FROM POC TO PRODUCTION

Chaitanya Pathak

Chief Product and Technology Officer –Analyttica Datalabs

INSIGHTS FROM IMPLEMENTING ADVANCED RAG ARCHITECTURES

TALKING POINTS

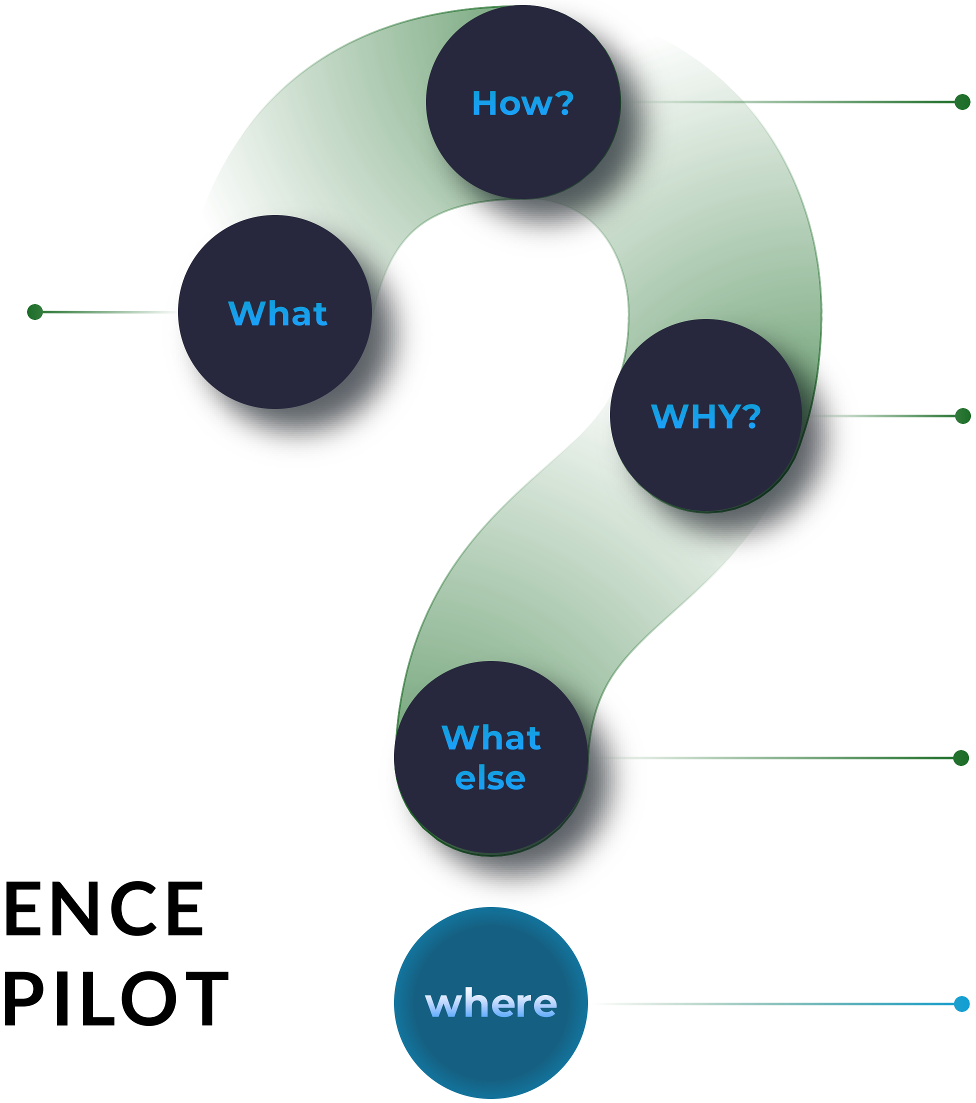
1. Context

2. Components

3. Prompts and RAG

4. Next Iteration

RAG BASED INFERENCE ENGINE - COPILOT



1. What can I do next

2. How can this be improved

3. Why are the results this way

4. What else can I do

5. Where can I find more details

EVALUATION LANDSCAPE

Approaches

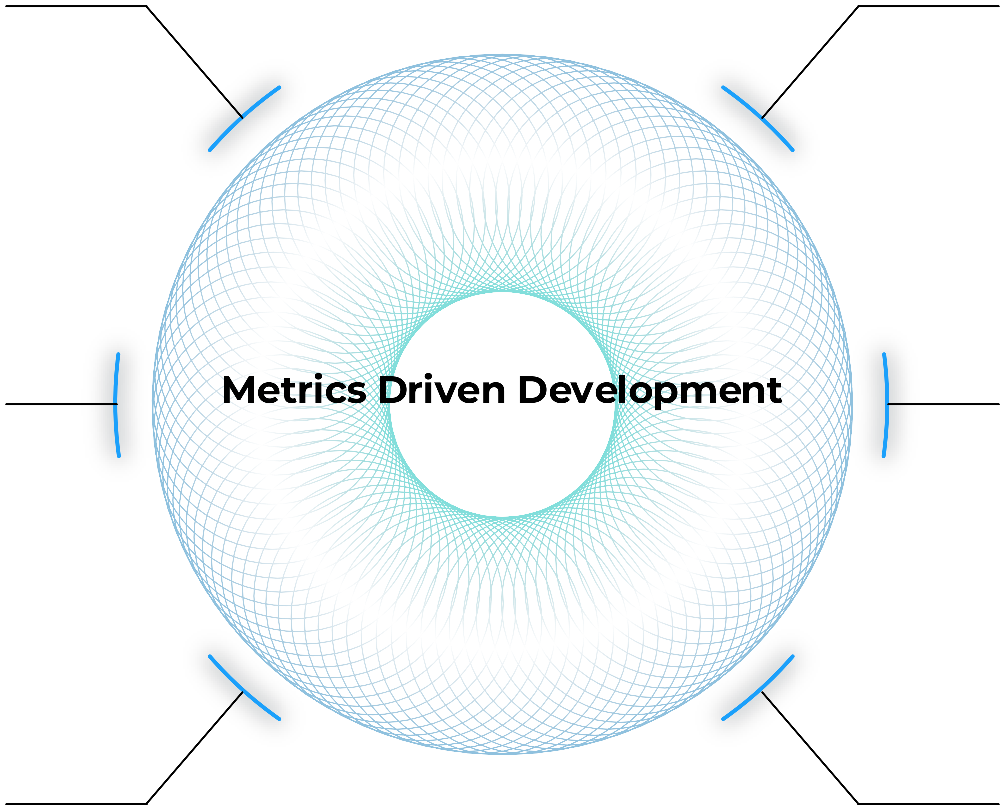
- Reference based
- LLM based
- Hybrid

Eval Data

- Building Gold standard data
- Human vs synthetic

IR Metrics

- Rank based (Hit Rate. MRR, NDCG
- Predictive (F1 score, precision , recall)
- Context Precision , Context Recall



Metrics Driven Development

Evaluation Scope

- End to end
- Component

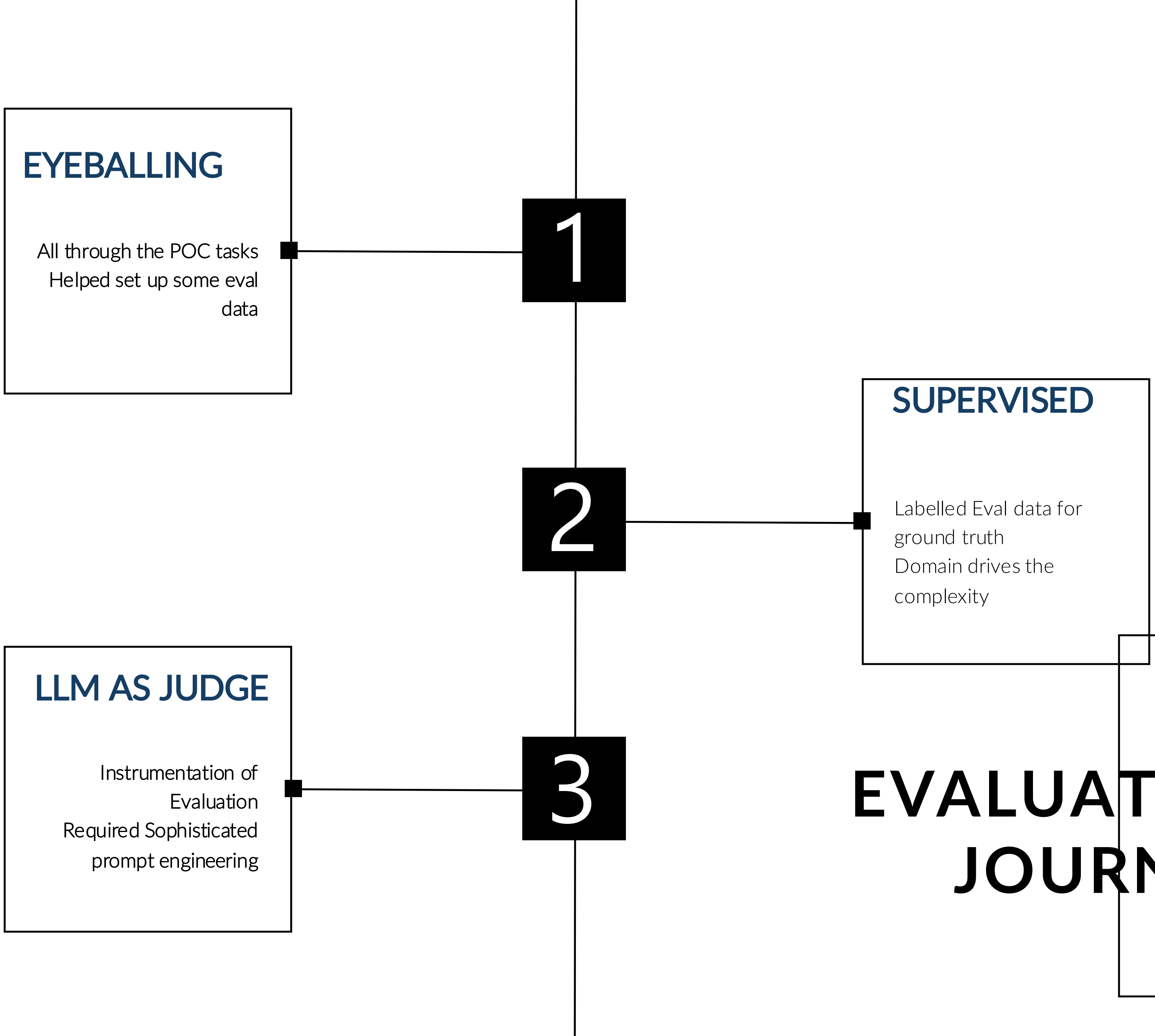
Generative

- Faithfulness
- Answer Relevancy

Design Considerations

- Deterministic vs nondeterministic
- Turns
- Prompt Flows

EVALUATION JOURNEY



1

EYEBALLING

All through the POC tasks
Helped set up some eval data

2

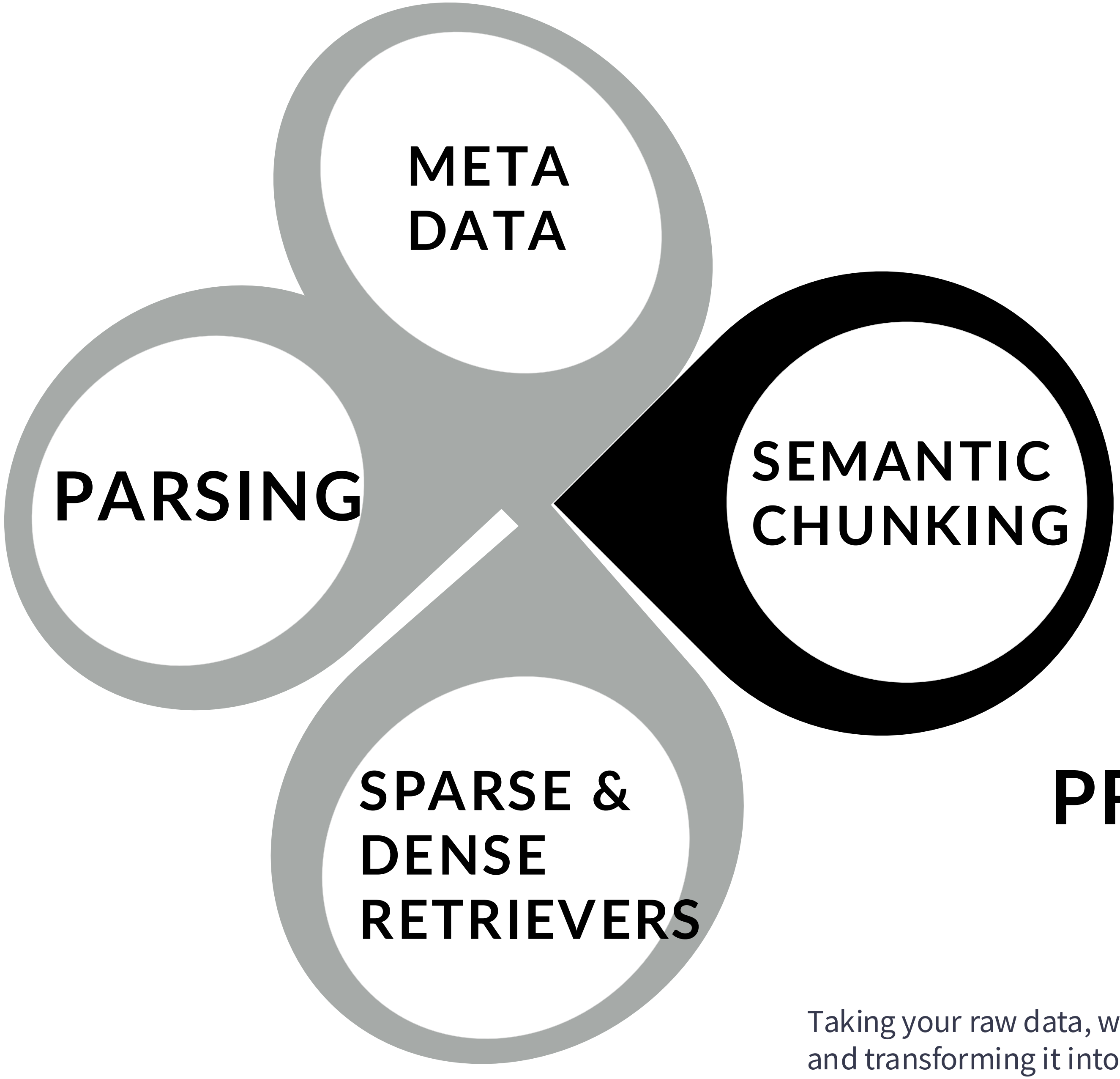
SUPERVISED

Labelled Eval data for ground truth
Domain drives the complexity

3

LLM AS JUDGE

Instrumentation of Evaluation
Required Sophisticated prompt engineering



DATA PREPROCESSING

Taking your raw data, which could be in various formats (documents, code, web pages, etc.), and transforming it into a format that your RAG system can understand and work with efficiently.

QUERY ENHANCEMENT –Benefits, Best practices and Tradeoffs

01

Intent

Helps s understand intent better

02

Breakdown

Breakdown complex queries into sub queries which are more consumable by the search , retrieval and LLM models

03

Tradeoffs

Latency and complexity

An essential part of Rag systems that improves retrieval accuracy and relevancy of responses

04

HyDE

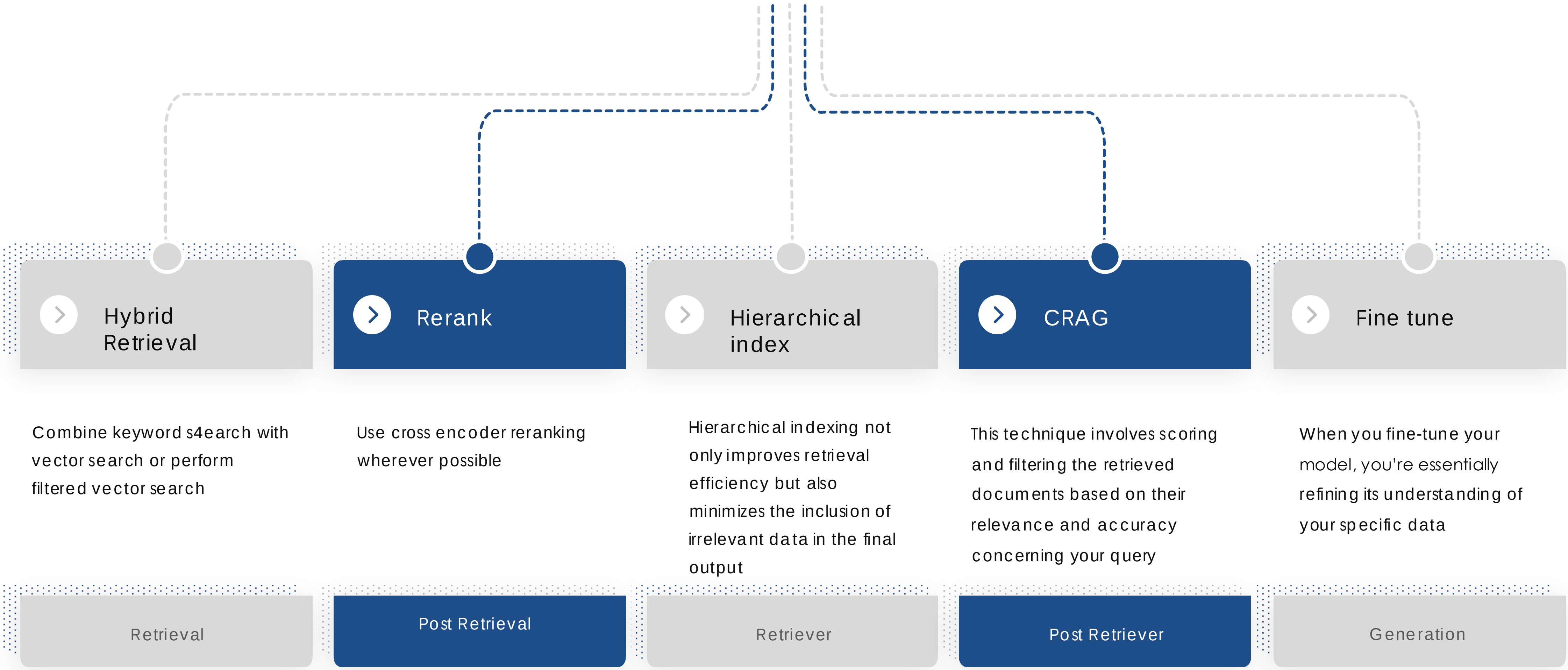
Generate documents / context wherever not available

05

Routing

This is essential best practice in a multi database / context scenarios

Retriever and Reranking



Hybrid Retrieval

Combine keyword s4earch with vector search or perform filtered vector search

Retrieval

Rerank

Use cross encoder reranking wherever possible

Post Retrieval

Hierarchical index

Hierarchical indexing not only improves retrieval efficiency but also minimizes the inclusion of irrelevant data in the final output

Retriever

CRAG

This technique involves scoring and filtering the retrieved documents based on their relevance and accuracy concerning your query

Post Retriever

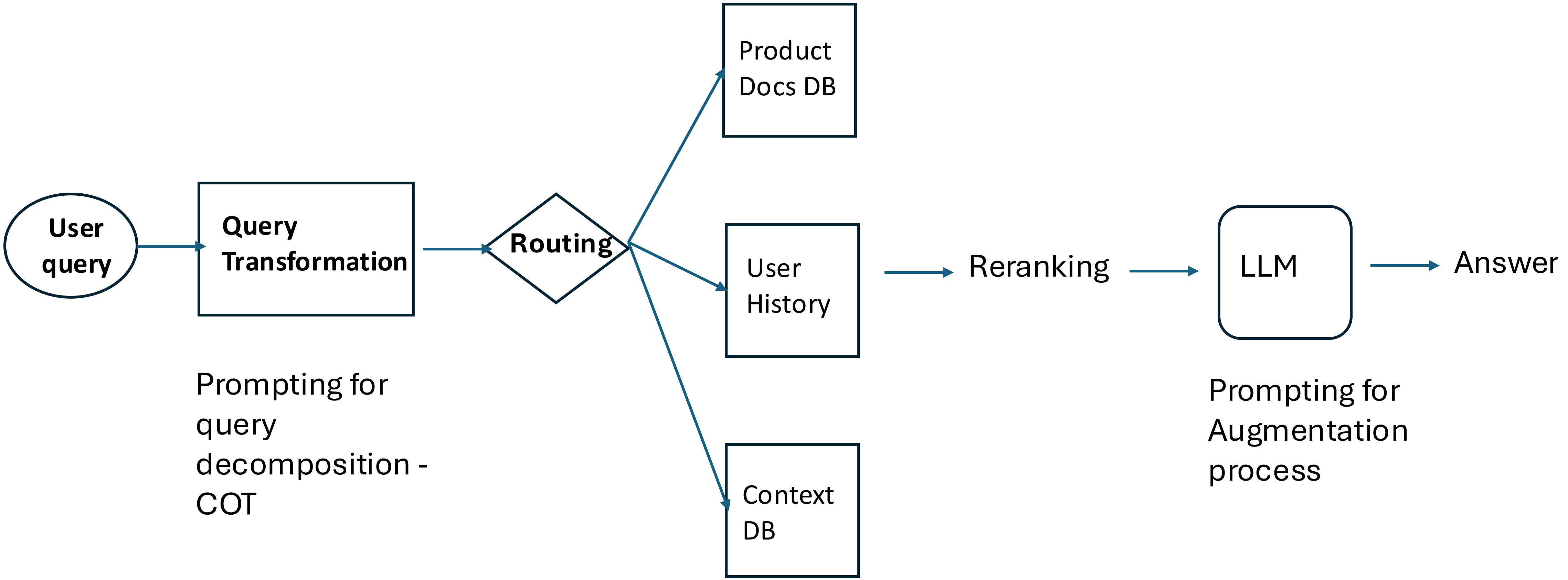
Fine tune

When you fine-tune your model, you're essentially refining its understanding of your specific data

Generation

LEAPS PROGRAMS

Prompt engineering and RAG architecture



Prompting for query decomposition - COT

Prompting for Augmentation process

Take away - careful prompt engineering will be employed for component(s) of RAG pipeline that uses LLM

NEXT STEPS

- Move towards modular RAG architecture
- Full versioning of prompts
- Unit testing – Promptfoo
- Agentic abilities – use of tools

HAPPY TO TAKE QUESTIONS
CHAITANYA.PATHAK@ANALYTTICA.COM