Agentic Relevance Tuning

Stavros Macrakis, OpenSearch @ AWS
Daniel Wrigley, OpenSource Connections

Search is perfect!

— no CEO nowhere nohow

Relevance

- Everyone wants it
- No one knows how to achieve it
- It’s too much work

More specifically

Customers can’t find iPhones

Revenue is down

Boss says “just use AI”

THE VIRTUOUS CIRCLE OF SEARCH IMPROVEMENT

Search processing

- Lexical search
- Semantic search
- Sparse neural

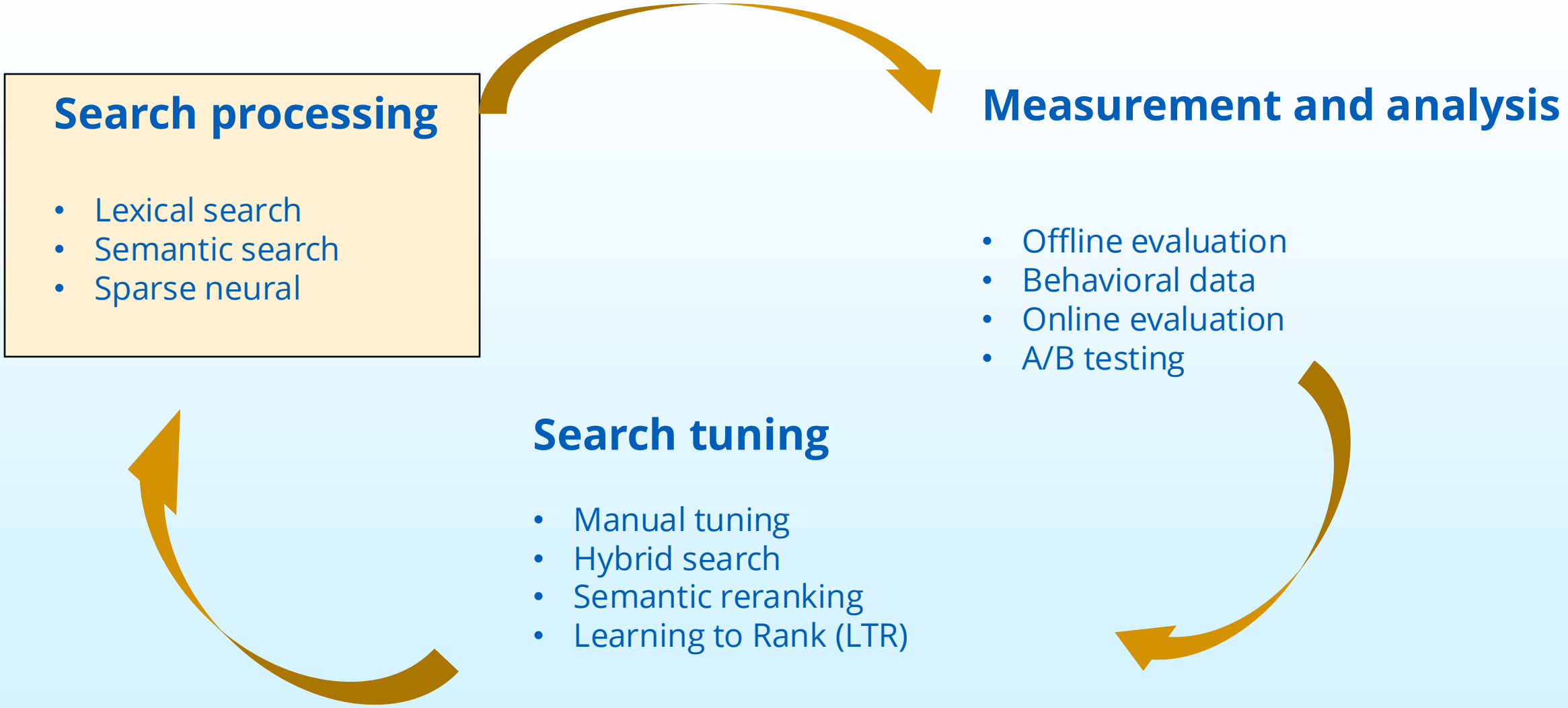
Measurement and analysis

- Offline evaluation
- Behavioral data
- Online evaluation
- A/B testing

Search tuning

- Manual tuning
- Hybrid search
- Semantic reranking
- Learning to Rank (LTR)

Lots of Progress



Lots of Progress

Search processing

- Lexical search
- Semantic search
- Sparse neural

- Embeddings
- Fast aNN
- Compression
- Caching strategies
- GPU Acceleration

- Faster lexical search

- Sparse neural

Search training

- Manual tuning
- Hybrid search
- Semantic reranking
- Learning-to-Rank (LTR)

- LLM-based query understanding and rewriting
- LLM-based result synthesis (RAG)

Measurement and analysis

- Offline evaluation
- Behavioral data
- Online evaluation
- A/B testing

Some Progress

Search processing

- Lexical search
- Semantic search
- Sparse neural

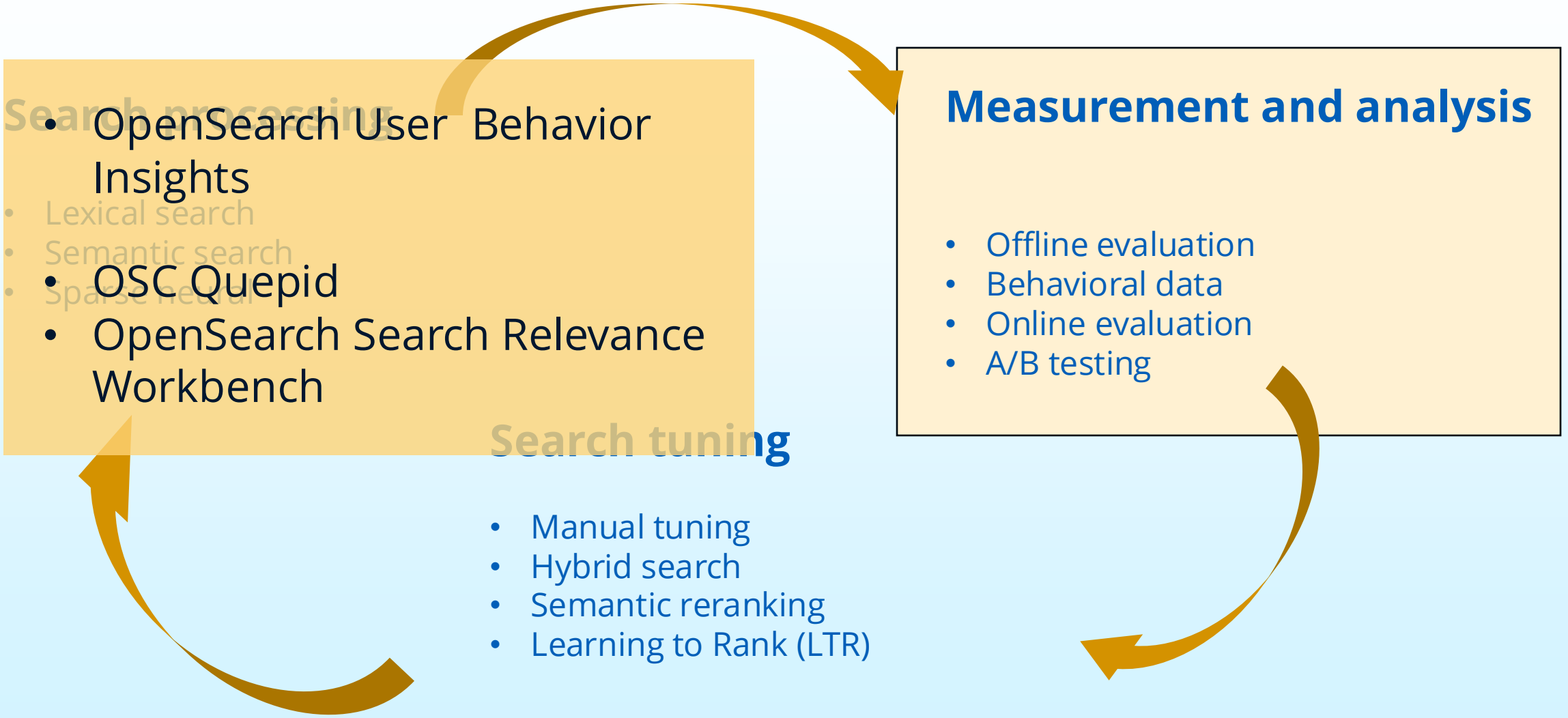
Measurement and analysis

- Offline evaluation
- Behavioral data
- Online evaluation
- A/B testing

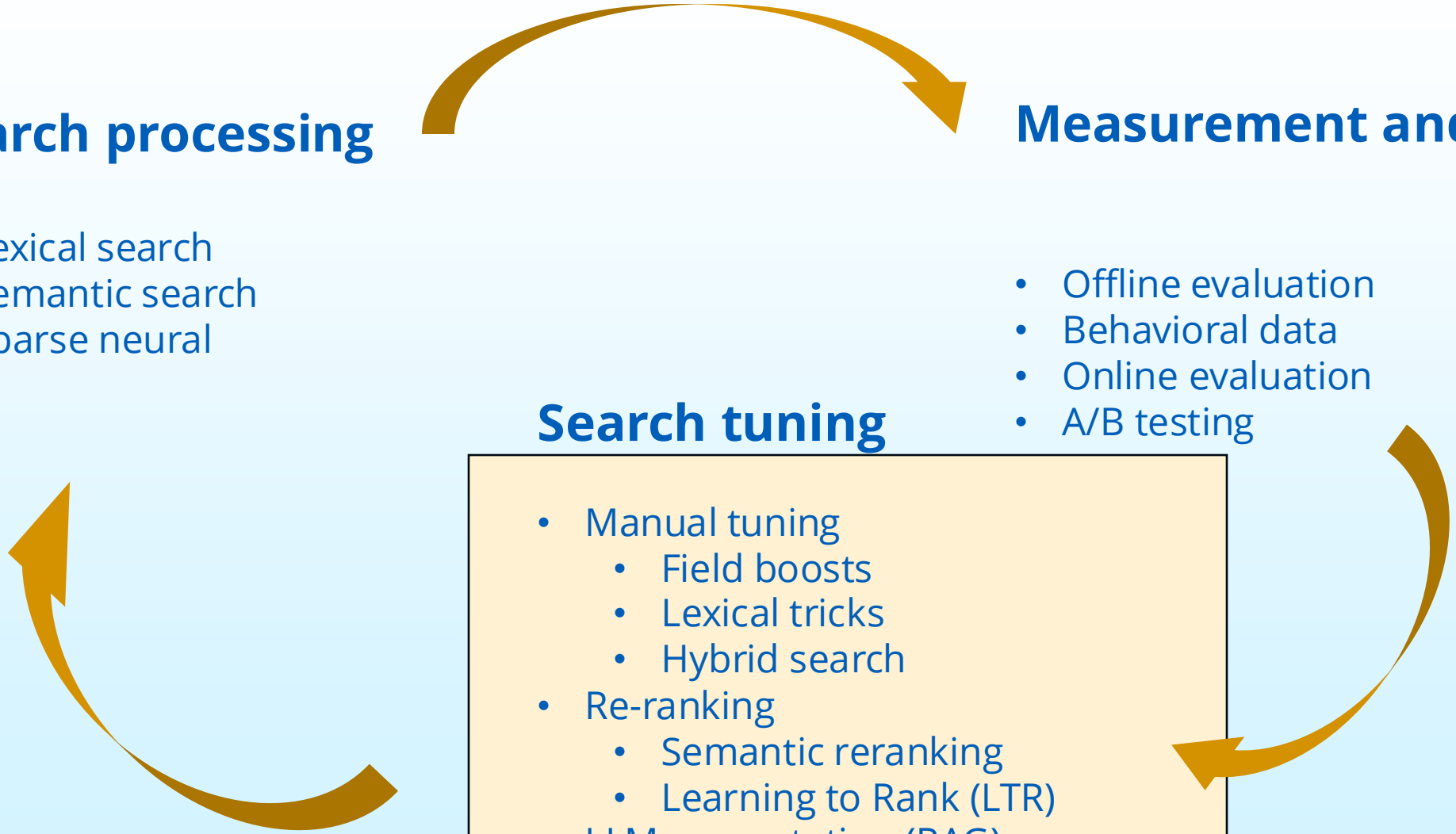
Search tuning

- Manual tuning
- Hybrid search
- Semantic reranking
- Learning to Rank (LTR)

Some Progress



NOT MUCH PROGRESS



Not much progress

- Reranking for precision
  - Learning to Rank
  - Cross-encoders

Doesn't address recall

Search processing

- Lexical search
- Semantic search
- Sparse neural

Measurement and analysis

- Offline evaluation
- Behavioral data
- Online evaluation
- A/B testing

Search tuning

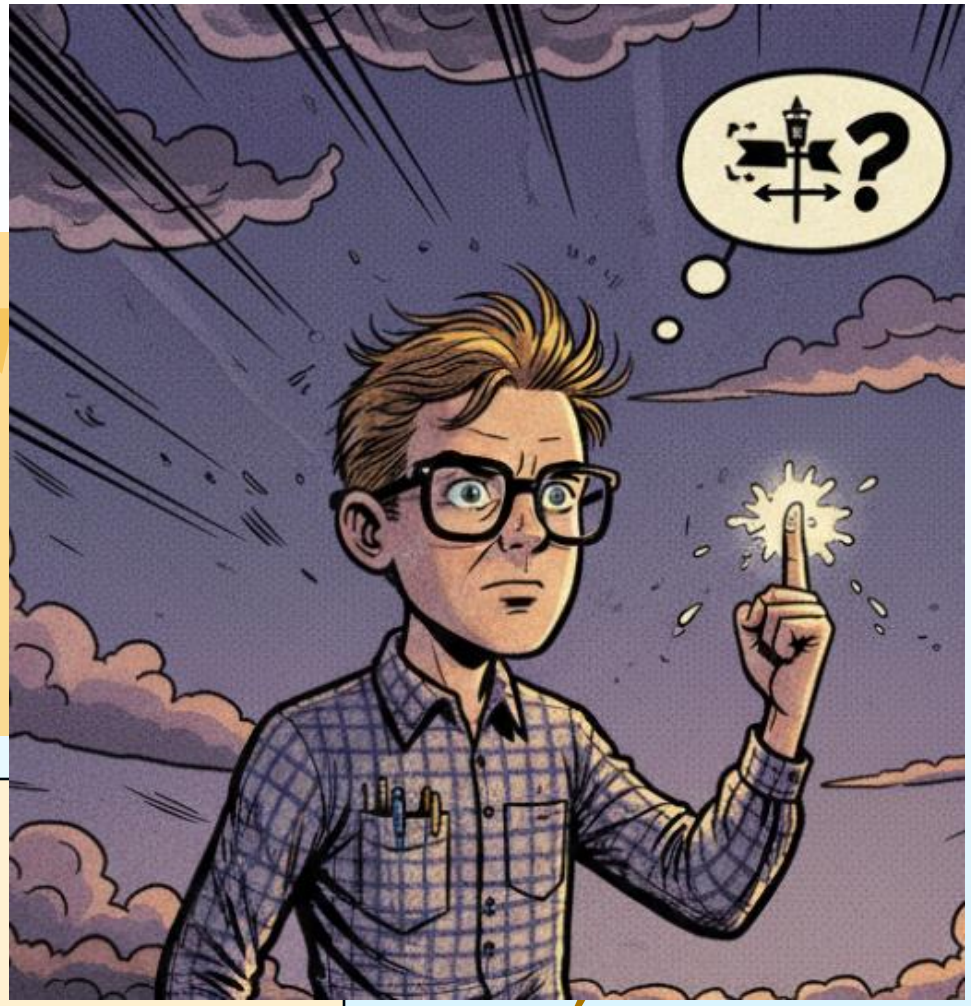
- Manual tuning
  - Field boosts
  - Lexical tricks
  - Hybrid search
- Re-ranking
  - Semantic reranking
  - Learning to Rank (LTR)
- LLM presentation (RAG)

Not much progress

Search processing

- Lexical search
- Semantic search
- Sparse neural

- Reranking for precision
  - Learning to Rank
  - Cross-encoders
- Finger in the air?



Search tuning

- Manual tuning
  - Field boosts
  - Lexical tricks
  - Hybrid search
- Re-ranking
  - Semantic reranking
  - Learning to Rank (LTR)
- LLM presentation (RAG)

The manual process

Identify a problem

Generate hypotheses

Create search configurations

Run off-line A/B tests

Run on-line A/B tests (months)

Deploy

Iterate on a scale of months

Can we do better?

THE AUTOMATIC PROCESS

Identify a problem — library of business cases

Generate a hypothesis — library of proposed fixes

Create search configurations — rewrite queries

Run off-line A/B tests — use SRW tools,

with and without judgments

Run on-line interleaved A/B test — with real-time feedback

Deploy

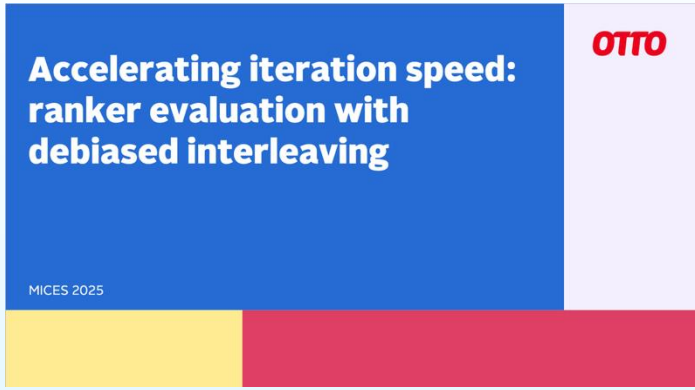
OpenSearch

Iterate on a scale of days

Relax must match
Add click data
Add phrase matching
Collect data (UBI)
Improve synonyms
Change field weights
Add ngrams
Add/remove fields
Change field mapping
Rewrite query (boost, filter ...)
Add lemmatization
Boost exact matches v. keyword
Map to structure of data
Add entity extraction for entity
(query)
Re-rank results
Semantic
Chunking strategy
...tune model; finetune
...ting (products v. accessories,
date/freshness)
Add taxonomy; engage external
taxonomist
Retrieval (hybrid; sparse)
Faceted search
Add multimedia data
Personalize results
Stock status, shipping, tax (post
filtering)
Use better images
Add entities to corpus
Add query terms to a field on
clicked X
Use machine learning to figure out
the value of terms
Spell checking
Compounding/decompounding
Security trimming
Add geo
Finetune model; hyperparameters
Map to structure of data
Change field mapping
Add lemmatization
Add entities to corpus
Re-rank results
Use machine learning to
the value of terms
Semantic
Faceted search
Remove query clauses
Add entity extraction for
(query)
Chunking strategy
Retrieval (hybrid; sparse
Add geo
Add click data
Rewrite query (boost, fi

OpenSearch

A QUICK WORD ON INTERLEAVED A/B TESTING



Classic A/B test:

- partition *users* into A group and B group
- Problem: each user issues different queries
- Very slow statistical significance

Interleaved A/B test:

- Combine A and B results for *each user*
- Faster statistical significance

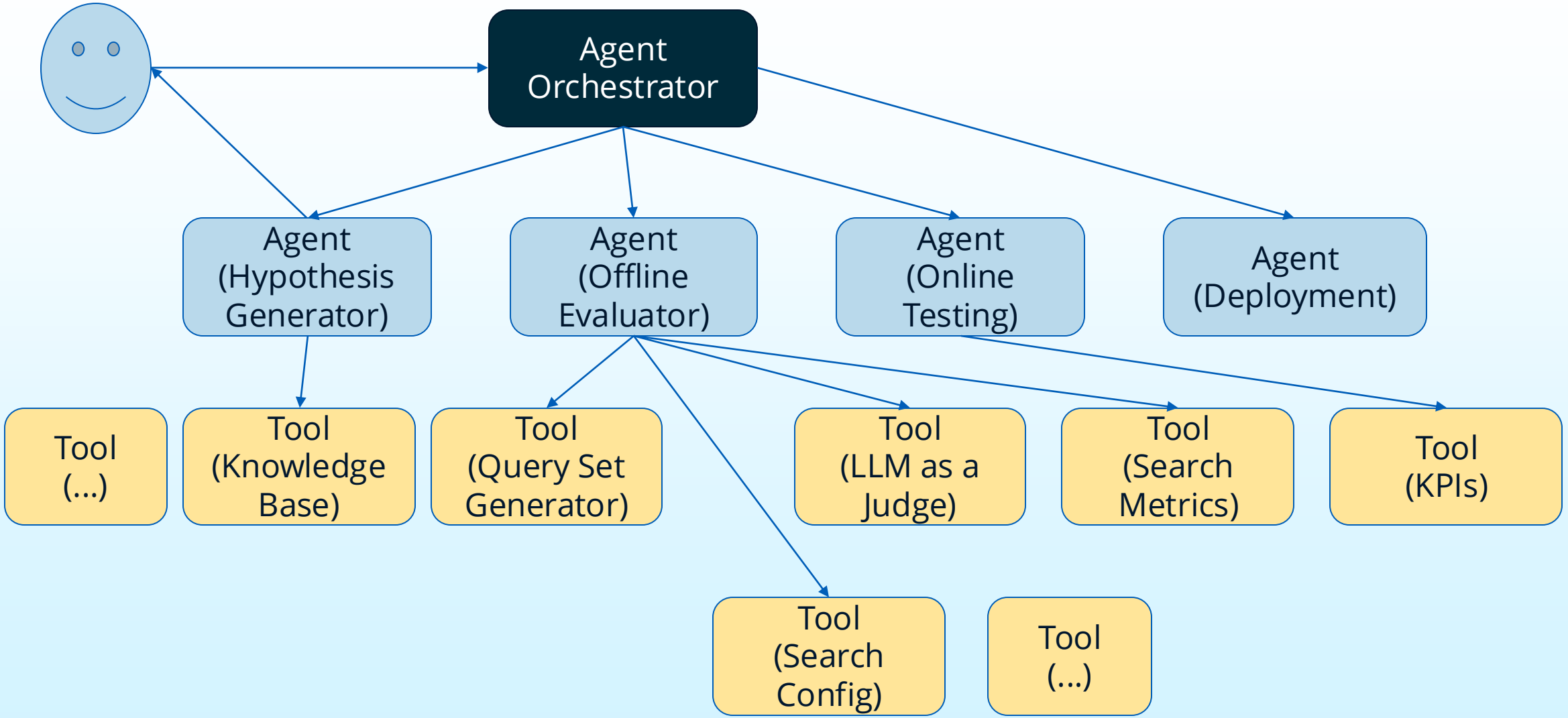
Much much faster feedback

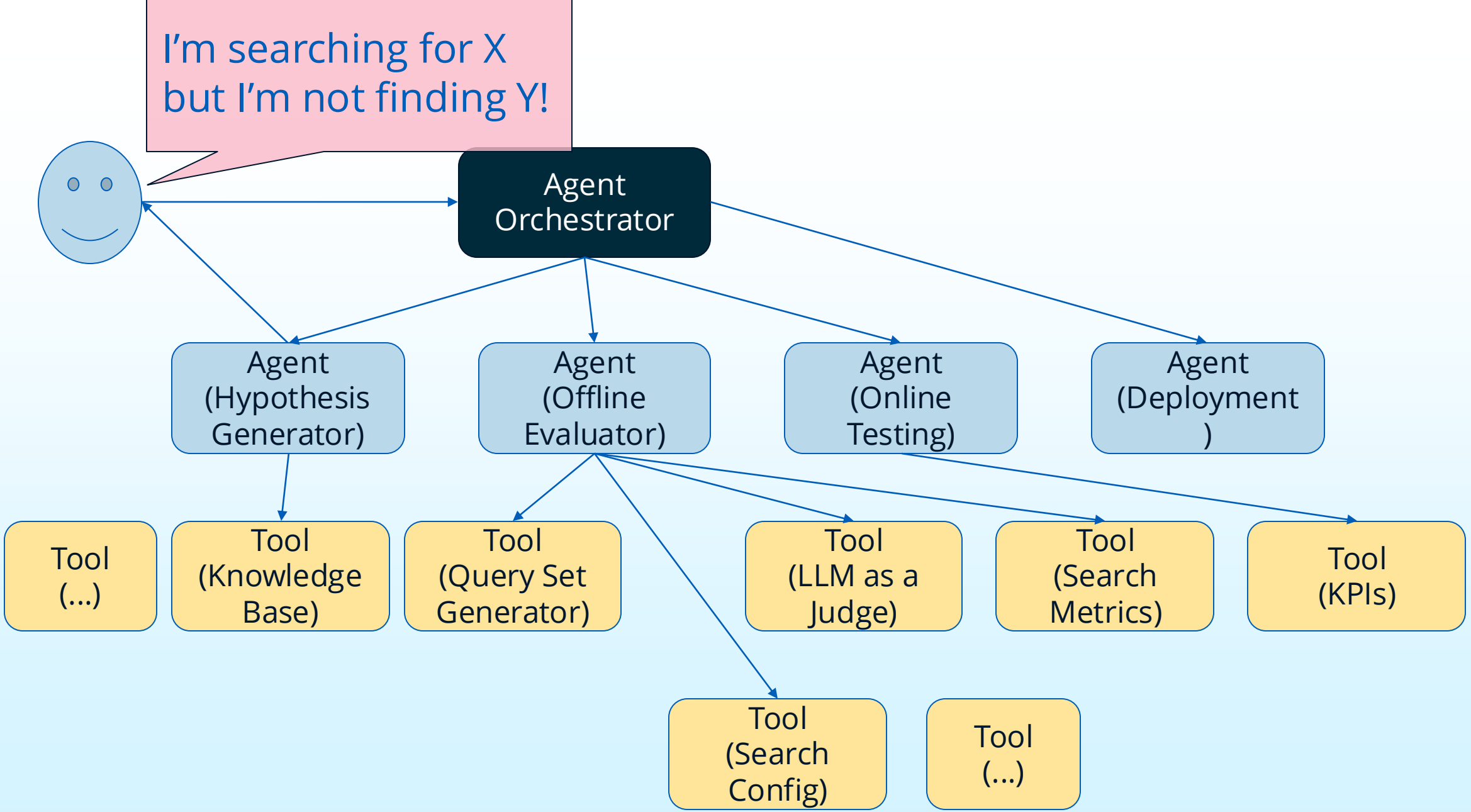
OpenSearch

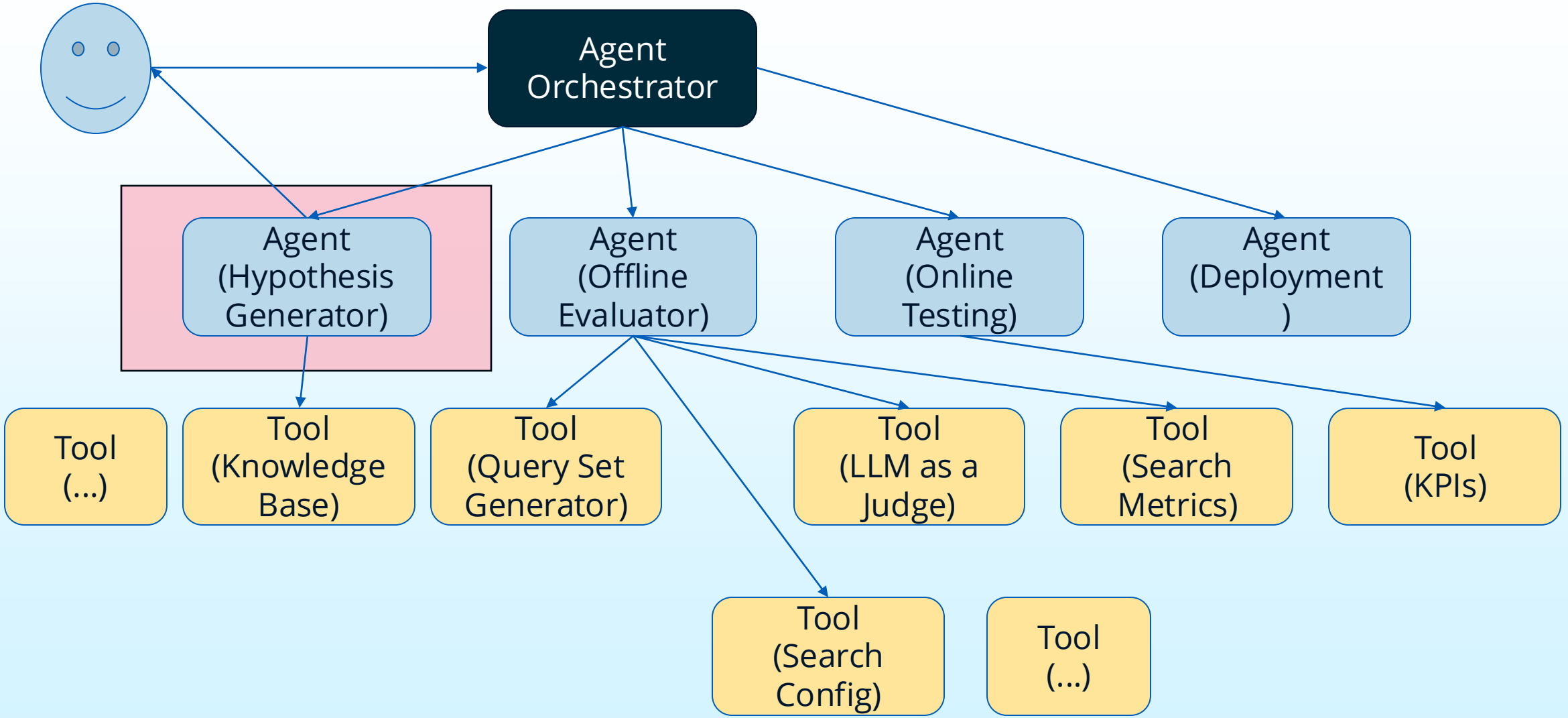
Using agents

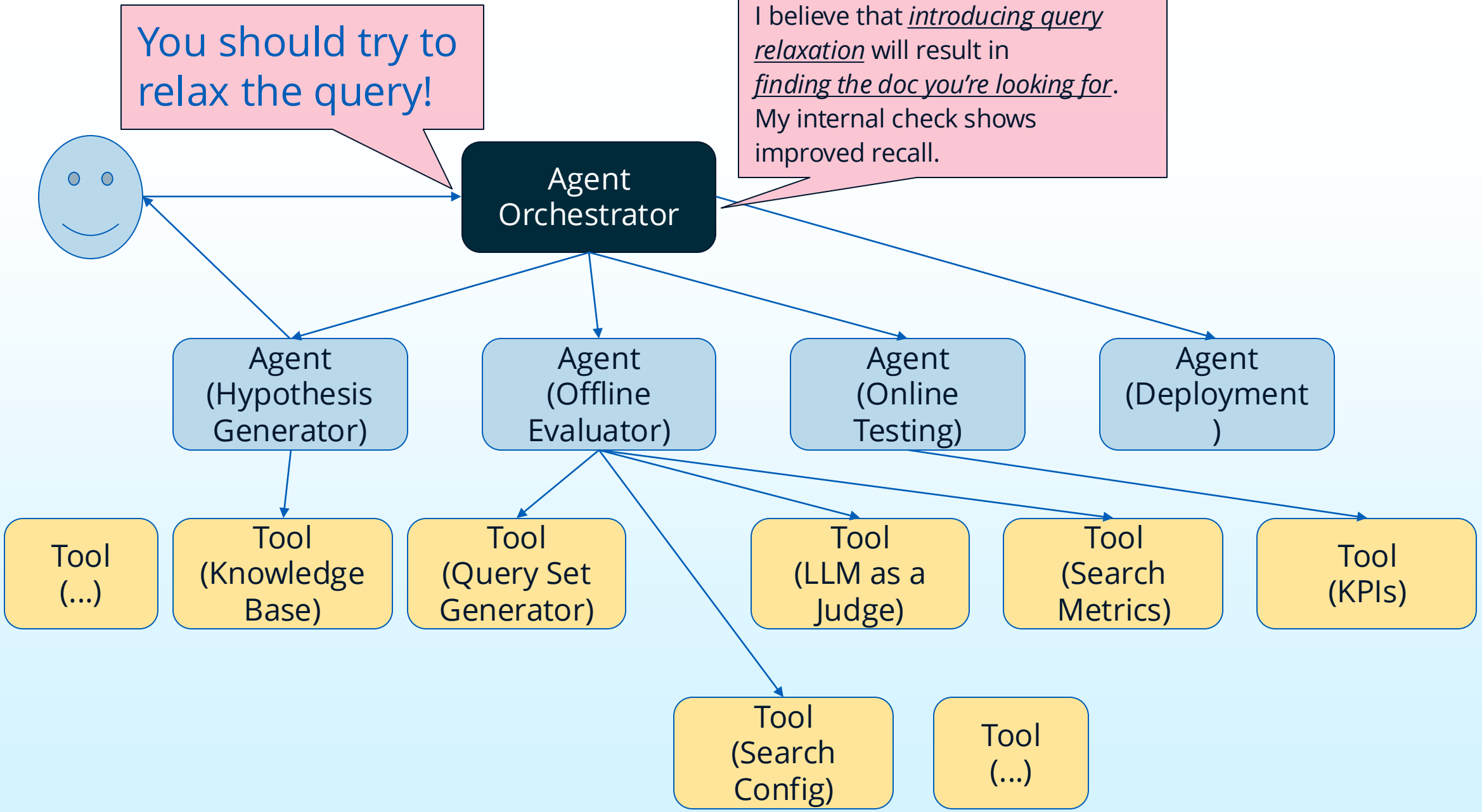
Orchestration and user interaction
Business Problem
Search Strategy
Offline Experiment
Online Experiment runner

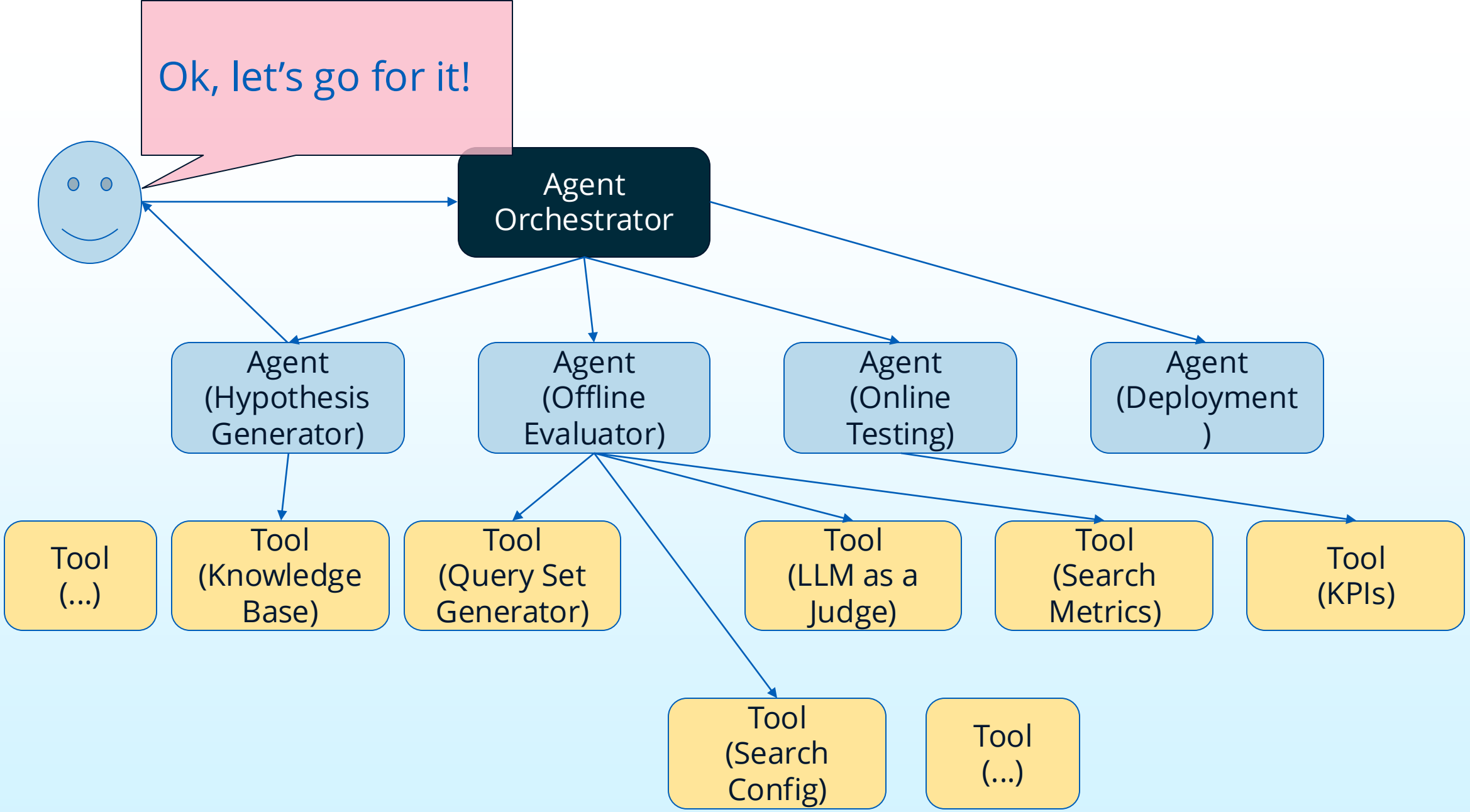
Agents, tools, and modules

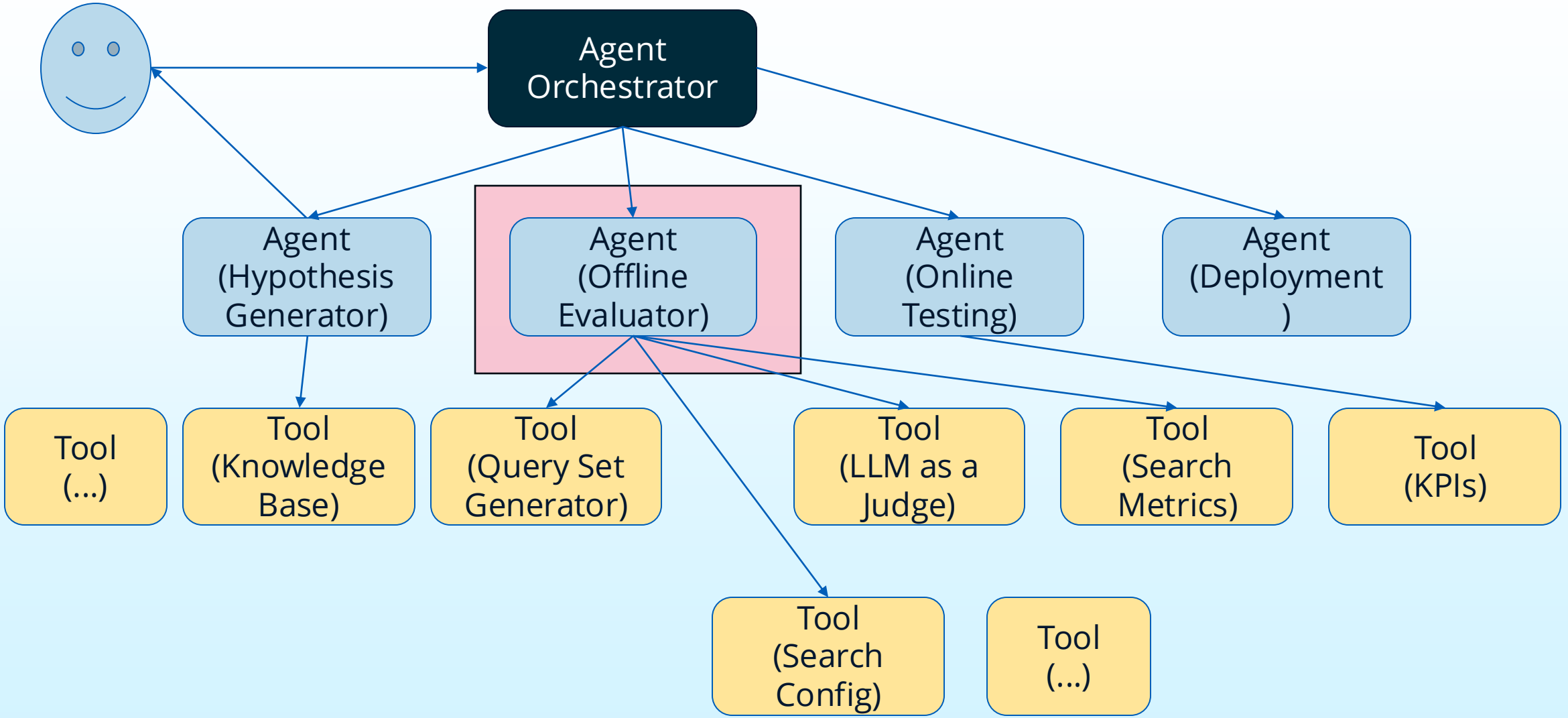

Agent Orchestrator
Agent (Hypothesis Generator)
Agent (Offline Evaluator)
Agent (Online Testing)
Agent (Deployment)
Tool (...)
Tool (Knowledge Base)
Tool (Query Set Generator)
Tool (LLM as a Judge)
Tool (Search Metrics)
Tool (KPIs)
Tool (Search Config)
Tool (...)

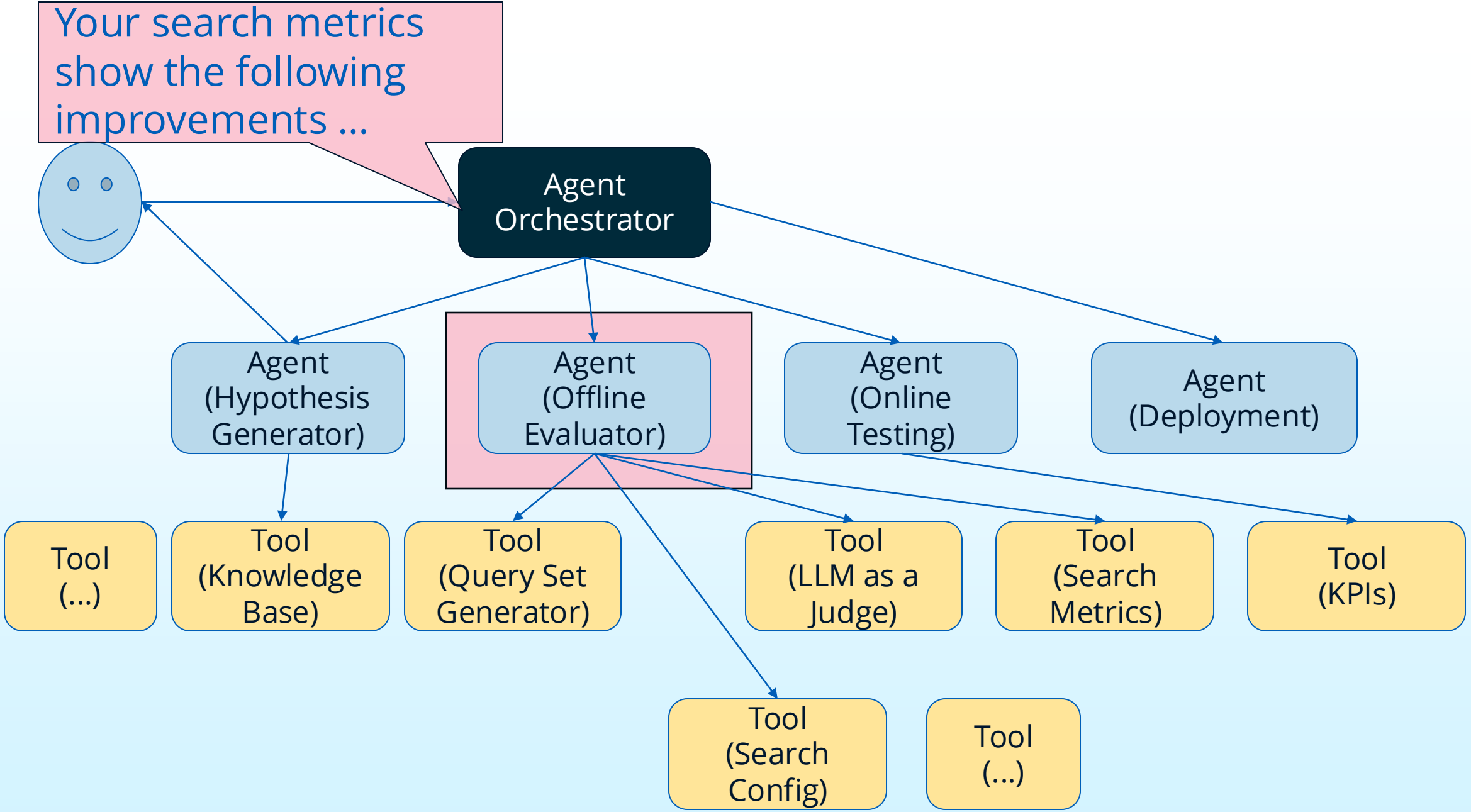

I’m searching for X
but I’m not finding Y!
Agent
Orchestrator
Agent
(Hypothesis
Generator)
Agent
(Offline
Evaluator)
Agent
(Online
Testing)
Agent
(Deployment
)
Tool
(...)
Tool
(Knowledge
Base)
Tool
(Query Set
Generator)
Tool
(LLM as a
Judge)
Tool
(Search
Metrics)
Tool
(KPIs)
Tool
(Search
Config)
Tool
(...)

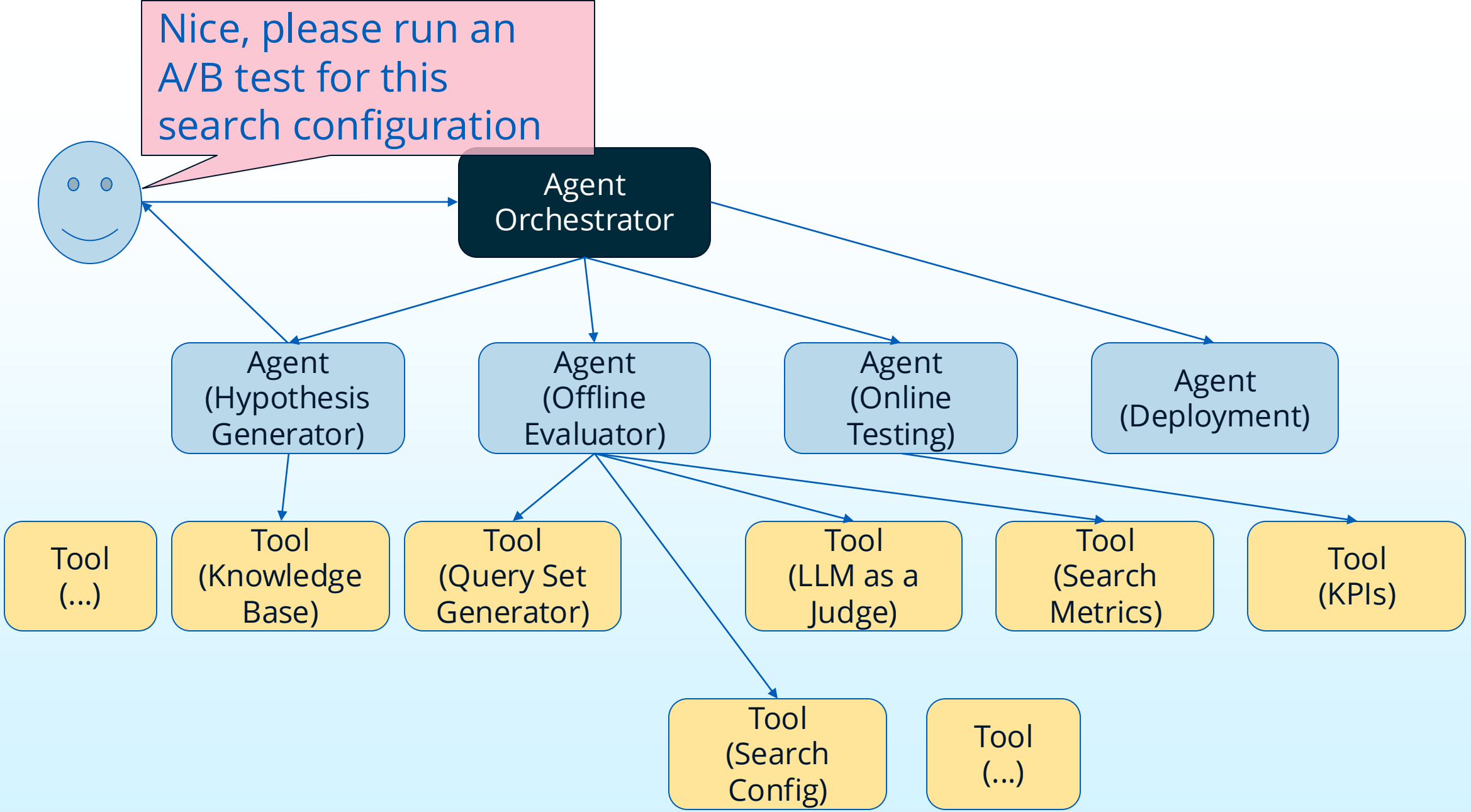

Agent Orchestrator
Agent (Hypothesis Generator)
Agent (Offline Evaluator)
Agent (Online Testing)
Agent (Deployment)
Tool (...)
Tool (Knowledge Base)
Tool (Query Set Generator)
Tool (LLM as a Judge)
Tool (Search Metrics)
Tool (KPIs)
Tool (Search Config)
Tool (...)

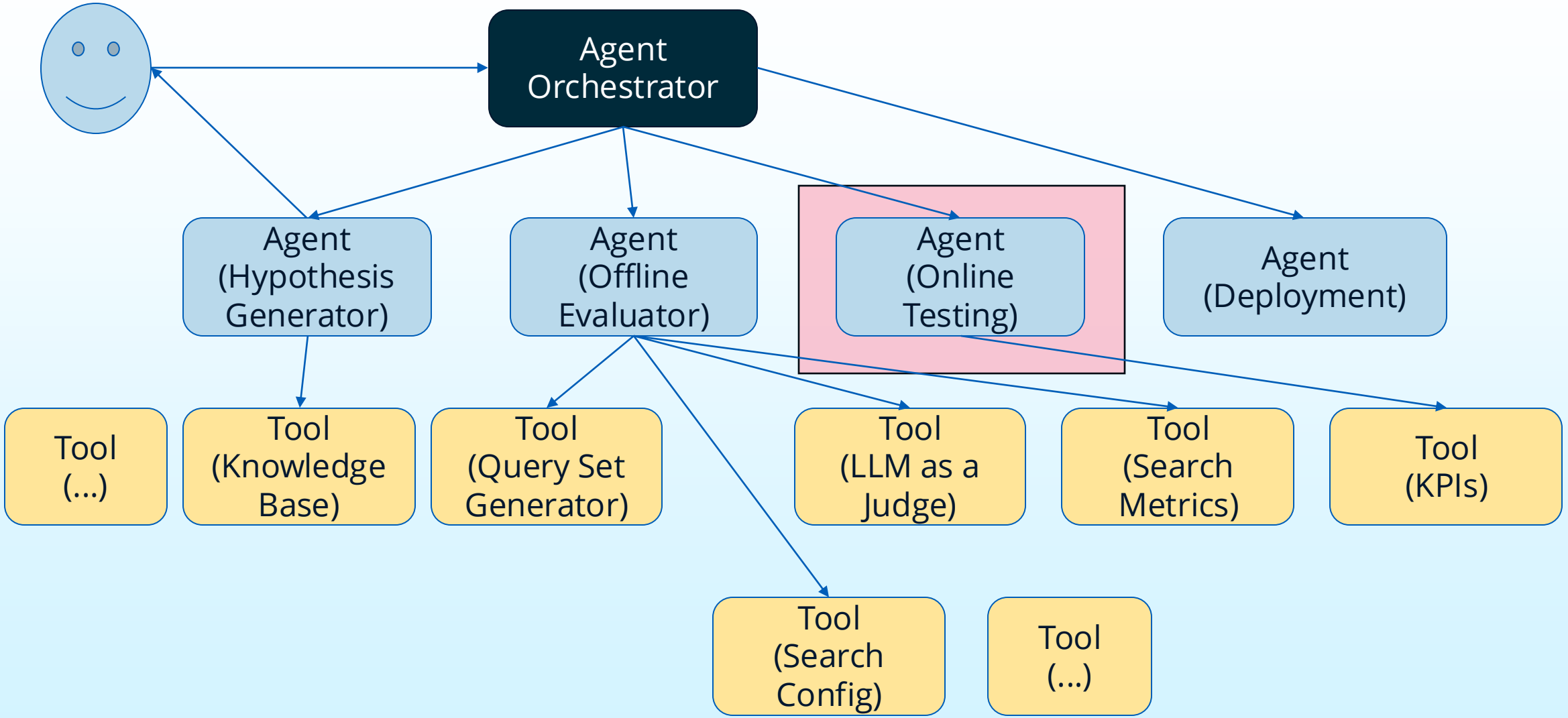

You should try to relax the query!
I believe that introducing query relaxation will result in finding the doc you're looking for. My internal check shows improved recall.
Agent Orchestrator
Agent (Hypothesis Generator)
Agent (Offline Evaluator)
Agent (Online Testing)
Agent (Deployment)
Tool (...)
Tool (Knowledge Base)
Tool (Query Set Generator)
Tool (LLM as a Judge)
Tool (Search Metrics)
Tool (KPIs)
Tool (Search Config)
Tool (...)

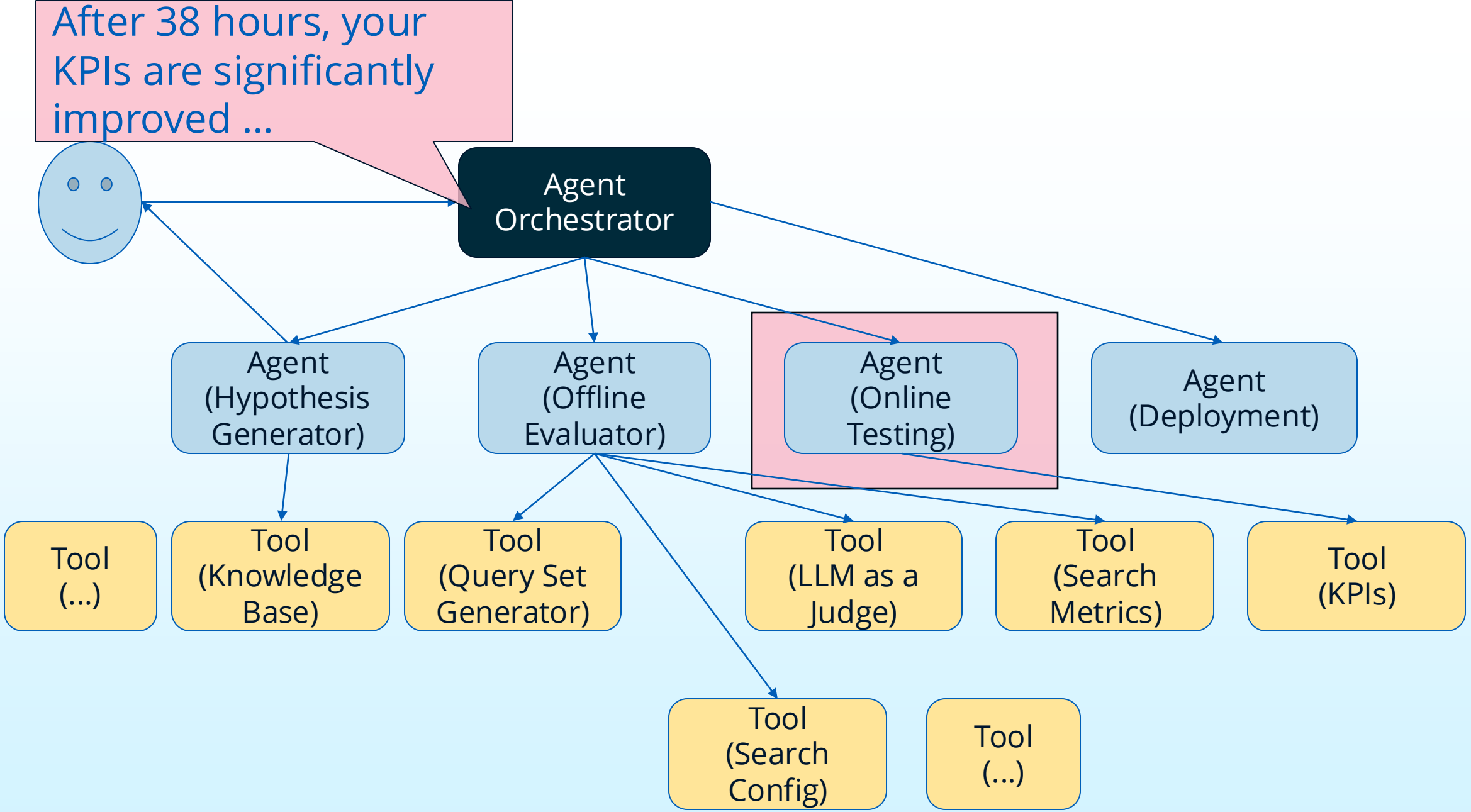

Ok, let’s go for it!
Agent Orchestrator
Agent (Hypothesis Generator)
Agent (Offline Evaluator)
Agent (Online Testing)
Agent (Deployment)
Tool (...)
Tool (Knowledge Base)
Tool (Query Set Generator)
Tool (LLM as a Judge)
Tool (Search Metrics)
Tool (KPIs)
Tool (Search Config)
Tool (...)

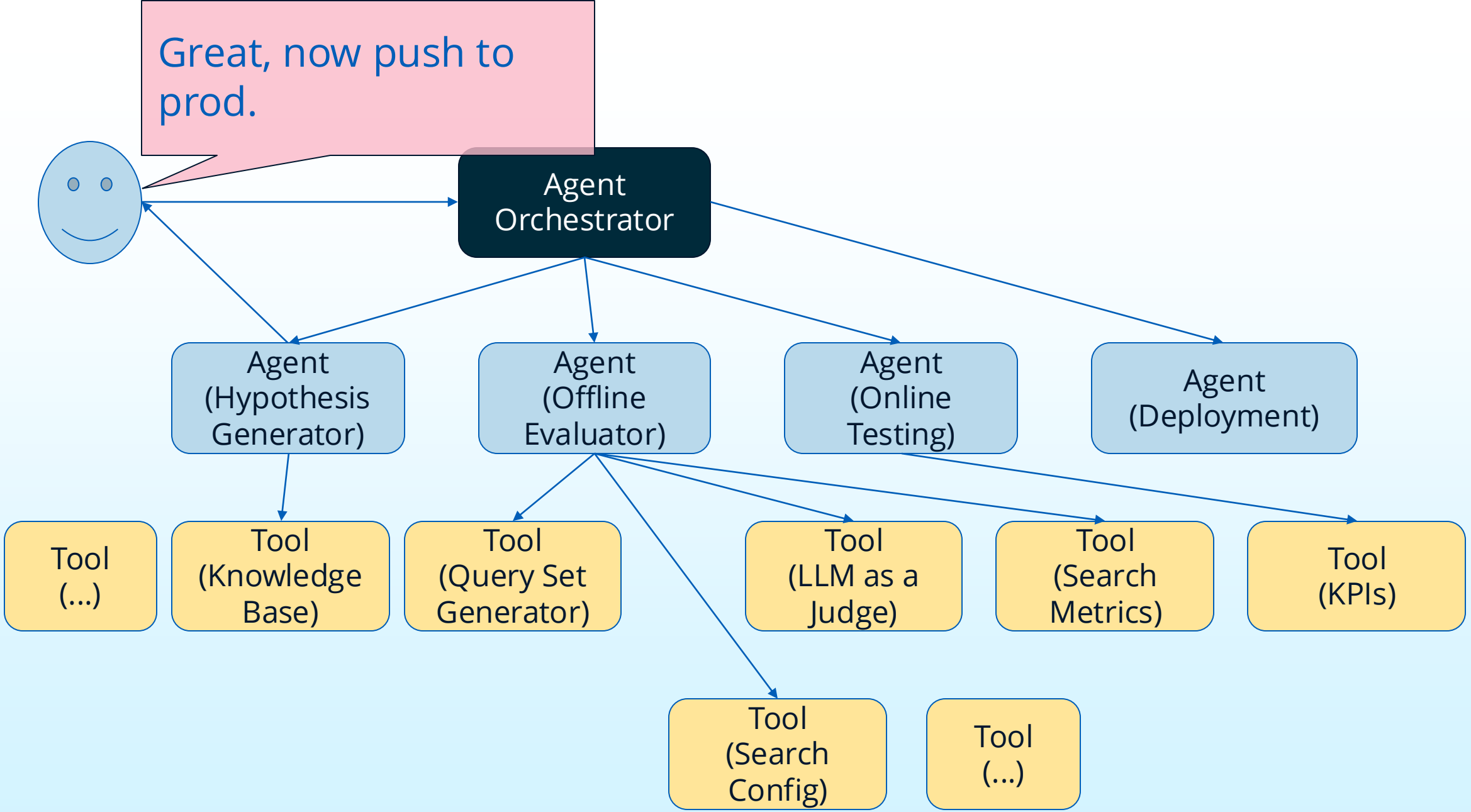

Agent Orchestrator
Agent (Hypothesis Generator)
Agent (Offline Evaluator)
Agent (Online Testing)
Agent (Deployment)
Tool (...)
Tool (Knowledge Base)
Tool (Query Set Generator)
Tool (LLM as a Judge)
Tool (Search Metrics)
Tool (KPIs)
Tool (Search Config)
Tool (...)

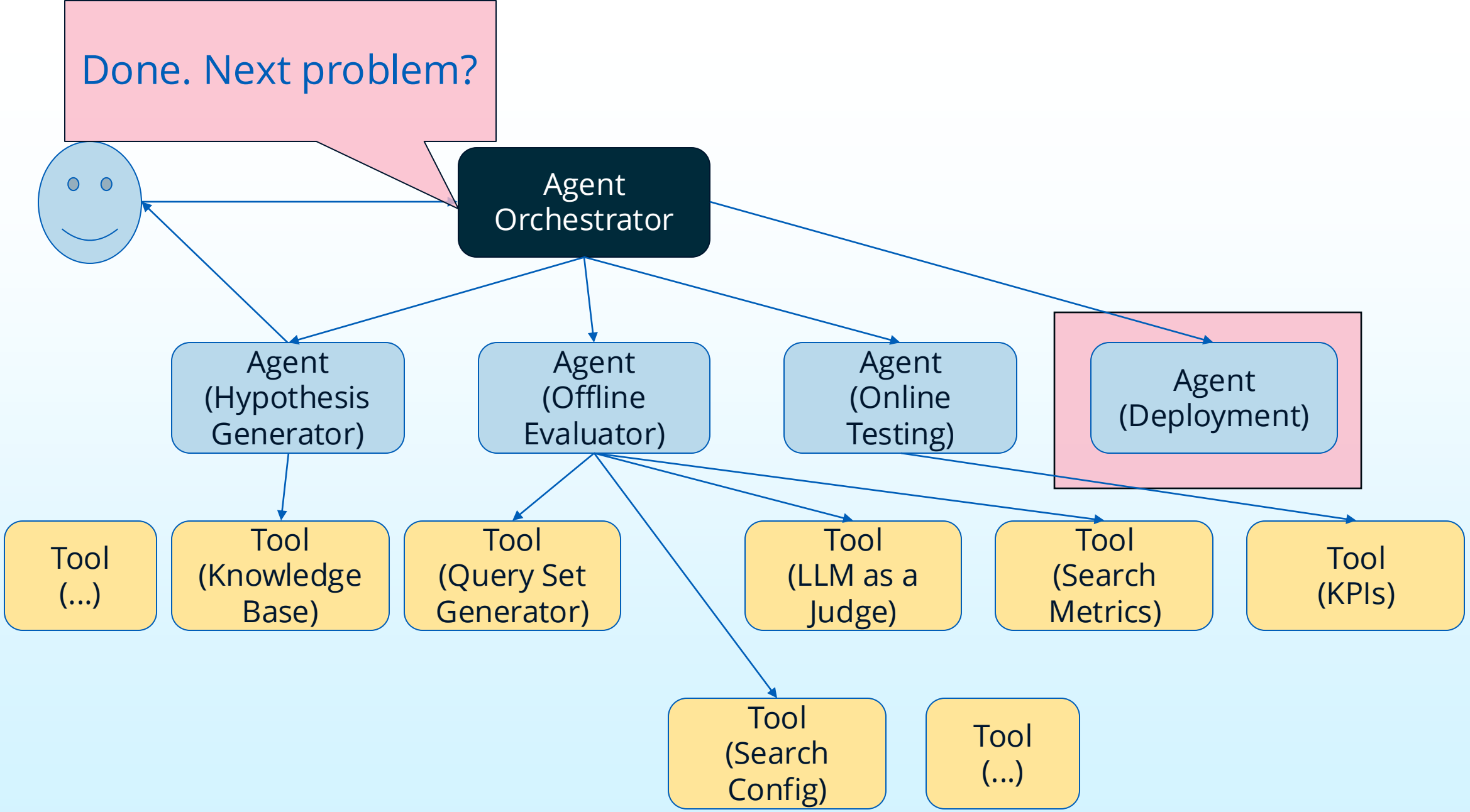

Your search metrics show the following improvements ...
Agent Orchestrator
Agent (Hypothesis Generator)
Agent (Offline Evaluator)
Agent (Online Testing)
Agent (Deployment)
Tool (...)
Tool (Knowledge Base)
Tool (Query Set Generator)
Tool (LLM as a Judge)
Tool (Search Metrics)
Tool (KPIs)
Tool (Search Config)
Tool (...)


Nice, please run an A/B test for this search configuration
Agent Orchestrator
Agent (Hypothesis Generator)
Agent (Offline Evaluator)
Agent (Online Testing)
Agent (Deployment)
Tool (...)
Tool (Knowledge Base)
Tool (Query Set Generator)
Tool (LLM as a Judge)
Tool (Search Metrics)
Tool (KPIs)
Tool (Search Config)
Tool (...)


Agent Orchestrator
Agent (Hypothesis Generator)
Agent (Offline Evaluator)
Agent (Online Testing)
Agent (Deployment)
Tool (...)
Tool (Knowledge Base)
Tool (Query Set Generator)
Tool (LLM as a Judge)
Tool (Search Metrics)
Tool (KPIs)
Tool (Search Config)
Tool (...)


After 38 hours, your KPIs are significantly improved ...
Agent Orchestrator
Agent (Hypothesis Generator)
Agent (Offline Evaluator)
Agent (Online Testing)
Agent (Deployment)
Tool (...)
Tool (Knowledge Base)
Tool (Query Set Generator)
Tool (LLM as a Judge)
Tool (Search Metrics)
Tool (KPIs)
Tool (Search Config)
Tool (...)


Great, now push to prod.
Agent Orchestrator
Agent (Hypothesis Generator)
Agent (Offline Evaluator)
Agent (Online Testing)
Agent (Deployment)
Tool (...)
Tool (Knowledge Base)
Tool (Query Set Generator)
Tool (LLM as a Judge)
Tool (Search Metrics)
Tool (KPIs)
Tool (Search Config)
Tool (...)


Done. Next problem?
Agent Orchestrator
Agent (Hypothesis Generator)
Agent (Offline Evaluator)
Agent (Online Testing)
Agent (Deployment)
Tool (...)
Tool (Knowledge Base)
Tool (Query Set Generator)
Tool (LLM as a Judge)
Tool (Search Metrics)
Tool (KPIs)
Tool (Search Config)
Tool (...)

Agentic Relevance Tuning

Agents to automate tuning

+

Interleaving for fast feedback

=

Fast, goal-driven relevance improvements

OpenSearch