Institute of Computing Technology
Chinese Academy of Sciences

Beyond Relevance:

Utility-Centric Retrieval in the LLM Era

Hengran Zhang · Minghao Tang · **Keping Bi** · Jiafeng Guo

Institute of Computing Technology, CAS · University of Chinese Academy of Sciences

SIGIR 2026 Tutorial · July 20, 2026 · 01:30 – 05:00 PM

https://stay-hungry-time.github.io/utility-tutorial/

Tutorial website

About the Presenters

Hengran Zhang

PhD student
ICT, CAS

Minghao Tang

Master student
ICT, CAS

Keping Bi

Faculty
ICT, CAS

Jiafeng Guo

Faculty
ICT, CAS

Tutorial Roadmap

180 min total

01 **Introduction & Foundations -Keping** — 40 MIN
Why retrieval objectives must change for LLM consumers

02 **What Is LLM-Centric Utility? -Keping** — 50 MIN
Task, model, context, and presentation dimensions

03 **Utility Modeling & Optimization -Hengran** — 40 MIN
Signals, labels, trainable components, and evaluation

04 **Utility in Agentic RAG -Keping** — 40 MIN
Utility in dynamic retrieval and agent loops

05 **Open Problems & Q&A -Keping** — 10 MIN
Open problems and discussion

Tutorial website

SECTION 1

Introduction and Foundations

1

Why retrieval objectives must change in the LLM era.

Foundations · LLM-Centric Utility · Modeling & Optimization · Agentic RAG · Open Problems

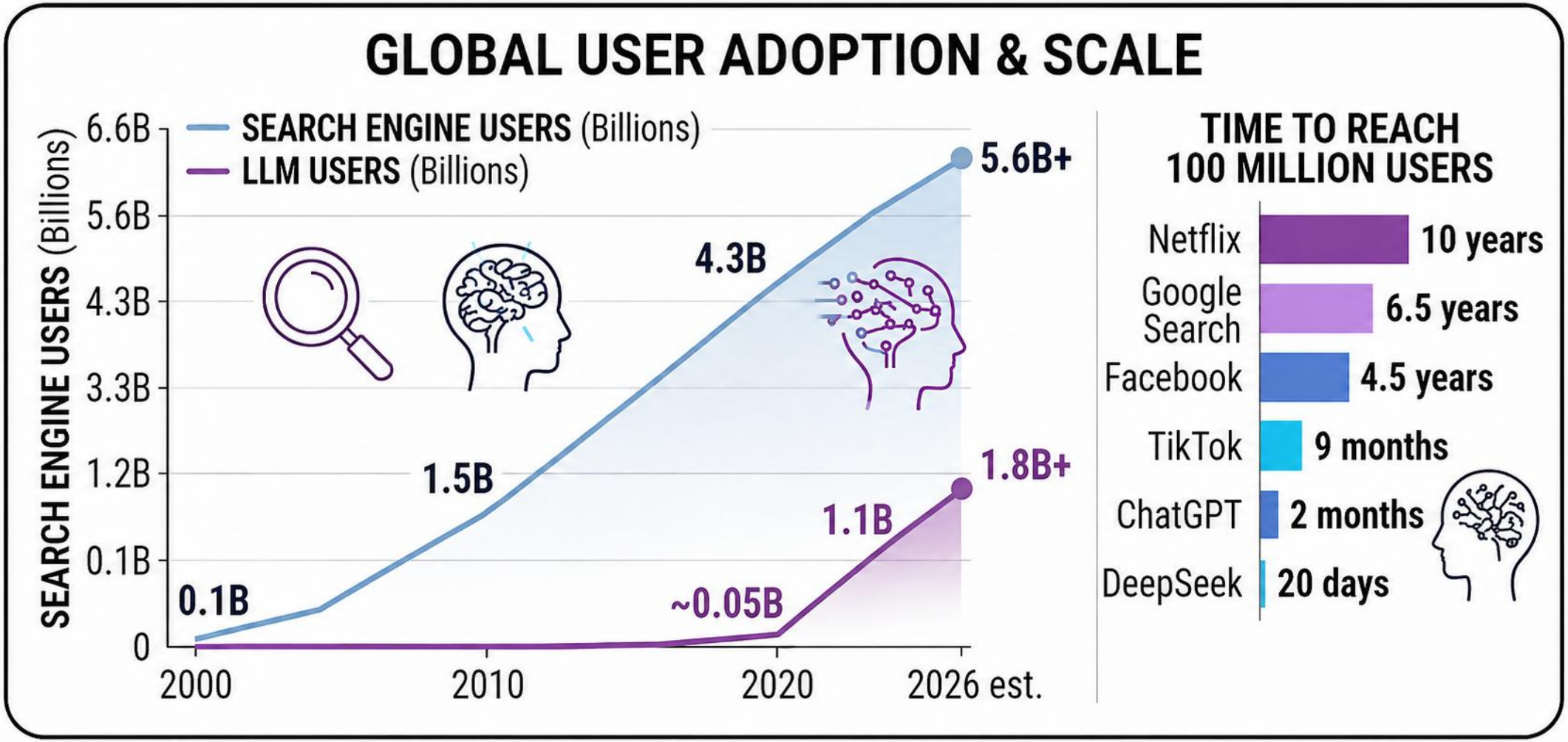
The Evolution of Information Access

1960s
Online Database Retrieval Era

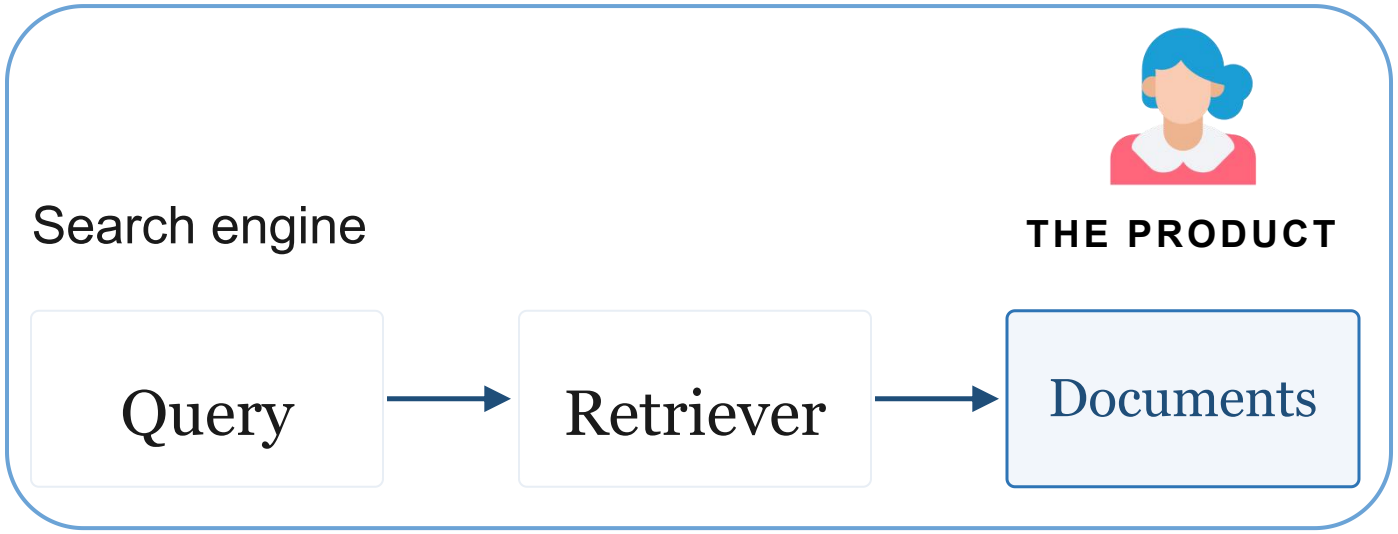
1990s
Web Search Era

2020s
Generative AI Era

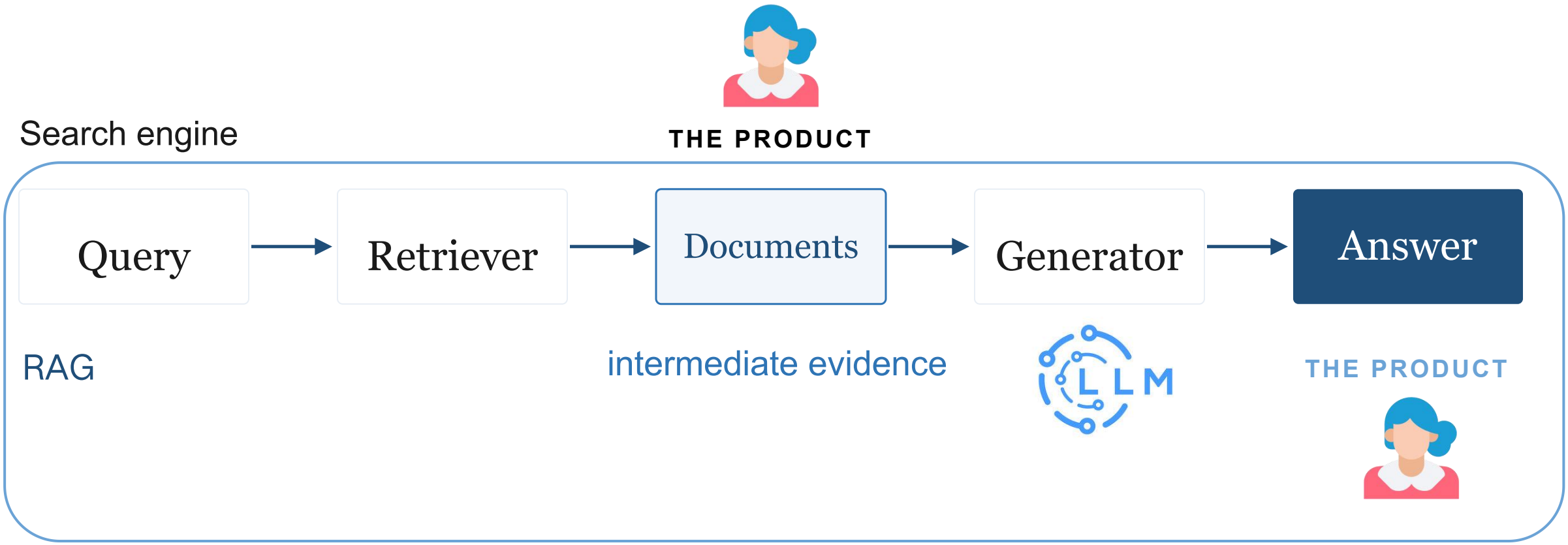
The Evolution of Information Access



RAG Changes the Product

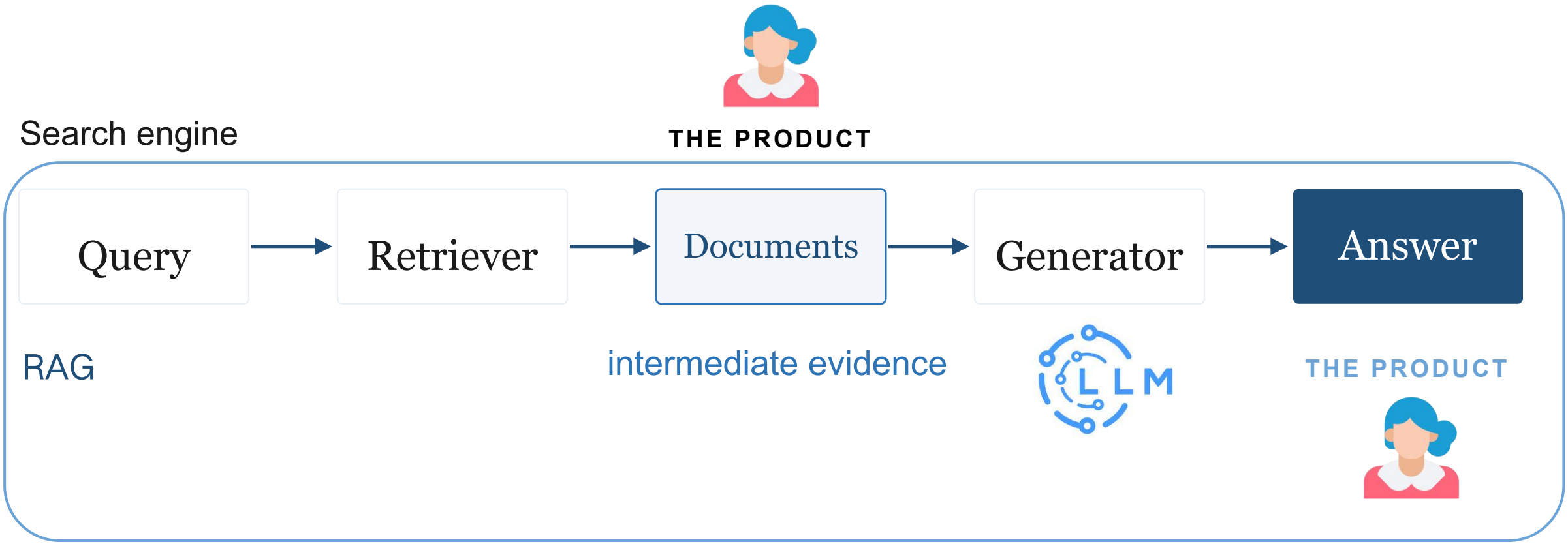


RAG Changes the Product



Lewis et al. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. NeurIPS, 2020.

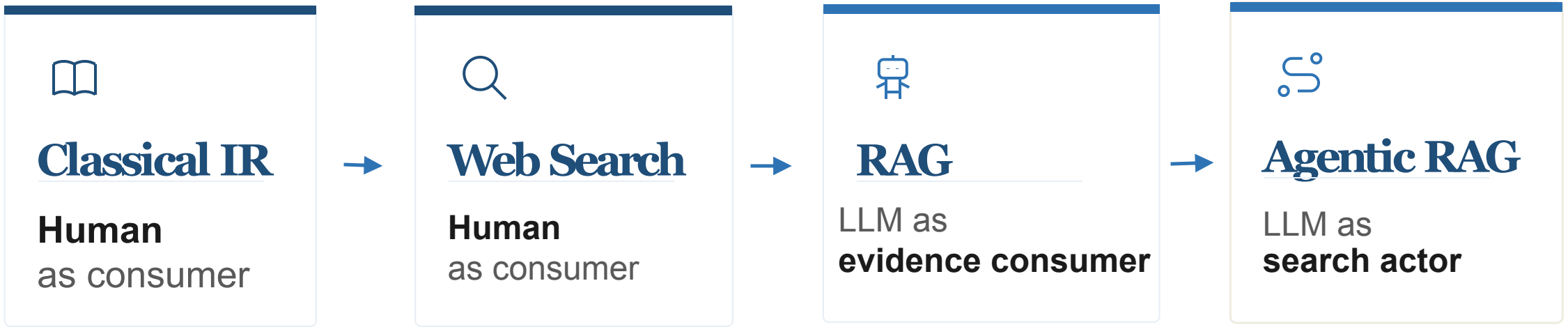
RAG Changes the Product



The consumer of retrieval results shifts from users to LLMs.

Lewis et al. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. NeurIPS, 2020.

From Relevance to Utility: the Consumer Evolves



From Relevance to Utility: the Consumer Evolves

Classical IR

Human
as consumer

→

Web Search

Human
as consumer

→

RAG

LLM as
evidence consumer

→

Agentic RAG

LLM as
search actor

Human searchers can skip, compare, and judge documents.
As the **LLM becomes the consumer**, retrieval must be judged by its contribution to what the model **generates.**

The Evolution of Retrieval Standards

RANKING STANDARD

Topical	Contextual	Fine-grained	Task utility	LLM utility
relevance defined	user & task context	Learning to Rank	scenario-oriented	serves generation

relevance → utility

1960s · 1970s · 2000s · 2010s · 2020s

DATABASE SEARCH SYSTEMS	WEB SEARCH ENGINES	LLM ASSISTANTS
SMART · Dialog · MEDLINE · AIM-TWX · MEDLARS	Google, Baidu 百度, yahoo!, Microsoft Bing	deepseek

From Relevance to Utility

Classical IR: Topical Relevance

- Topical match
- Aboutness
- Query–document relation
- Estimated usefulness — a proxy

Easy to annotate, reusable across collections, scalable to evaluate.

Multi-Dimensional Relevance

User

Task

Situation

Relevance depends on **who is asking, and why**.

Cooper. A Definition of Relevance for Information Retrieval. Information Storage and Retrieval, 1971.

Saracevic. Relevance: A Review of the Notion in Information Science. JASIS, 1975.

Relevance Is Multi-Dimensional

Topical	Is the document about the requested subject?
Cognitive	Does it add information the user does not already know and can understand?
Situational	Is it useful for the user's current task or decision?
Affective	Does it suit the user's interests, preferences, or goals?

…

Saracevic. Relevance: A Review of the Notion in Information Science. JASIS, 1975.

User-Centric Utility

RELEVANCE
aboutness

vs

UTILITY
value

Effectiveness Measures

Two sets of measures for evaluating the effectiveness of a search have been used, based on the two most often used criteria:

1. *Relevance:* the degree of fit between the question and the retrieved item. The criteria of *"aboutness"* is used.
2. *Utility:* the degree of actual usefulness of answers to an information seeker. The criteria used is the *value* to the information seeker.

- Distinguished from **topical relevance**
- Useful for a **task**
- Dependent on **context**
- Valuable relative to cost

Implicit. Hard to measure.

W. S. Cooper, "On Selecting a Measure of Retrieval Effectiveness," JASIS, 1973.
Saracevic et al. A Study of Information Seeking and Retrieving. JASIS, 1988.

Binary to Multi-Grade Relevance

Grade	Label	TREC Web Track short reading
0	Non	No useful information for the intended topic
1	Rel	Some on-page information
2	HRel	Substantial information
3	Key	Authoritative, comprehensive, top-result worthy
4	Nav	Direct navigational target for an entity query

Supports gain-based ranking metrics such as DCG and nDCG.

Collins-Thompson et al. TREC 2014 Web Track Overview. TREC, 2014.

Multi-Grade Relevance As A Bridge

Binary relevance

Relevant vs. non-relevant.
A single yes/no label.

→

Multi-grade relevance

Non · Rel · HRel · Key · Nav.
Degrees and types of gain.

→

Utility

How much the evidence helps the information seeker.

Multi-grade relevance tells us how good a document looks for a query.
Utility tells us what it contributes to the user, context, and task.

Multi-Grade Relevance ≠ Utility

Aspect	Multi-grade relevance	Utility
Basic object	Query–document pair	User / model / task outcome
Judgment timing	Usually before use	During or after use
Context reliance	Often limited	Strong: context, state, set
Redundancy	Usually not captured	Central: repeats add little
Negative effect	Represented as low relevance	Can be explicitly harmful
Consumer	Human assessor	Human, LLM, agent

Judged by a third person, following an objective standard and context-independent.

Why Relevance Won Historically

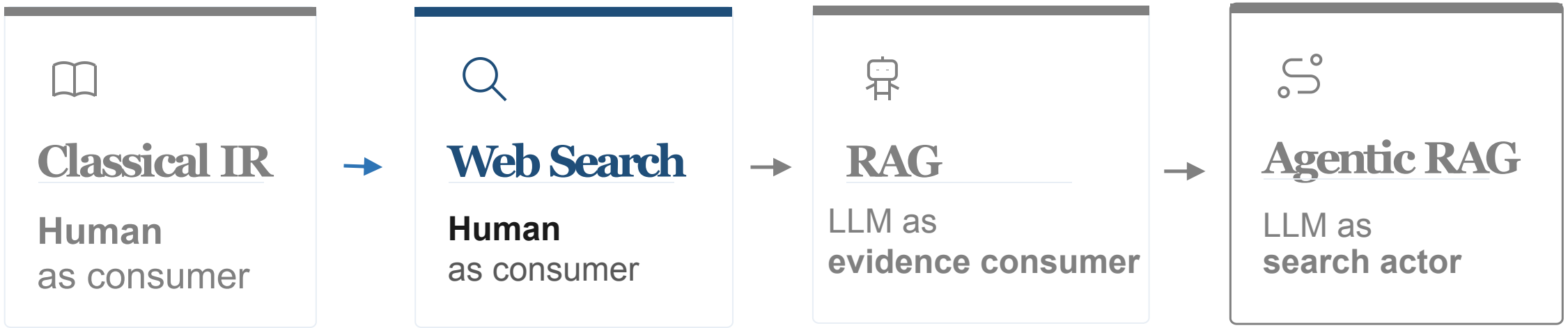
RELEVANCE

- Easier judgments
- Reusable benchmarks
- Scalable evaluation

UTILITY

- Task-dependent outcomes
- Costly to measure
- No reusable ground truth

Relevance won as a **practical proxy**.


Classical IR
Human
as consumer
Web Search
Human
as consumer
RAG
LLM as
evidence consumer
Agentic RAG
LLM as
search actor

User-Centric Utility via Behavior

WHAT WE OBSERVE

- Clicks
- Dwell time
- Reformulation
- Session length

...

approximate →

THE UTILITY THEY STAND IN FOR

- Attraction
- Satisfaction
- Task progress
- Abandonment

Behavior is a proxy for utility, gathered at scale.

Jung et al. Click Data as Implicit Relevance Feedback in Web Search. Information Processing & Management, 2007.
Kelly & Belkin. Display Time as Implicit Feedback: Understanding Task Effects. SIGIR, 2004.
Buscher et al., Segment-Level Display Time as Implicit Feedback. SIGIR, 2009;
Luo et al. Does Document Relevance Affect the Searcher's Perception of Time?," WSDM, 2017.

Behavioral Signals Are Biased

- ⚠ Position bias
- ⚠ Snippet-attractiveness bias
- ⚠ Long dwell caused by confusion
- ⚠ Satisfaction without a click

Each one can **bias** the utility signal we read from behavior.

Ai et al. Unbiased Learning to Rank with Unbiased Propensity Estimation. SIGIR, 2018.
Ye et al. Why Don't You Click: Understanding Non-Click Results in Web Search with Brain Signals. SIGIR, 2022.
Liu et al. "Satisfaction with Failure" or "Unsatisfied Success": Investigating the Relationship between Search Success and User Satisfaction. WWW, 2018.

Beyond Individual Document Utility

Whole-Page Utility

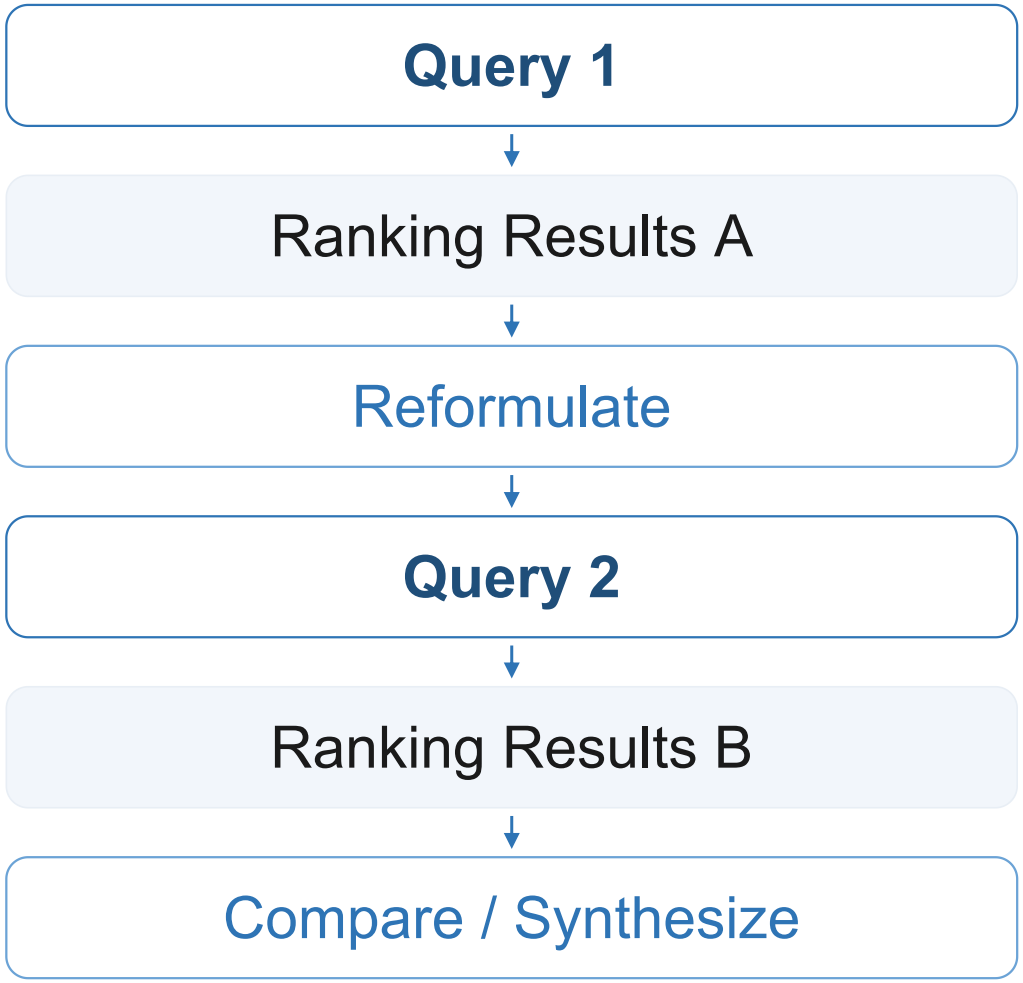
Search results are consumed as a **page**

Wang et al. Optimizing Whole-Page Presentation for Web Search. WSDM, 2016.

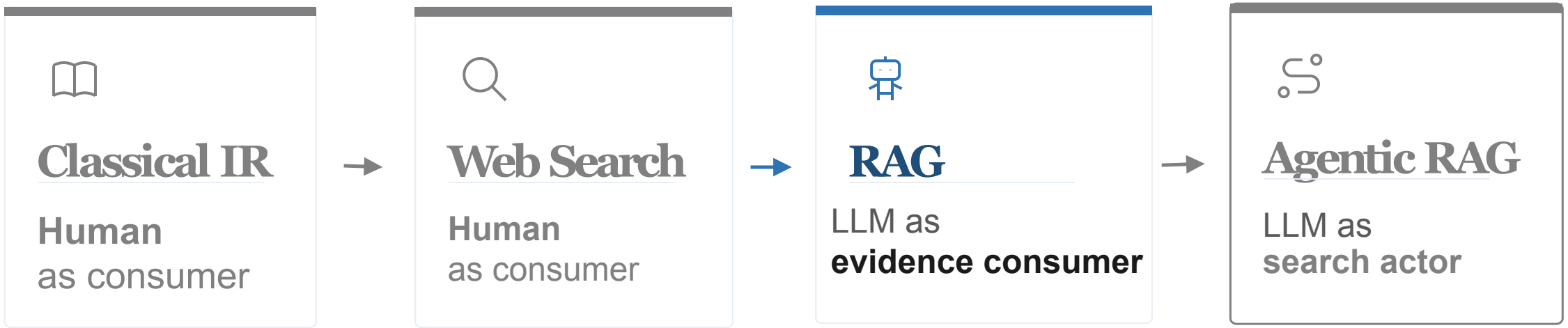
Beyond Individual Document Utility

Session Utility

Utility emerges across the **session**

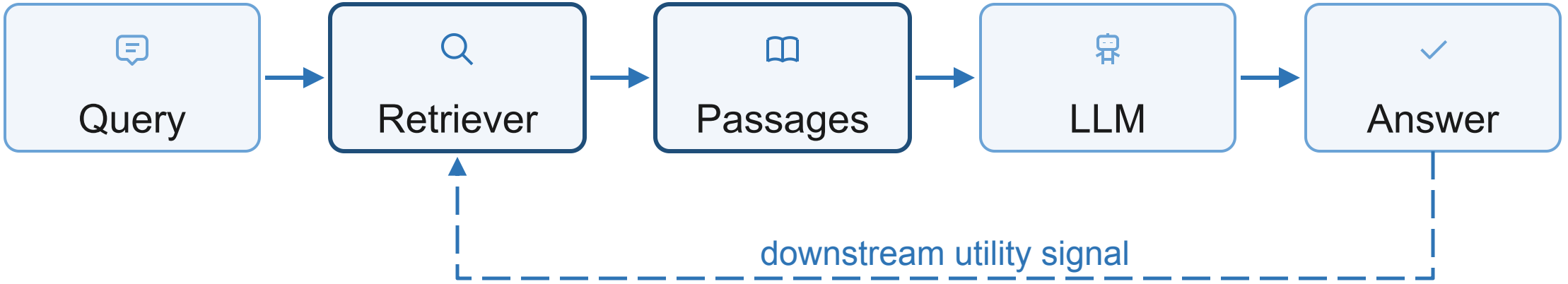


Liu et al. Investigating Cognitive Effects in Session-Level Search User Satisfaction. KDD, 2019.


Classical IR
Human
as consumer
Web Search
Human
as consumer
RAG
LLM as
evidence consumer
Agentic RAG
LLM as
search actor

LLM-Centric Utility

Retrieval is a component inside a generation pipeline.



Retrieved information is useful if it **improves the target LLM's output.**

What Counts as Improvement?

MORE
correct

MORE
faithful

MORE
efficient

Also: more complete, better reasoned, properly attributed, safer.

What Counts as Improvement?

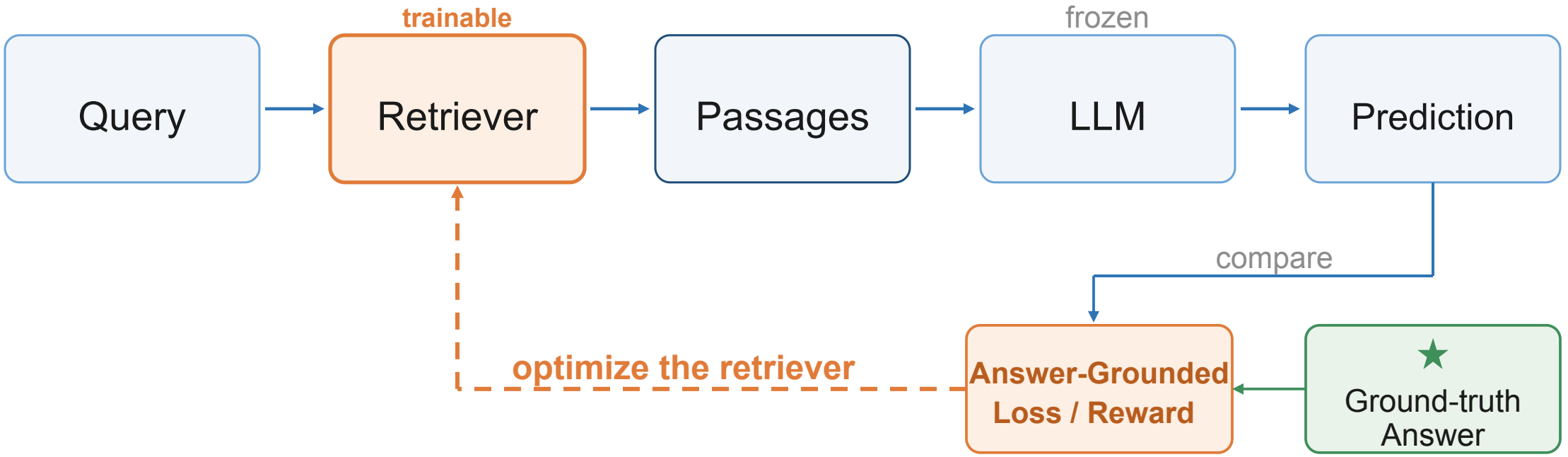
MORE **correct**

MORE faithful

MORE efficient

Also: more complete, better reasoned, properly attributed, safer.

Retrieval Optimized for the Answer



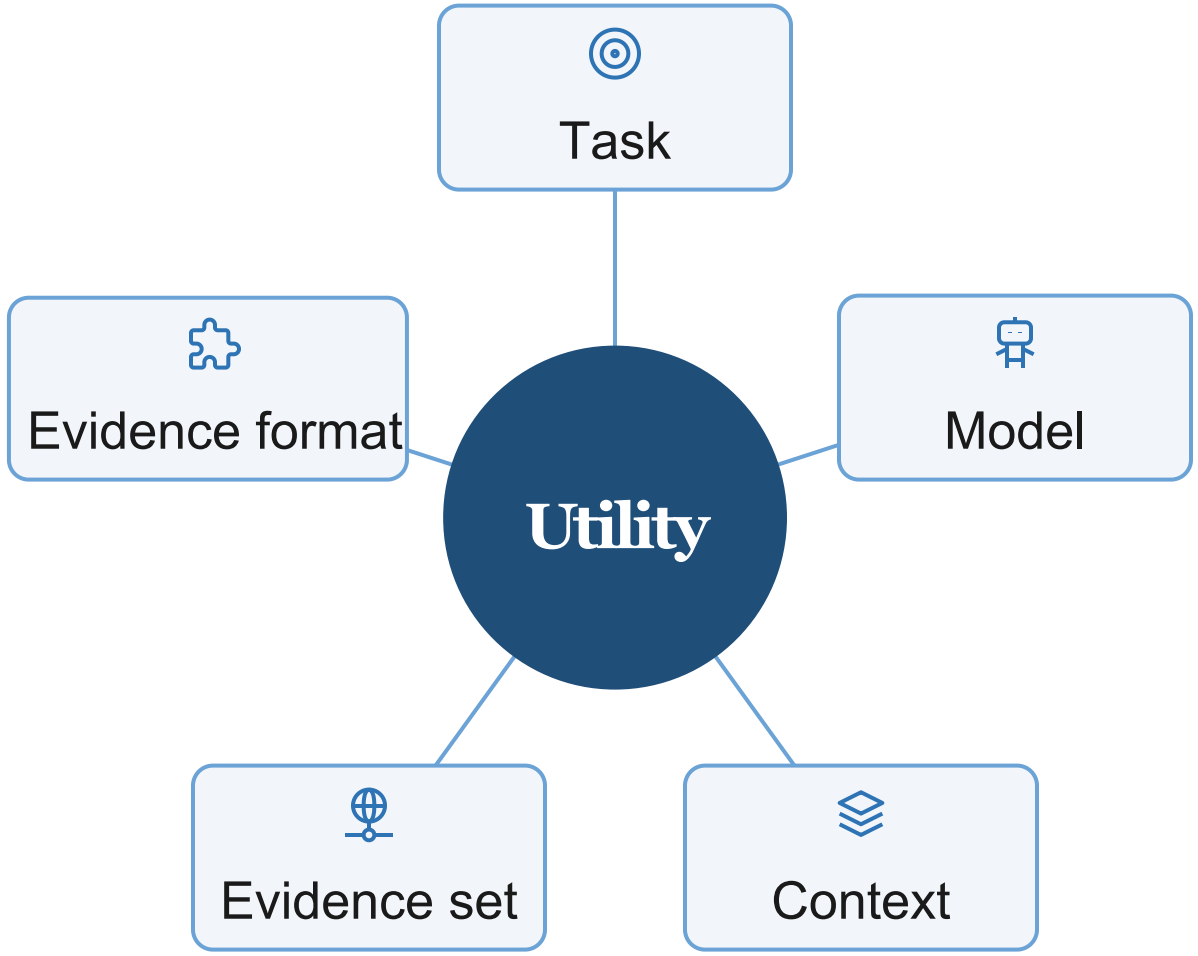
> The training target is answer quality, not topical relevance.

Shi et al. REPLUG: Retrieval-Augmented Black-Box Language Models. NAACL, 2024.
Bai et al. GripRank: Bridging the Gap between Retrieval and Generation via the Generative Knowledge Improved Passage Ranking. CIKM, 2023.

The Consumer Changes: Human → LLM

HUMAN CONSUMER	LLM CONSUMER
Skips irrelevant results	May consume **all** context
Notices source quality	May **over-trust** context
Compares sources	May **merge conflicts**
Reformulates the query	Needs a **policy or prompting**
Judges documents	Produces an answer **judged by the user**

Multi-Dimensional LLM-Centric Utility



Multi-Dimensional LLM-Centric Utility

- **Task** — does the evidence support the downstream task?
- **Model** — which LLM will consume it, and what it already knows.
- **Context** — what evidence and reasoning state already exist.
- **Evidence set** — do the passages jointly cover the need?
- **Evidence format** — is the form one the LLM can use?

When Relevance Fails

Missing answer

On topic, no answer-bearing fact

Redundant evidence

Repeats what is already known

Harmful evidence

Outdated, conflicting, unsupported

Wrong consumer

Depends on the target model

…

Relevant but Non-Answering

Q: When did Paris Agreement take effect?

RETRIEVED PASSAGE

The Paris Agreement is a legally binding international treaty on climate change. It brings countries together to reduce greenhouse-gas emissions, adapt to the effects of climate change, an….

no effective date

✓ Passage is on topic

✗ Answer-bearing fact is missing

■ **Utility is low**

On-topic is not the same as answer-bearing.

Model-Dependent Usefulness

Question: What is the capital of Australia?

SAME RETRIEVED PASSAGE

Canberra is the capital of Australia. Sydney is the largest city; Melbourne was once the temporary seat.

Weak model	→	**evidence helps answer correctly**
Strong model	→	**evidence mostly repeats memory**
Small model	→	**Sydney and Melbourne can distract**

Same evidence, same query, different consumer model — different utility.

Relevant but Redundant

Question: When did the James Webb Space Telescope launch, and from where?

PASSAGE P1

The Webb telescope launched on **December 25, 2021**, on an Ariane 5 from **Kourou, French Guiana**.

≈

same facts

PASSAGE P2

Webb lifted off on **Christmas Day 2021**, aboard Ariane 5 from **Kourou, French Guiana**.

P2 is relevant, but its **marginal utility is low**.

Relevant but Harmful

Question: How many planets are officially recognized in the Solar System?

AUTHORITATIVE EVIDENCE

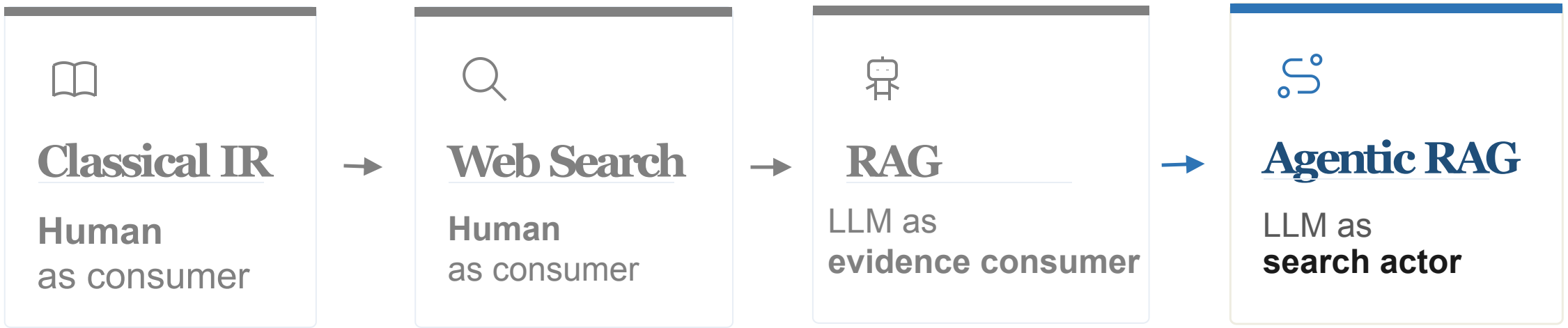
IAU 2006 definition: the Solar System has **eight planets**; Pluto is a dwarf planet.

STALE RETRIEVED PASSAGE

A pre-2006 page lists Mercury through Pluto as the **nine planets** of the Solar System.

Both passages are on topic.

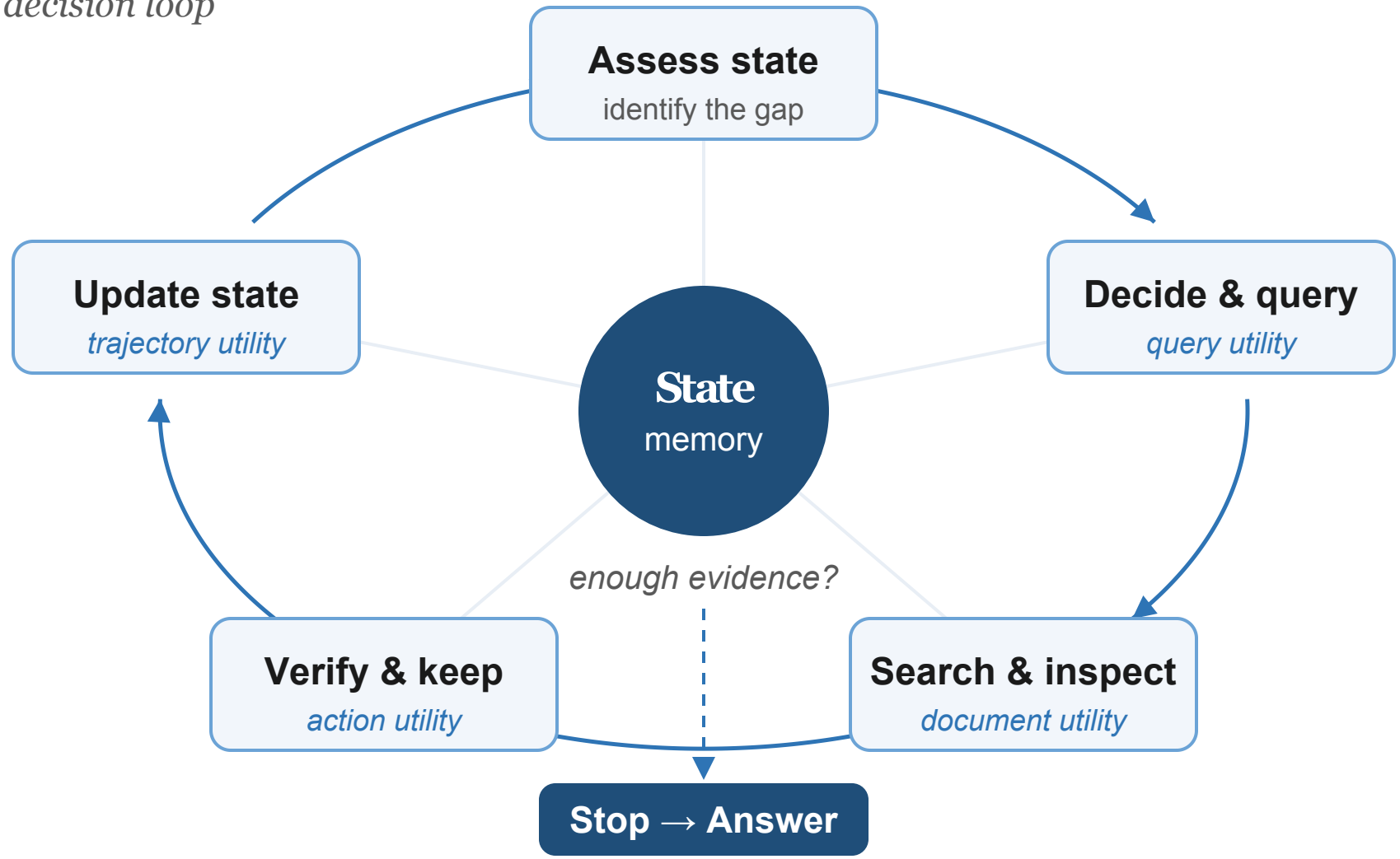
But one pulls generation toward the wrong answer.



Static RAG changes who consumes retrieval.
Agentic RAG changes when and why retrieval happens.

How Agentic RAG Works

Retrieval as a decision loop



Static RAG vs. Agentic RAG

Dimension	Static RAG	Agentic RAG
Timing	Retrieve once, then generate	Retrieve & verify across steps
Control flow	Query → top-k → answer	Plan → search → reason → repeat
State	Fixed retrieved context	Evolving state & memory
Utility question	Does evidence help the answer?	Does this action help progress?

Static RAG already made relevance insufficient; agentic RAG raises the bar again.

Agent-Centric Utility

- One-shot retrieval asks for useful **context before** generation.
- Agentic retrieval asks for useful **progress during** problem-solving.

A passage earns utility when it —

- answers the current subquestion
- verifies or corrects a claim
- reveals the next missing fact
- rules out an incorrect path

Utility depends on the agent's current state and the decision being made.

Three Utility Objects in Agentic RAG

Query utility

Right thing to search for now?

e.g., target the missing entity or date

Document utility

Does this evidence help the step?

e.g., locally relevant, answer-supporting

Action utility

Search, verify, summarize, or stop?

e.g., verify when sources disagree

Trajectory utility

Did the whole sequence succeed?

e.g., supported answer, fewer calls

supplies training signals of the first three

From Evidence Utility to Agentic Utility

1. **Human-centric search** — utility tied to user task success.
2. **Static RAG** — utility tied to whether evidence helps the output.
3. **Agentic RAG** — utility follows each query, document, and action.

⟼ Section 2 formalizes utility; Section 4 revisits it for agents.

Worked Example — Deep Research

Which recent papers train retrievers for deep-research agents — and how do they differ?

Static RAG behavior

- Retrieve top-k papers once
- Generate comparison from fixed context
- Risk: may miss a key paper or detail

Agentic RAG behavior

- Search, inspect titles & abstracts
- Separate agentic vs deep-research scope
- Verify it truly trains a retriever
- Summarize the final evidence set

✓ Useful if it supports the comparison — not if it merely says “agentic search.”

A Related View: Denoising-First IR

LLM-oriented IR should **reduce** harmful or low-value context.

NOISE COMES FROM

Corpus · Retriever · Context assembly · Agent loops

Dai et al. LLM-Oriented Information Retrieval: A Denoising-First Perspective. SIGIR, 2026.

A Related View: Denoising-First IR

LLM-oriented IR should **reduce** harmful or low-value context.

NOISE COMES FROM

Corpus · Retriever · Context assembly · Agent loops

Noise wastes tokens, attention, latency, and reasoning capacity.

- Wastes limited context length.
- Distracts attention from answer-bearing evidence.
- Adds latency and cost.
- Causes unsupported, hallucinated answers.

Dai et al. LLM-Oriented Information Retrieval: A Denoising-First Perspective. SIGIR, 2026.

Denoising vs. Utility-Centric

Denoising: remove the harmful

Utility-centric: optimize the **useful**

What to retrieve, select, transform, order, and optimize —
for this model, task, and context.

Section 1 — Takeaways

1. Retrieval now serves generation: the LLM is the consumer.
2. Relevance is necessary, but no longer sufficient.
3. Utility depends on task, model, context, set, and format.
4. Agentic RAG adds query and action utility in addition to document.

⟼ Next — Section 2: what is LLM-centric utility?

Tutorial Roadmap

180 min total

01 **Introduction & Foundations -Keping** — 40 MIN
Why retrieval objectives must change for LLM consumers

02 **What Is LLM-Centric Utility? -Keping** — 50 MIN
Task, model, context, and presentation dimensions

03 **Utility Modeling & Optimization -Keping** — 40 MIN
Signals, labels, trainable components, and evaluation

04 **Utility in Agentic RAG -Keping** — 40 MIN
Utility in dynamic retrieval and agent loops

05 **Open Problems & Q&A -Keping** — 10 MIN
Open problems and discussion

Tutorial website