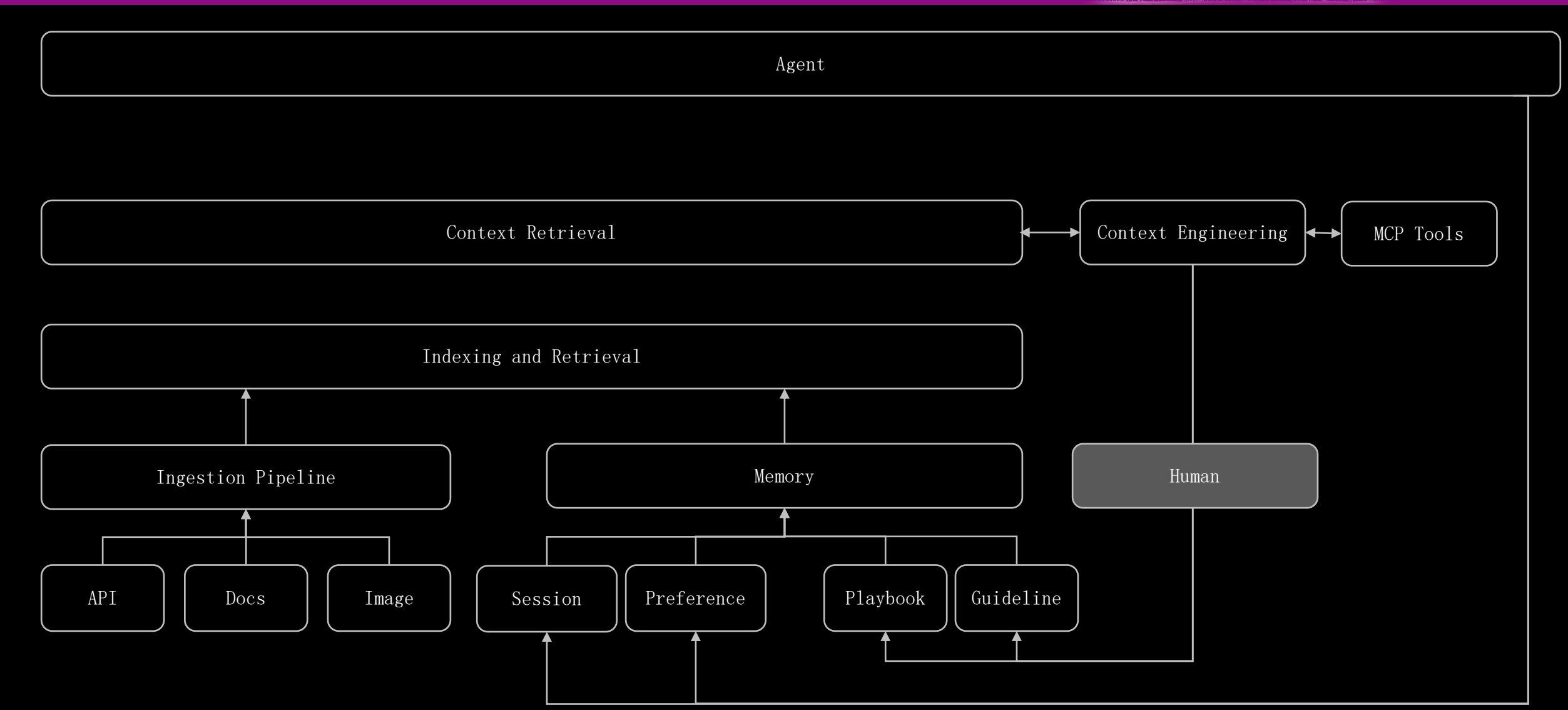
当下的 Memory 定义



Tools 也是需要 Retrieval 的



Tools 和 Guideline

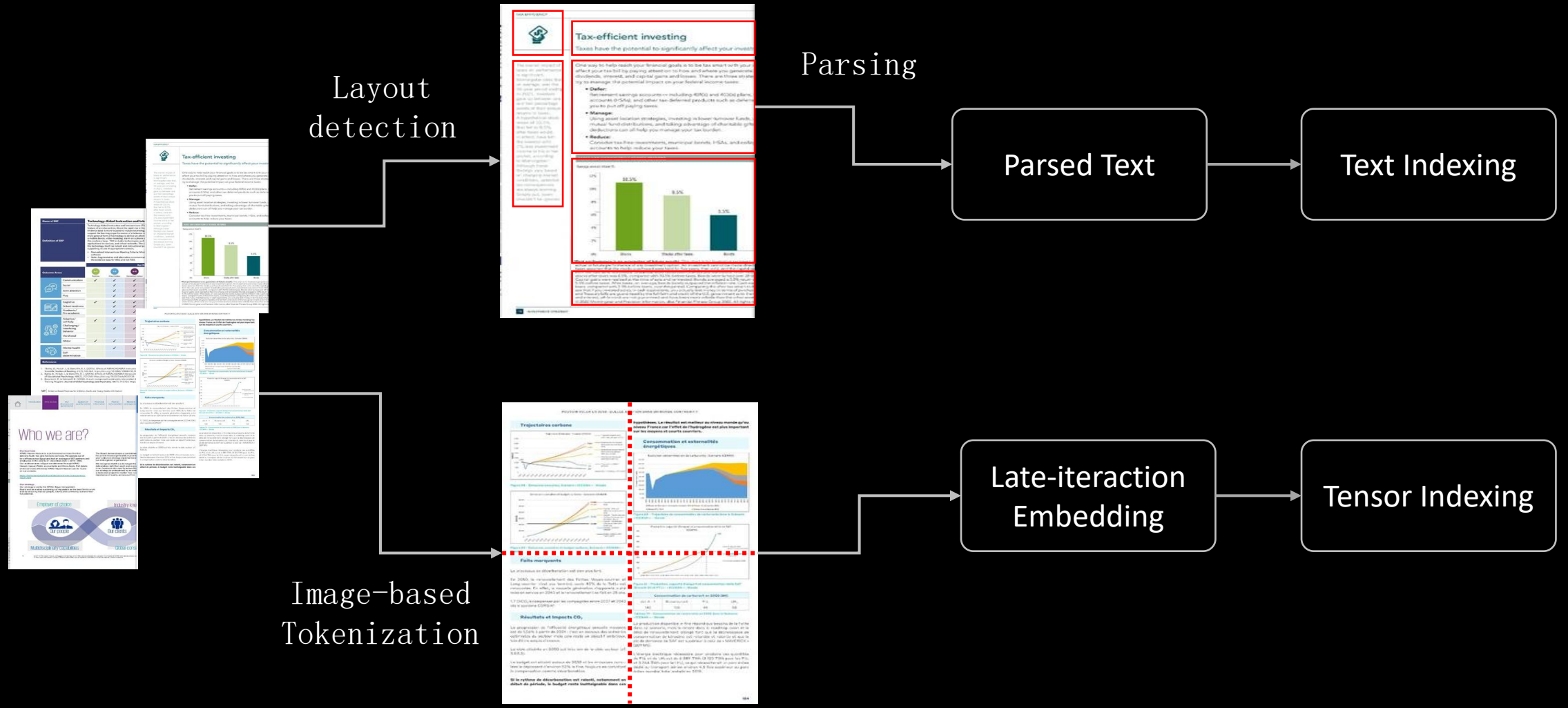


企业 AI 落地的下一个范式转移

维度	Context Engineering （现状）	Context Platform （未来）
Context 创建	FDE 完成	大多自动化
Context 组装	在 Workflow 当中进行 Prompt 硬编码	Context Retrieval
Context 维护	Vendor 手动维护	客户自主维护

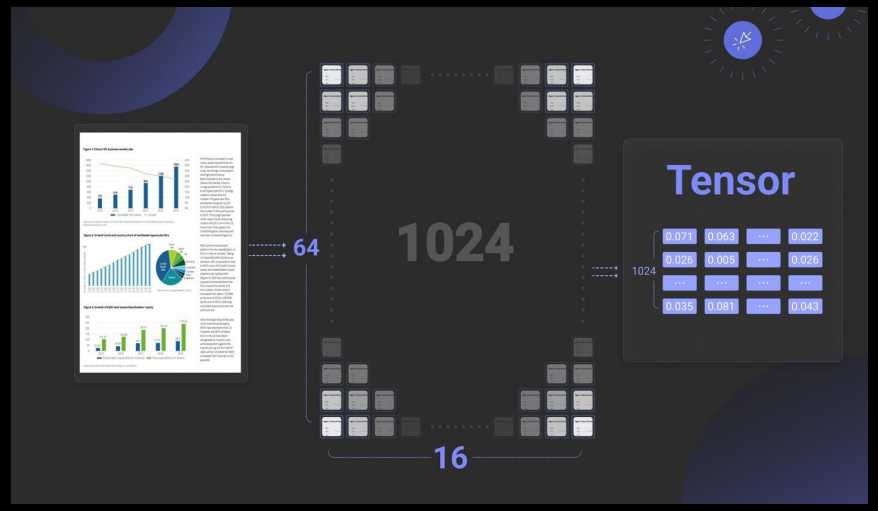
- Context 的"产品化"将是解锁下一代AI应用的关键
- 从 RAG 到 Memory 到 Context，代表了 LLM 和 Agent 应用范式的成熟

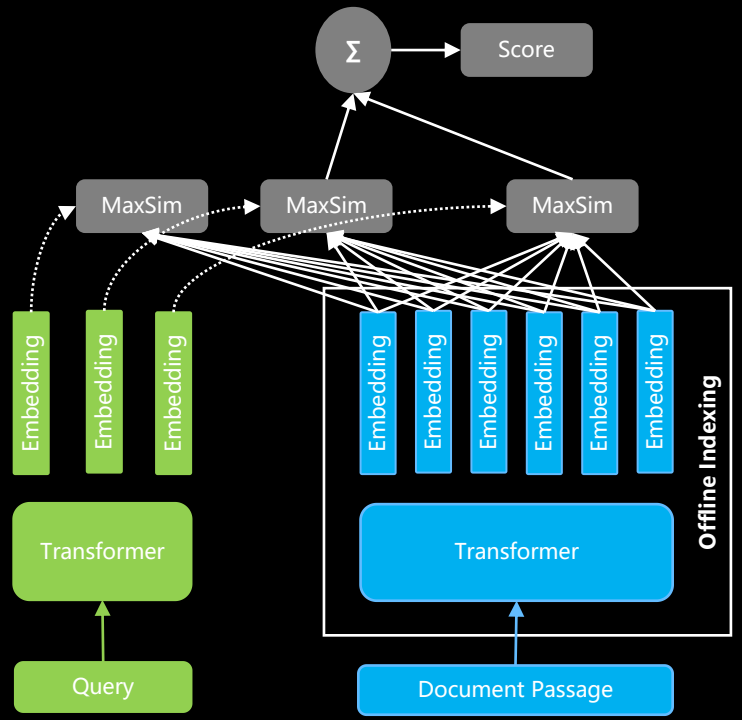
关于多模态和跨模态 Retrieval 的那些事情



关于多模态和跨模态 Retrieval 的那些事情

单向量还是多向量（Late Interaction）





Google DeepMind

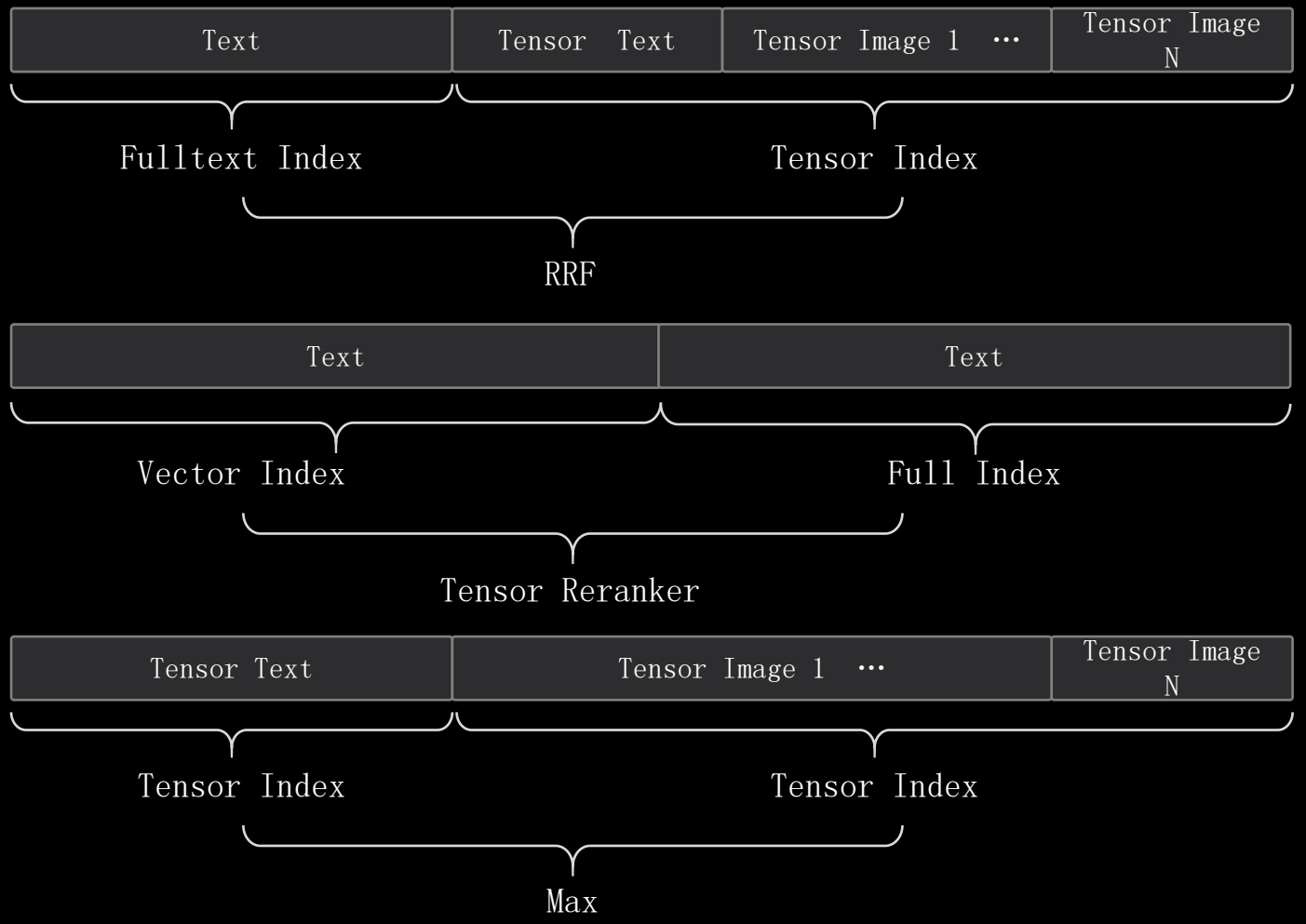
On the Theoretical Limitations of Embedding-Based Retrieval

Orion Weller[*,1,2], **Michael Boratko**[1], **Iftekhar Naim**[1] **and Jinhyuk Lee**[1]
[1]Google DeepMind, [2]Johns Hopkins University

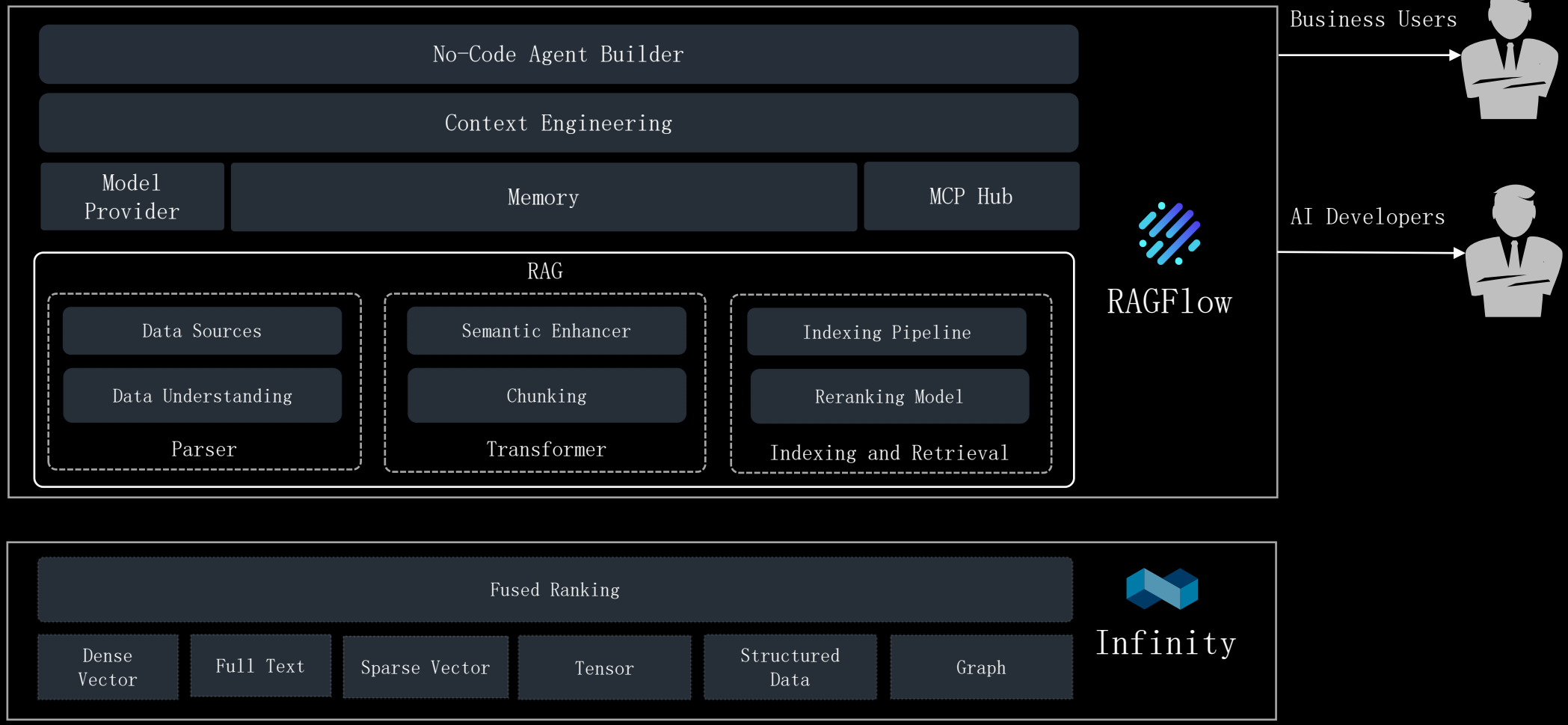
- 图文结合任务优势显著
- 单向量检索存在瓶颈

关于多模态和跨模态 Retrieval 的工程挑战

- 工程化挑战
  - 召回单元
  - 存储膨胀/计算复杂度
    - 训练对二值量化友好的 Embedding
    - Token 剪枝
      - MUVERA
      - Token Clustering
      - 改造VLM，根据注意力分布输出剪枝结果
      - 重新训练，直接在Tokenizer输出少的Token（ModernVBERT）



RAG is Dead? Long Live RAG !



开源社 kaiyuanshe

AI Infra
CCF开源发展技术委员会

COSCon'25
第十届中国开源年会

众智开源 | Open Source, Open Intelligence

Thanks