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.

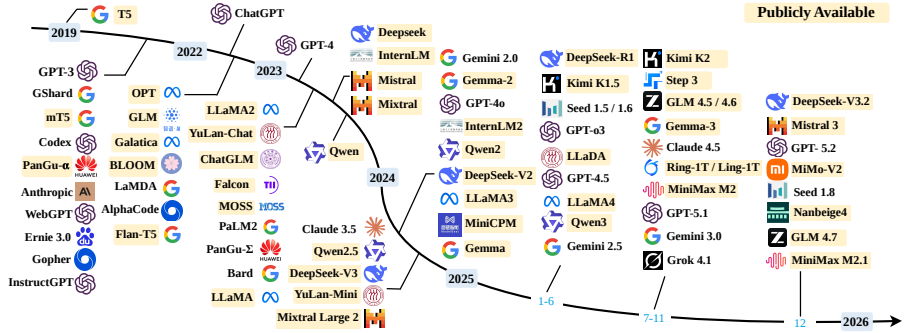
Foundations · LLM-Centric Utility · Modeling & Optimization · Agentic RAG · Open Problems

The Evolution of Information Access

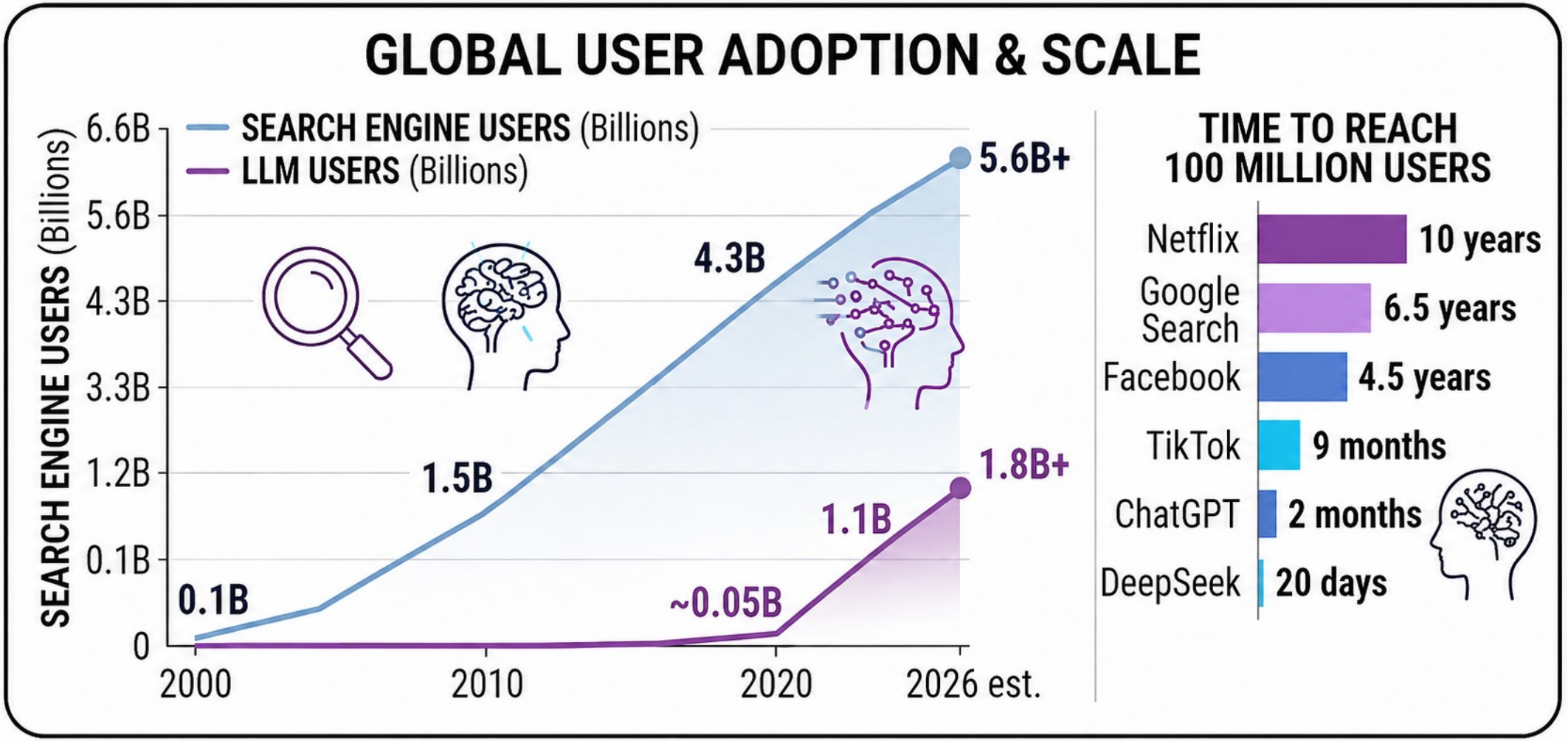
1960s
Online Database Retrieval Era

1990s
Web Search Era

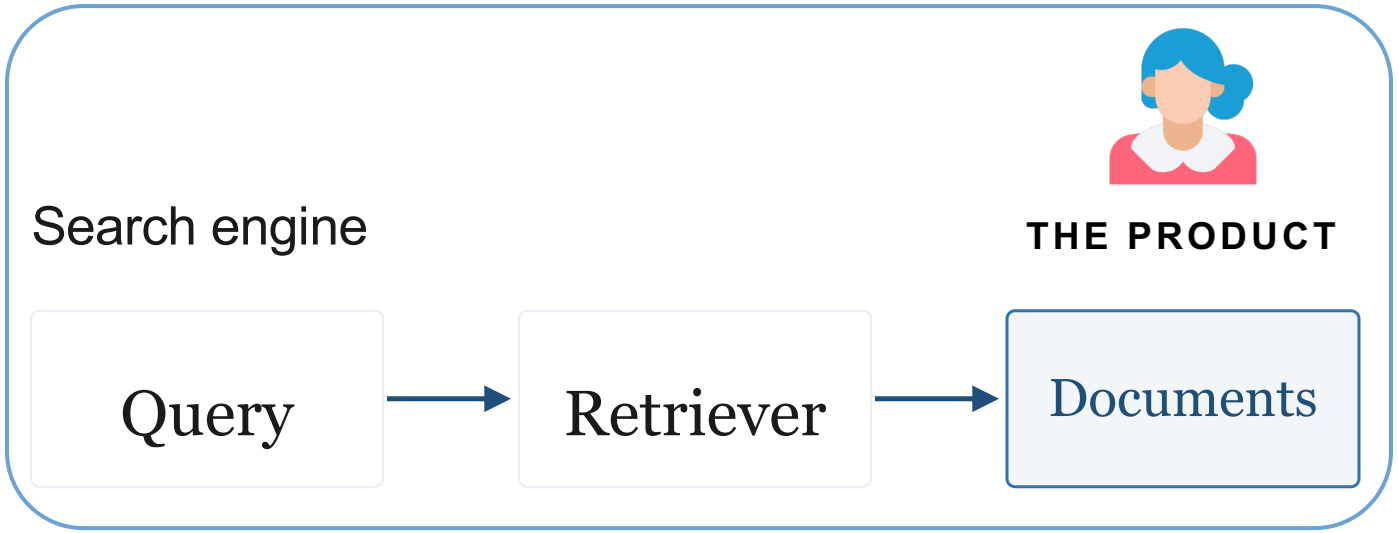
2020s
Generative AI Era



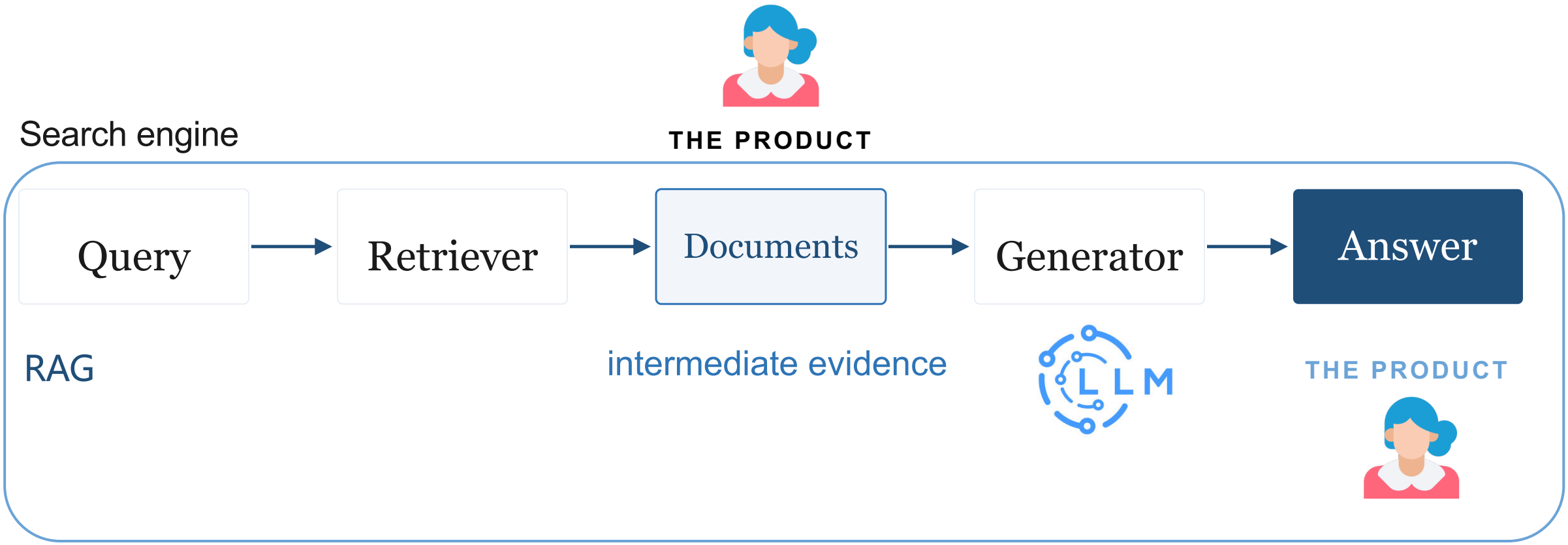
The Evolution of Information Access



RAG Changes the Product

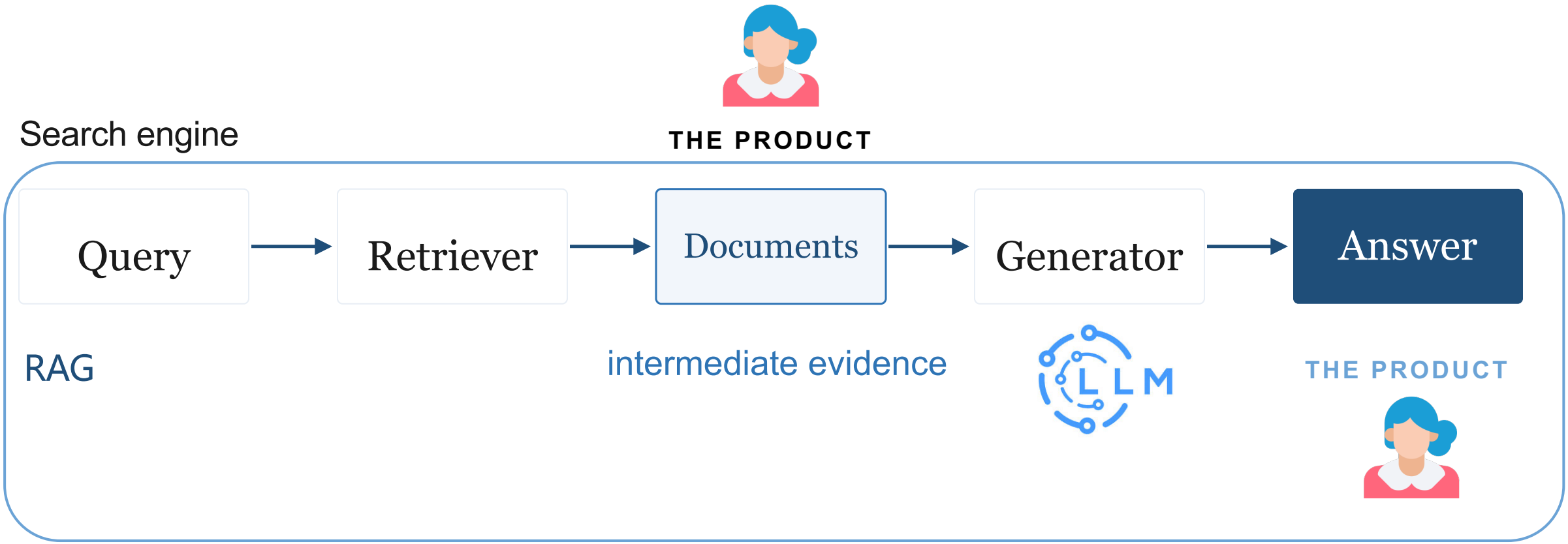


RAG Changes the Product



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

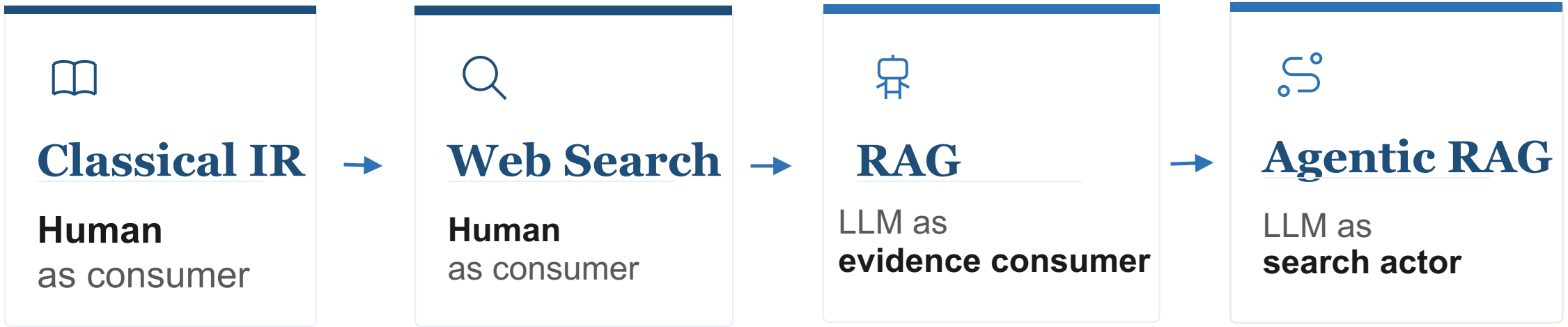
RAG Changes the Product



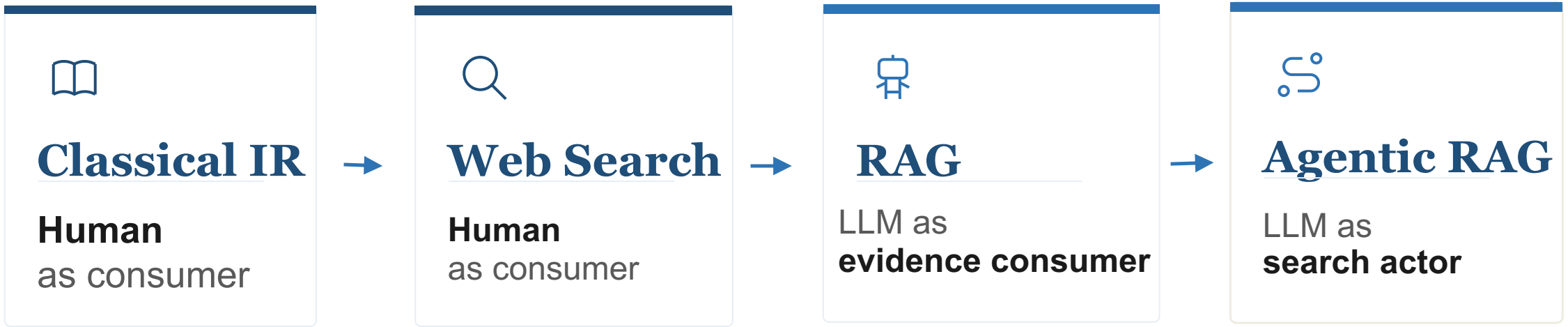
The consumer of retrieval results shifts from users to LLMs.

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



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

1960s	1970s	2000s	2010s	2020s
Topical relevance defined	Contextual user & task context	Fine-grained Learning to Rank	Task utility scenario-oriented	**LLM utility** serves generation

relevance → utility

SMART · Dialog · MEDLINE · AIM-TWX · MEDLARS

DATABASE SEARCH SYSTEMS

Google · Baidu 百度 · yahoo! · Microsoft Bing

WEB SEARCH ENGINES

deepseek

LLM ASSISTANTS

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

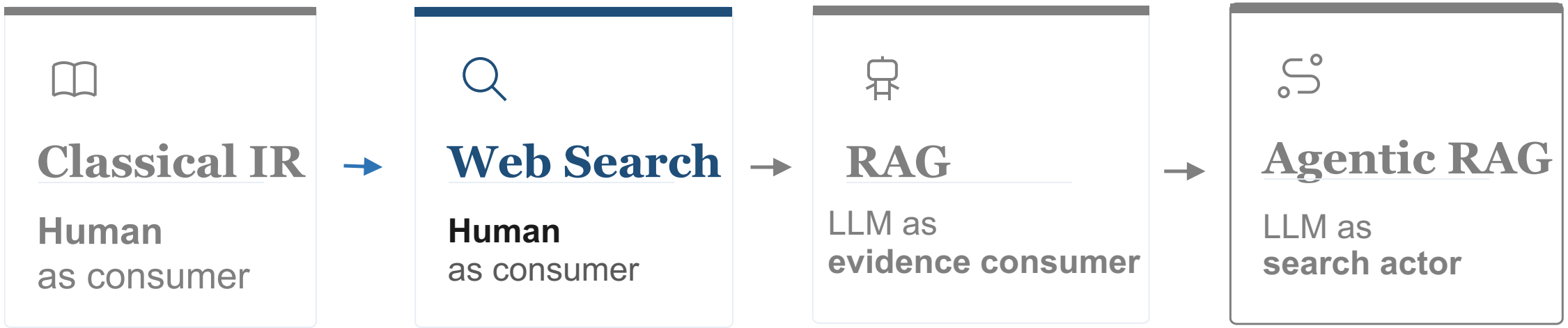
RELEVANCE

- 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		THE UTILITY THEY STAND IN FOR
Clicks	approximate →	Attraction
Dwell time		Satisfaction
Reformulation		Task progress
Session length		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

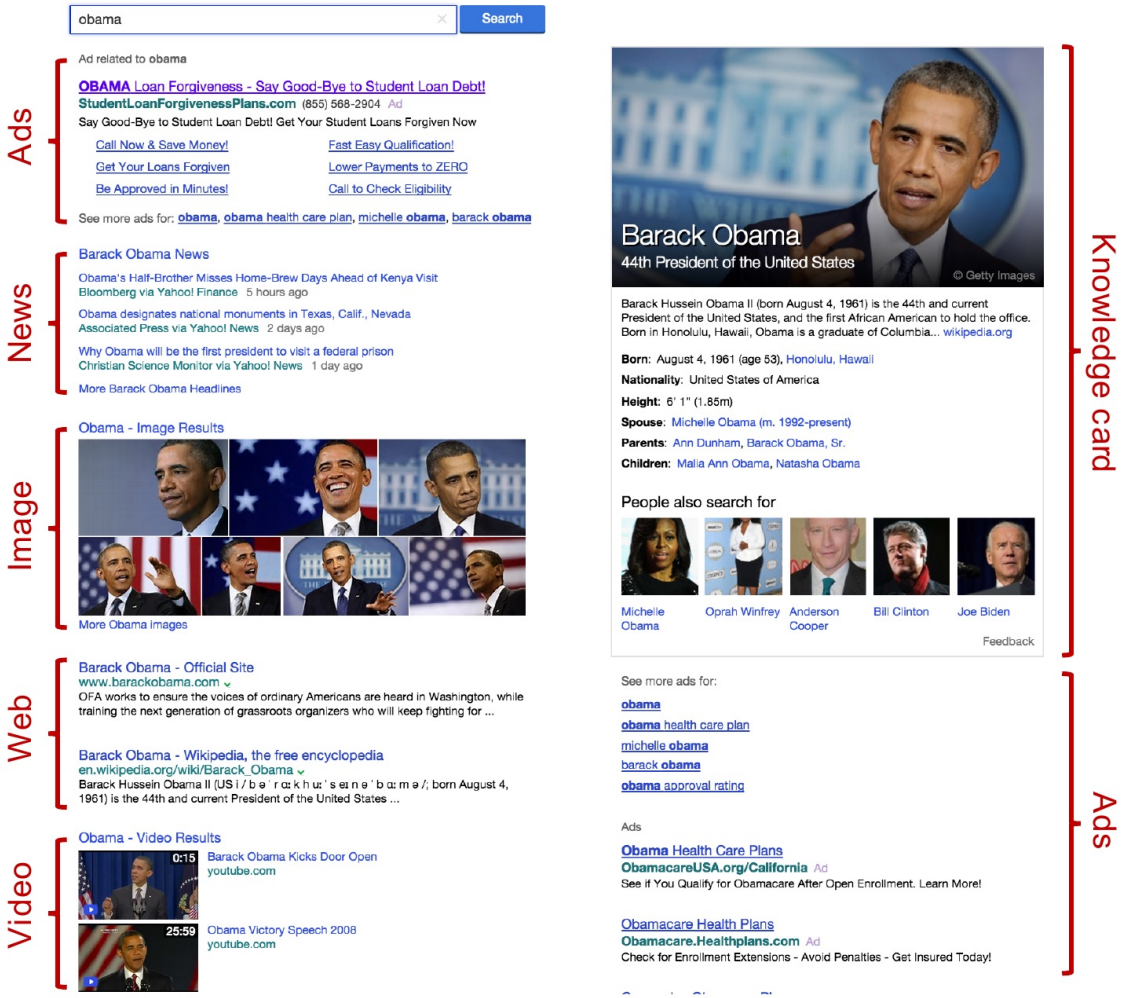
Each one can **bias** the utility signal we read from behavior.

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**

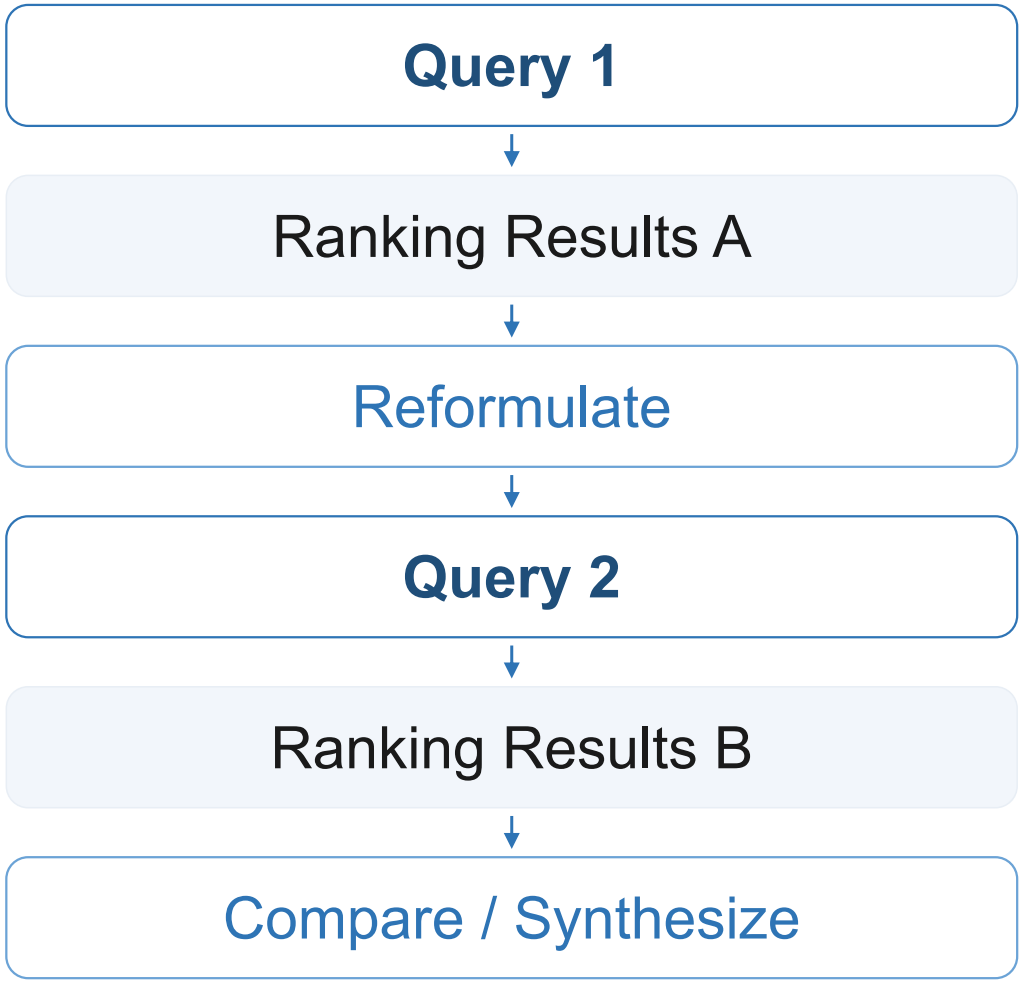


Wang et al. Optimizing Whole-Page Presentation for Web Search. WSDM, 2016.

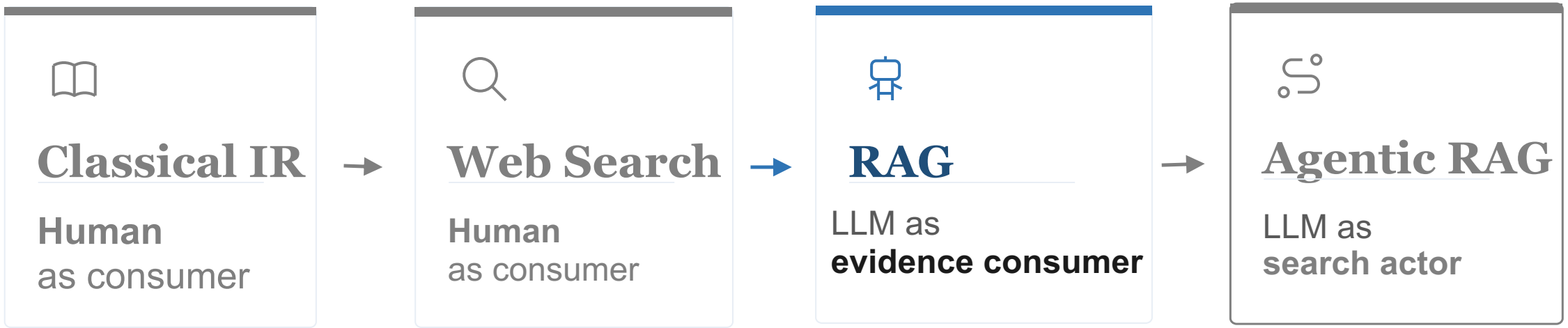
Beyond Individual Document Utility

Session Utility

Utility emerges across the **session**

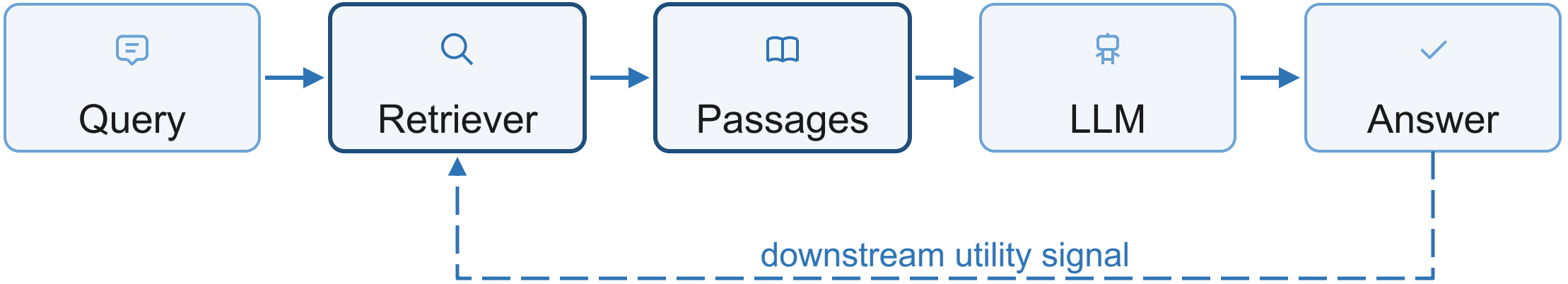


Liu et al. Investigating Cognitive Effects in Session-Level Search User Satisfaction. KDD, 2019.


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?

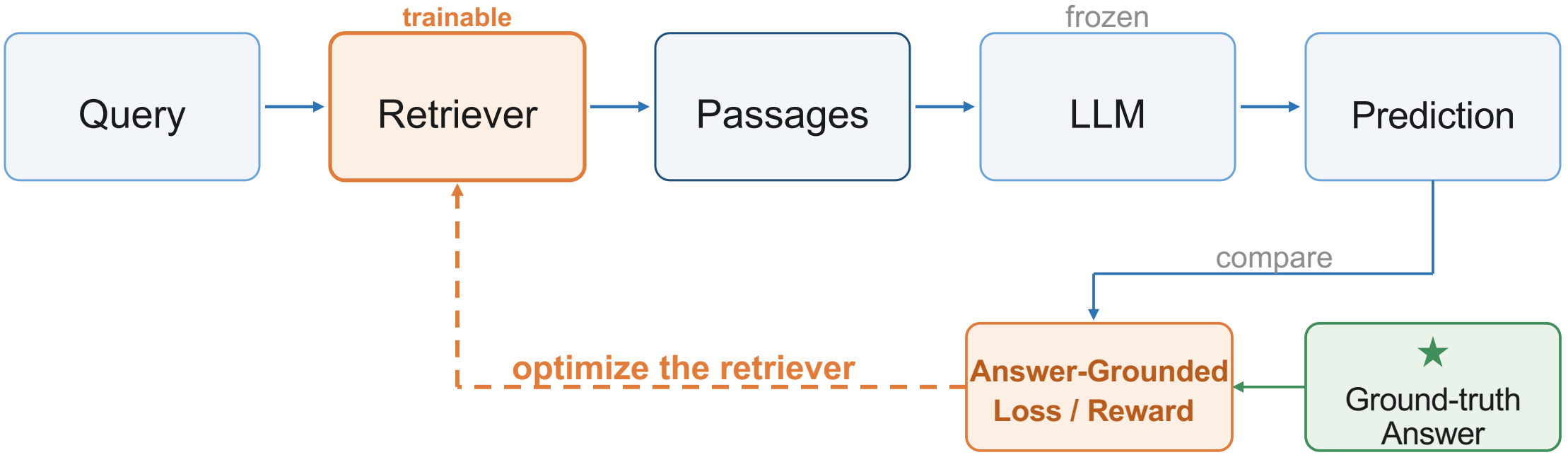
MORE **correct**

MORE faithful

MORE efficient

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

Retrieval Optimized for the Answer



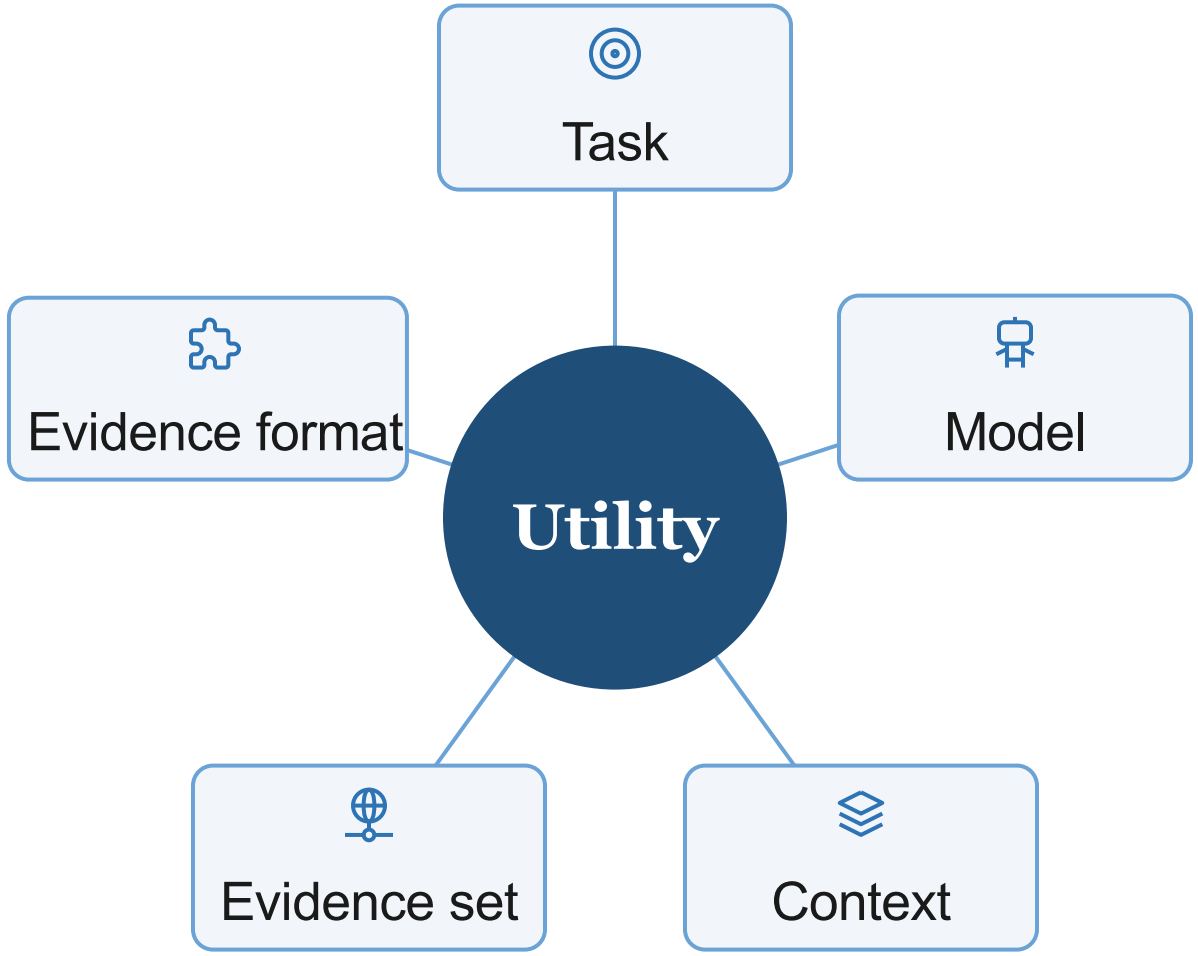
> The training target is answer quality, not topical relevance.

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

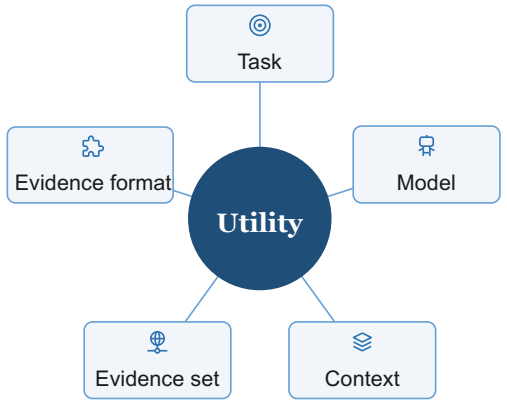
HUMAN CONSUMER	LLM CONSUMER
Skips irrelevant results	May consume **all** context
Notices source quality	May **over-trust** context
Compares sources	May **merge conflicts**
Reformulates the query	Needs a **policy or prompting**
Judges documents	Produces an answer **judged by the user**

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

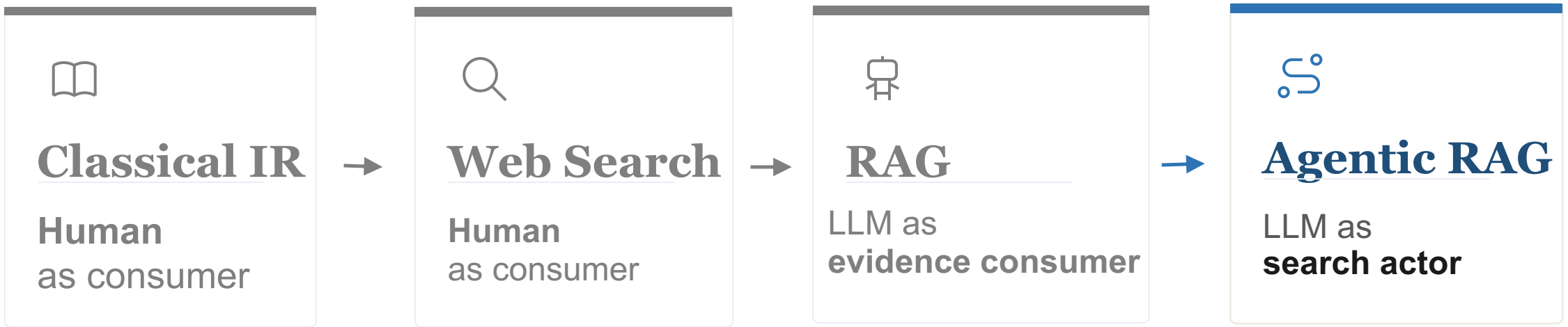
IAU 2006 definition: the Solar System has **eight planets**; Pluto is a dwarf planet.

STALE RETRIEVED PASSAGE

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

Both passages are on topic.

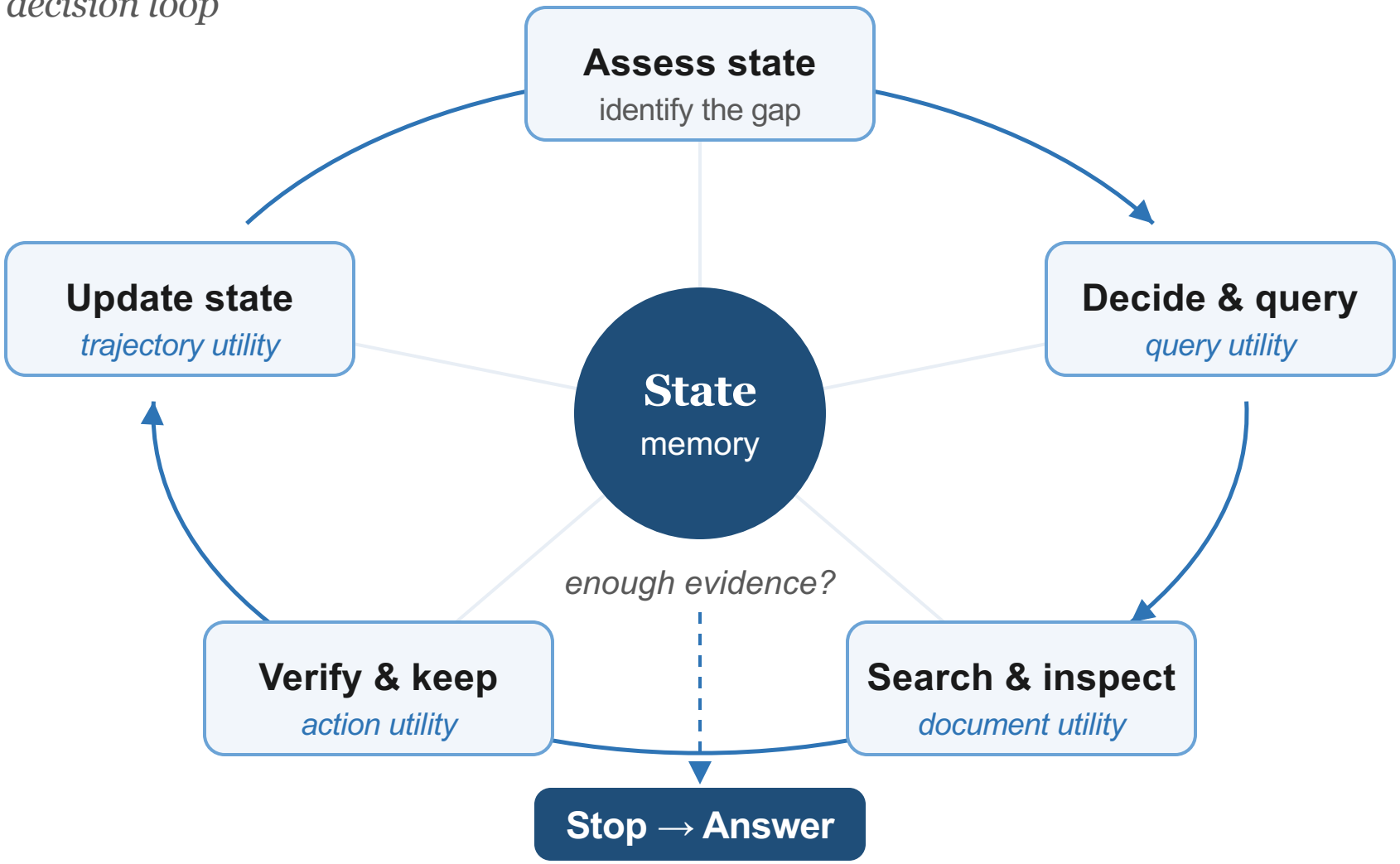
But one pulls generation toward the wrong answer.



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.

An 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

SECTION 2

What Is LLM-Centric Utility?

A working definition — and the five dimensions that make it concrete.

A Working Definition

> LLM-centric utility is the extent to which retrieved information improves a **target LLM**'s downstream generation, reasoning, verification, or action quality — under a specific **task**, **model**, **context**, **evidence set**, and **evidence format**.

Every highlighted term is a dimension — this section walks through them one by one.

Unpacking the Definition

The definition **is** the roadmap — each phrase names one dimension.

"a specific task" → **Task**

What downstream problem?

"a target LLM" → **Model**

Which LLM consumes it?

"context" → **Context**

What is already known?

"evidence set" → **Evidence set**

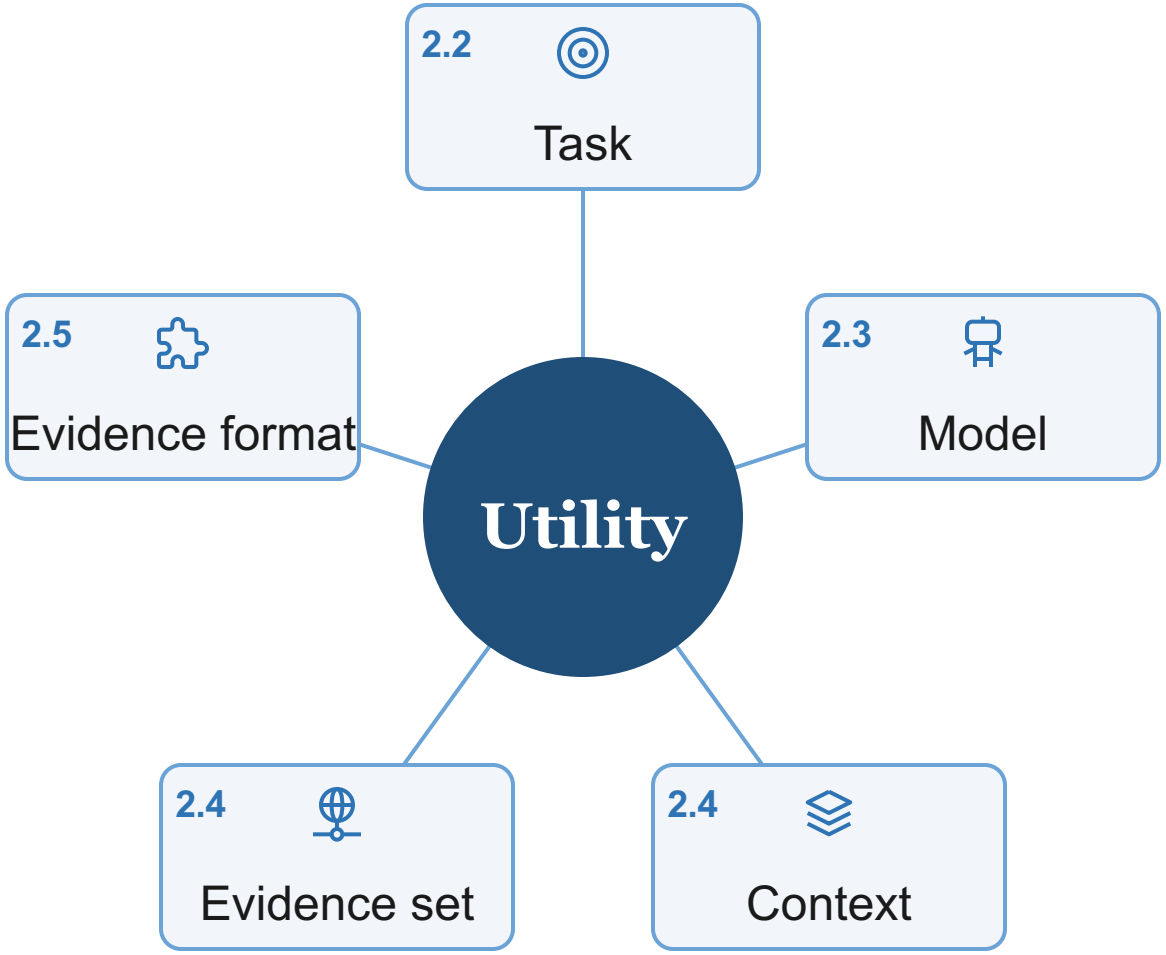
Does the set jointly cover the need?

"evidence format" → **Format**

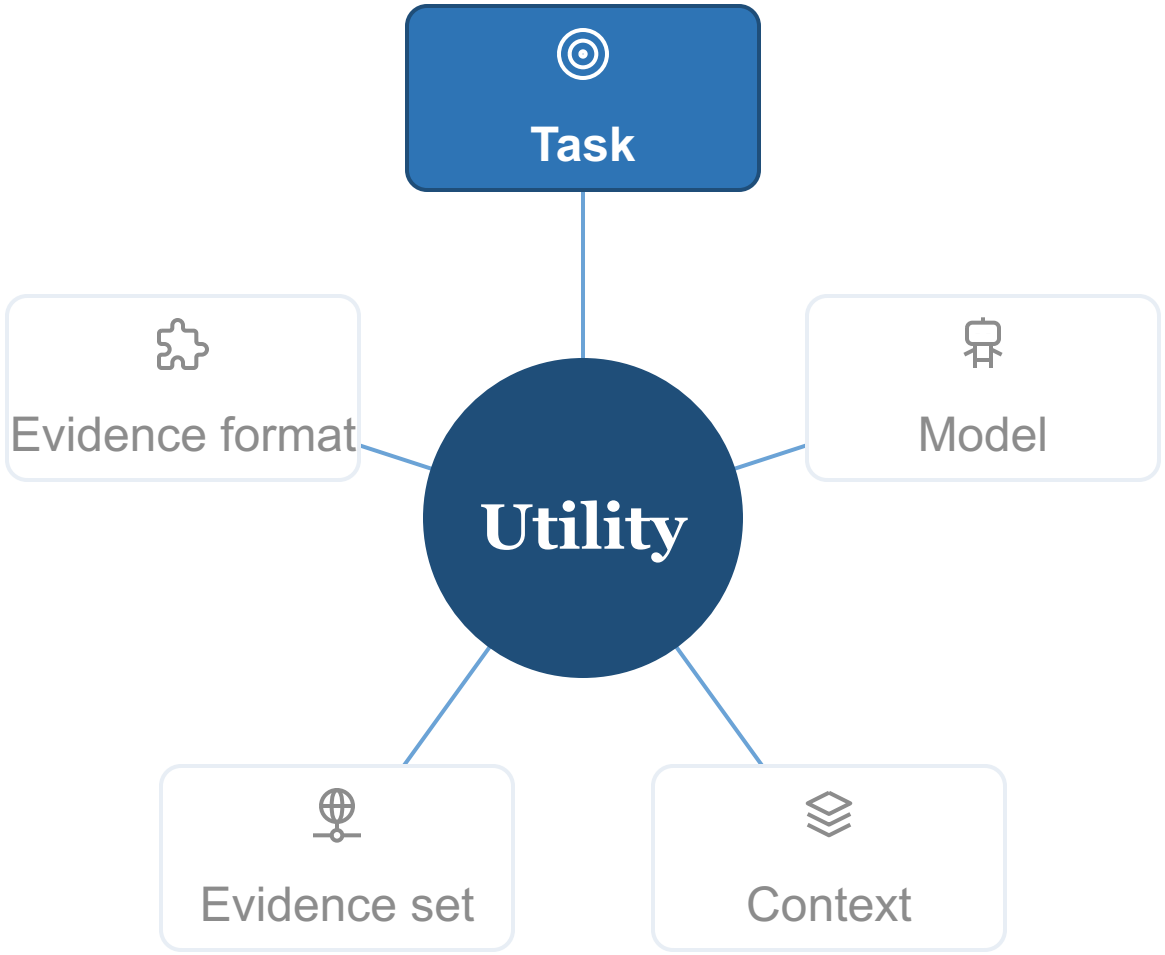
Is the form usable?

One definition, five dimensions — the roadmap for the next fifty minutes.

The Road Through This Section



2.2 · The Task Dimension



What counts as **useful** evidence is defined by the downstream task.

Task-Specific Utility

What counts as **useful evidence** in different downstream tasks?

Factoid QA

contains or entails the answer

Non-Factoid & Complex QA

improves the answer or report

Fact Verification

supports the verdict

Code Generation

helps the code run

Tool & Skill Retrieval

enables the next action

Factoid QA

For factoid QA, evidence is useful when it…

- **contains or entails the answer** — not just the topic;
- helps the model **output the correct answer**;
- supports **grounding and citation** of that answer;
- reduces the model's **answer uncertainty**.

> Topical relevance ≠ answer-bearing support

Factoid QA: Answer Support

SECTION 1 example revisited

Q: When did the Paris Agreement take effect?

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

⚠ **No effective date stated**

- ✓ On topic
- ✗ Contains the answer
- ✗ Supports citing the date
- ✗ Reduces answer uncertainty

→ Utility for generation: **low**

Section 1 showed you this failure — the Task dimension names why it fails.

Complex QA

No single ground-truth answer — utility is judged by the quality of the **generated answer or report**. Useful evidence helps the model produce:

Comprehensive answer

All aspects of the question are covered.

Multi-aspect synthesis

Pieces combine into one coherent account.

Balanced comparison

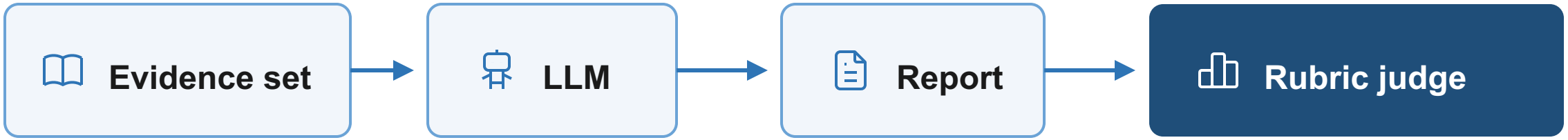
Competing views are weighed fairly.

Well-supported explanation

Claims are grounded in cited evidence.

Deep Research: Report Quality

Evidence utility is **mediated by the final report** the agent writes.



RUBRIC DIMENSIONS

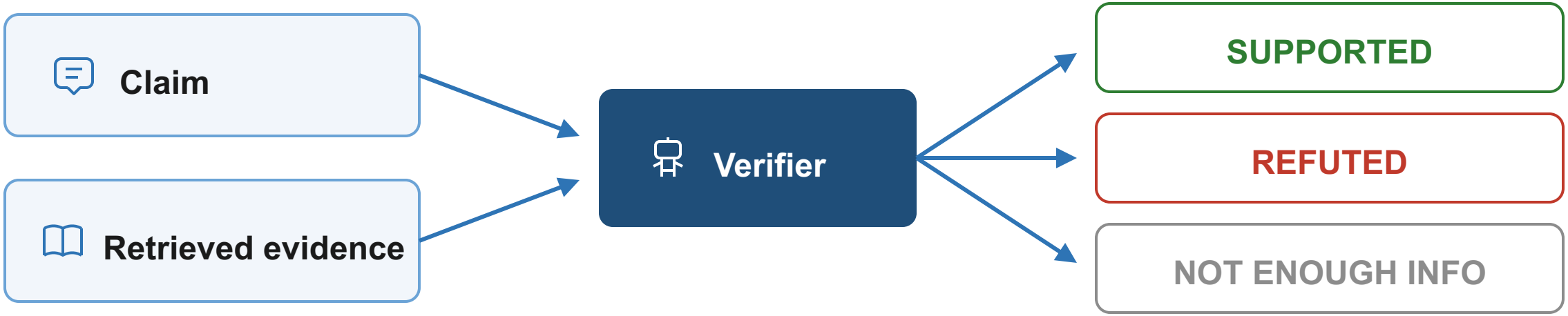
- comprehensiveness
- insight / depth
- instruction-following
- readability

Rubric-based **LLM-as-judge** scores report quality; citations are checked separately.

Du et al. DeepResearch Bench: A Comprehensive Benchmark for Deep Research Agents. ICLR, 2026.

Fact Verification

Useful evidence is **decision support** for a claim.



USEFUL VERIFICATION EVIDENCE IS

- relevant to the claim
- sufficient for the verdict
- faithful to the explanation
- from an authoritative source

Zhang et al. From Relevance to Utility: Evidence Retrieval with Feedback for Fact Verification. Findings of EMNLP, 2023.
Hu et al. Read It Twice: Towards Faithfully Interpretable Fact Verification by Revisiting Evidence. SIGIR, 2023.

Code: Run, Not Match

Utility is **not similarity** to the current code context.

× Looks similar

```
def parse_config(path):
# a near-duplicate helper,
# same tokens, no new facts
```

Text overlap with the context — but nothing the model can act on.

Actually helps

· API usage · dependency constraints
· repository structure · test behavior
· worked examples · repair hints

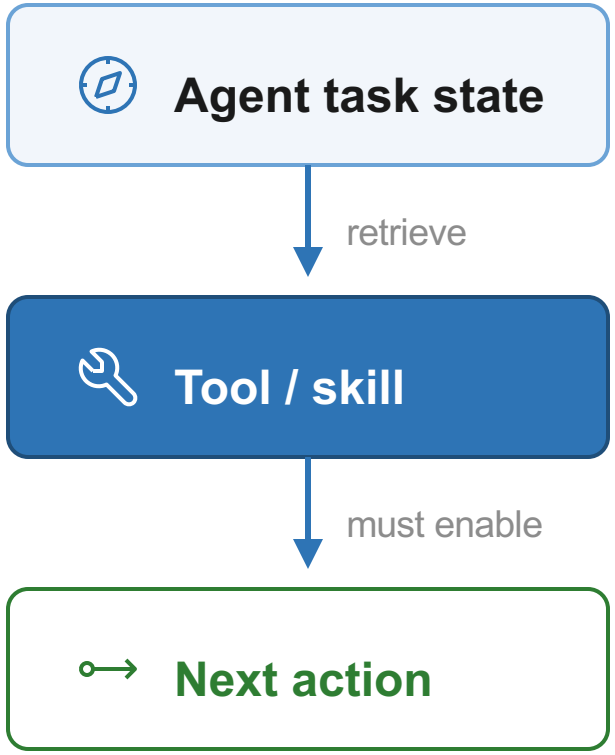
Whatever helps produce executable, context-compatible code next step.

Signal = **execution**: pass rate, repair success, preference over code.

Gao et al. Preference-Guided Refactored Tuning for Retrieval Augmented Code Generation. ASE, 2024.
Saberi and Fard. Context-Augmented Code Generation Using Programming Knowledge Graphs. arXiv, 2024.
Dong et al. SelfRACG: Enabling LLMs to Self-Express and Retrieve for Code Generation. EMNLP, 2025.

Tool & Skill Retrieval

A tool whose **description matches the query** is not necessarily useful.



A USEFUL TOOL OR SKILL:

- **enables the next action**
- matches the current environment
- exposes the needed API or affordance
- is reliable when invoked
- reduces uncertainty in the trajectory

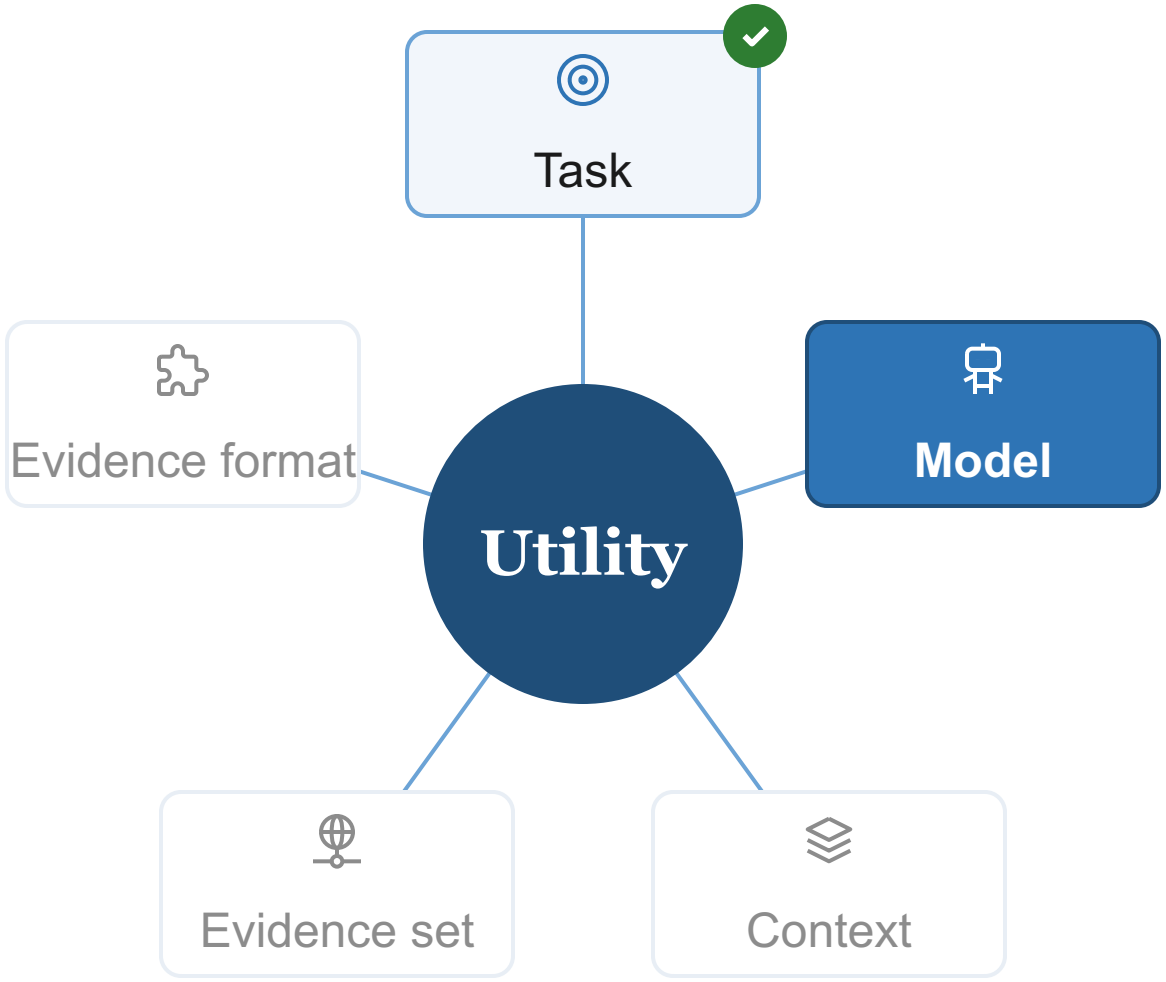
Underexplored — an open research direction. Agent-trajectory methods return in Section 4.

Task-Specific Utility: Recap

Task	Useful evidence means…	Evaluation signal
Factoid QA	contains or entails the answer	answer correctness
Complex QA	coverage, synthesis, decision usefulness	answer / report quality
Fact verification	sufficient, faithful decision support	verdict + explanation
Code generation	helps produce runnable, compatible code	execution / pass rate
Tool & skill retrieval	enables the next action	action success

Each task redefines both **"useful"** and **the signal that measures it**.

2.3 · The Model Dimension



Useful — for which model?

Utility can be judged for LLMs in general — or for the one model that will actually consume the evidence.

Two Views of the Model Dimension

ANY MODEL

LLM-agnostic utility

· **average** usefulness across generators

· one label serves many systems

· scales supervision cheaply

"Would this help an LLM?"

vs

THIS MODEL

LLM-specific utility

· judged for the **actual target model**

· same passage, different value per model

· aligns with the real consumer

"Would this help this LLM, here?"

As information consumers, LLMs may exhibit personalized preferences, much like humans.

LLM-Agnostic Utility

The scalable view: assume usefulness **transfers across generators**.

GOOD FIT FOR

- Scalable dataset construction
- General retriever training
- Reusable labels
- Simpler evaluation

TWO TYPES of AGNOSTIC LABELS

LLM utility judgments

with or without ground-truth answers

(Zhang et al.2026)

Generator's performance

The LM's answer likelihood indicates utility

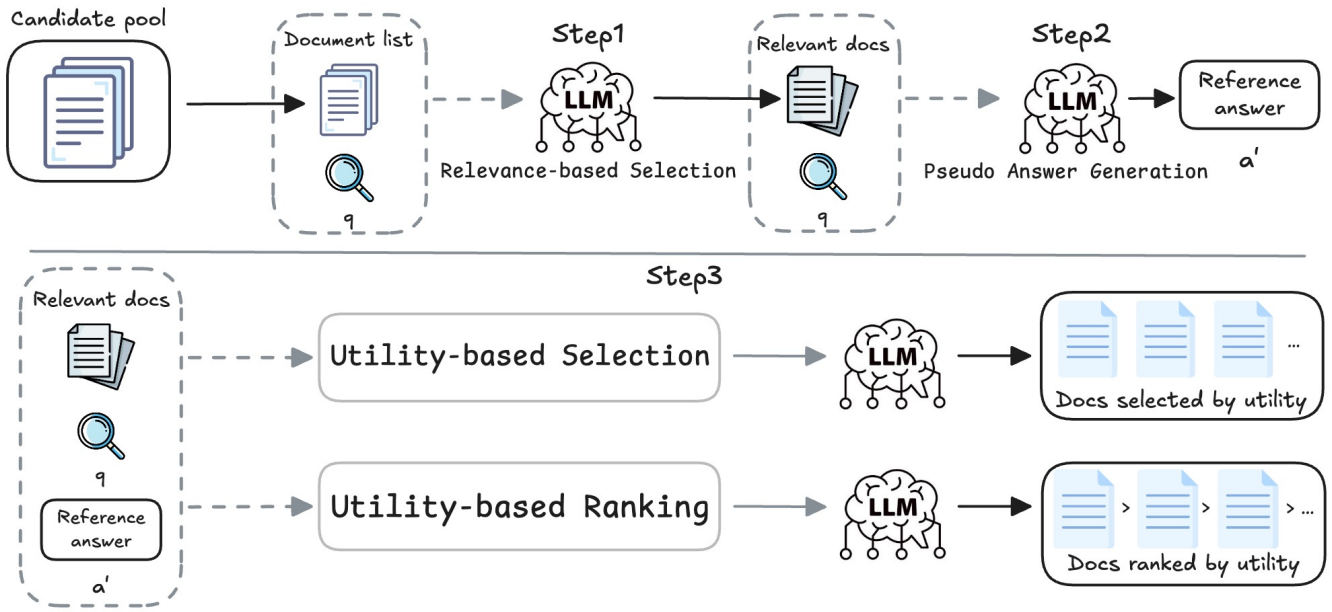
(Shi et al., 2025)

Zhang et al. Utility-Focused LLM Annotation for Retrieval and Retrieval-Augmented Generation. EMNLP, 2025.

Shi et al. REPLUG: Retrieval-Augmented Black-Box Language Models. NAACL, 2024.

Utility-Focused LLM Annotation

Label utility with LLM judgments — **generic labels indicating utility**



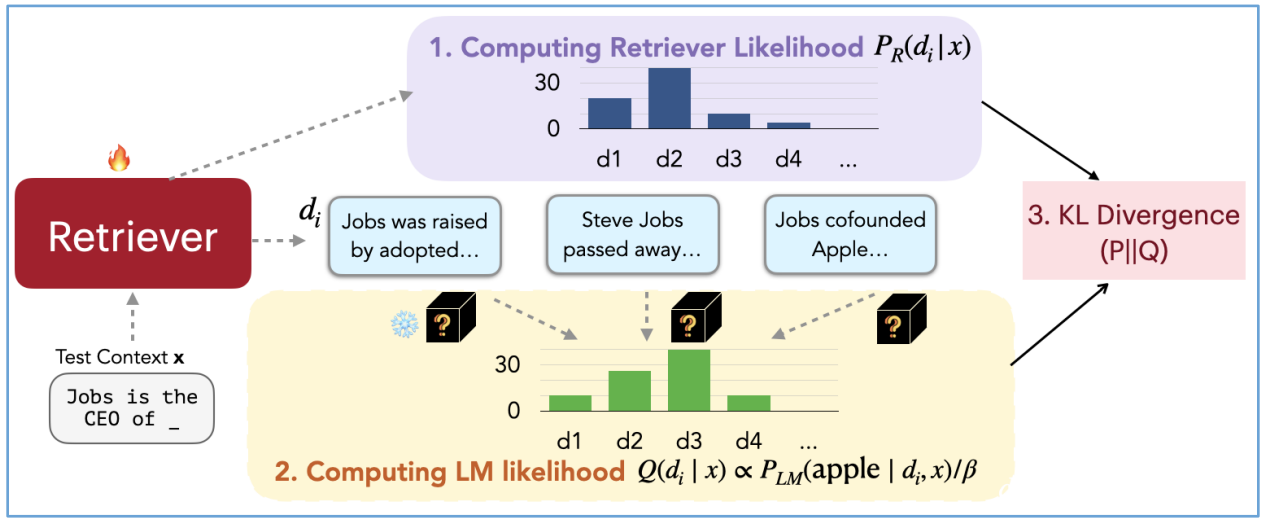
LLM judgments without manually annotated answers

Zhang et al. Utility-Focused LLM Annotation for Retrieval and Retrieval-Augmented Generation. EMNLP, 2025.

REPLUG: Performance as the Signal

The black-box LM tells the retriever **which passages help it answer.**

LM-SUPERVISED RETRIEVAL (LSR)



THE TRAINING SIGNAL

> **$Q(d \mid x) \propto P_{LM}(\text{answer} \mid d, x)$**
> the frozen black-box LM's own answer likelihood

1. Score each doc by that likelihood
2. Train retriever to match it (KL)

A generator-aware score — **reused to serve other generators.**

One LM's likelihood supervises the retriever (experimented on blackbox LLMs of same-families)

Shi et al. REPLUG: Retrieval-Augmented Black-Box Language Models. NAACL, 2024.

LLM-Agnostic — Risks

Generic labels may hide the target model's needs.

- What helps an average model may not help yours.
- The transfer assumption can quietly fail.

> Which is why we also need LLM-specific utility.

Canberra, Revisited

SECTION 1, REVISITED

Q: What is the capital of Australia?

Same passage: "Canberra is the capital. Sydney is the largest city..."

Weak model ✓

evidence helps it answer correctly

Strong model ≈

mostly repeats what it already knows

Small model ✗

Sydney and Melbourne distract it

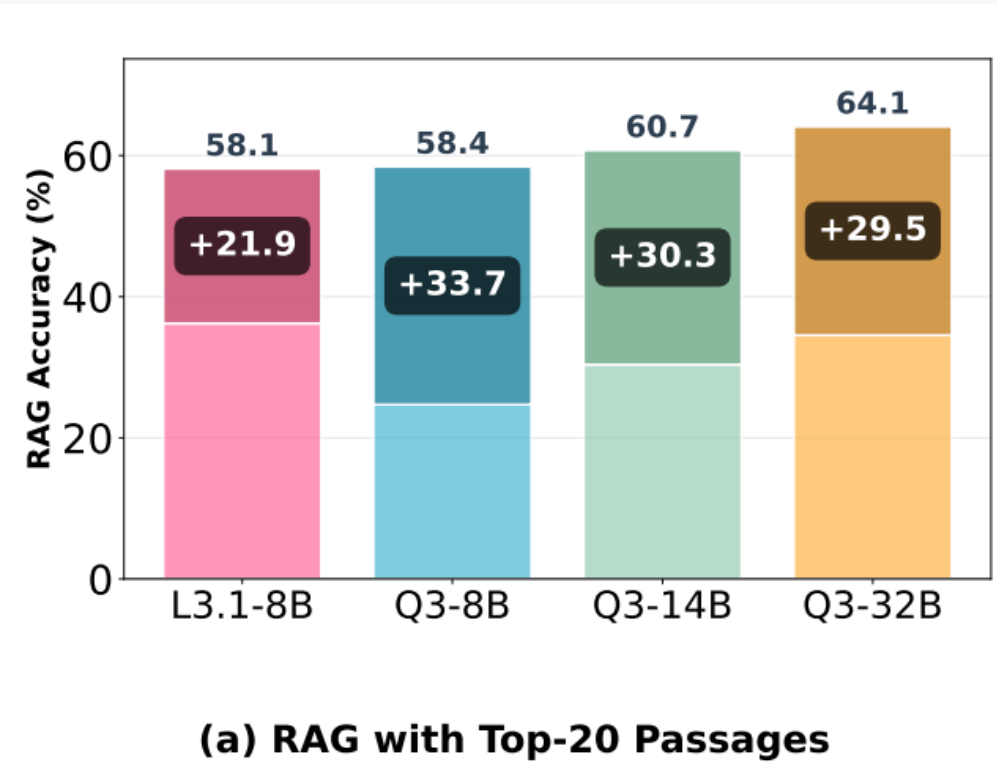
MODELS DIFFER IN

- parametric knowledge
- reading ability
- reasoning ability
- long-context behavior
- conflict sensitivity

LLM-Specific Utility: A New Perspective for Retrieval-Augmented Generation. 2025.

LLM-Specific Utility

Top-20 RAG lifts every generator on NQ — **but by very different margins**.



WHAT THE CHART SHOWS

- All four generators gain from top-20 RAG.
- But the gain differs sharply across models.
- Weaker models gain most; strong ones least.

Same passages, different value per model.

Utility must be judged for one model.

Zhang et al. LLM-Specific Utility: A New Perspective for Retrieval-Augmented Generation. arXiv, 2025.

LLM-Specific Utility

A passage is utilitarian only if it **improves the LLM's answer** without evidence.

A UTILITARIAN PASSAGE IMPLIES

1. The LLM does not already possess this knowledge.
2. The passage contains answer-related information.
3. The passage can be understood and used by the LLM.

AND

Utilitarian for this LLM

all three must hold

Miss any one, and it is not utilitarian for that model.

Zhang et al. LLM-Specific Utility: A New Perspective for Retrieval-Augmented Generation. arXiv, 2025.

A Gold-Label Utility Benchmark

Utility is **LLM-specific** — labeled per model by measured answer gain.

HOW WE LABEL UTILITY

$$u(d) = \mathbb{1}[\, \mathrm{Acc}(L(q,d)) > \mathrm{Acc}(L(q,\emptyset)) \,]$$

$$\text{gold set } G = \{\, d \in C : u(d) = 1 \,\}$$

A passage is **gold-utilitarian** for a model iff it lifts that model's answer over no-evidence.

HOW IT'S BUILT

4 LLMs · two families

Llama3.1-8B · Qwen3-8B / 14B / 32B

3 factoid QA datasets

Natural Questions · TriviaQA · MS MARCO-FQA

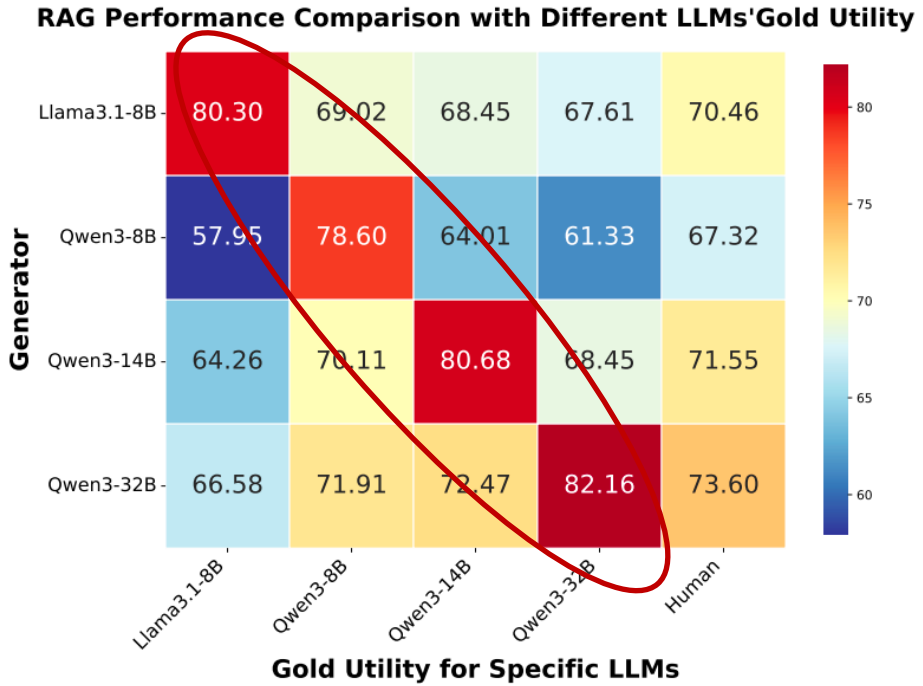
3 candidate pools / query

Top-20 Retrieved · Human Gold · Merged

Zhang, Bi, Guo, Zhang, Wang, Yin, Cheng. LLM-Specific Utility for Retrieval-Augmented Generation. arXiv:2510.11358 (2026).

LLM-Specific Utility: The Evidence

Golden utility **does not transfer** — each LLM has its own.



1st **The model's own gold passages**

Wins its row by 7–20 points.

2nd **Human-annotated positives**

Useful to a human is not always usable by the model. Relatively stable

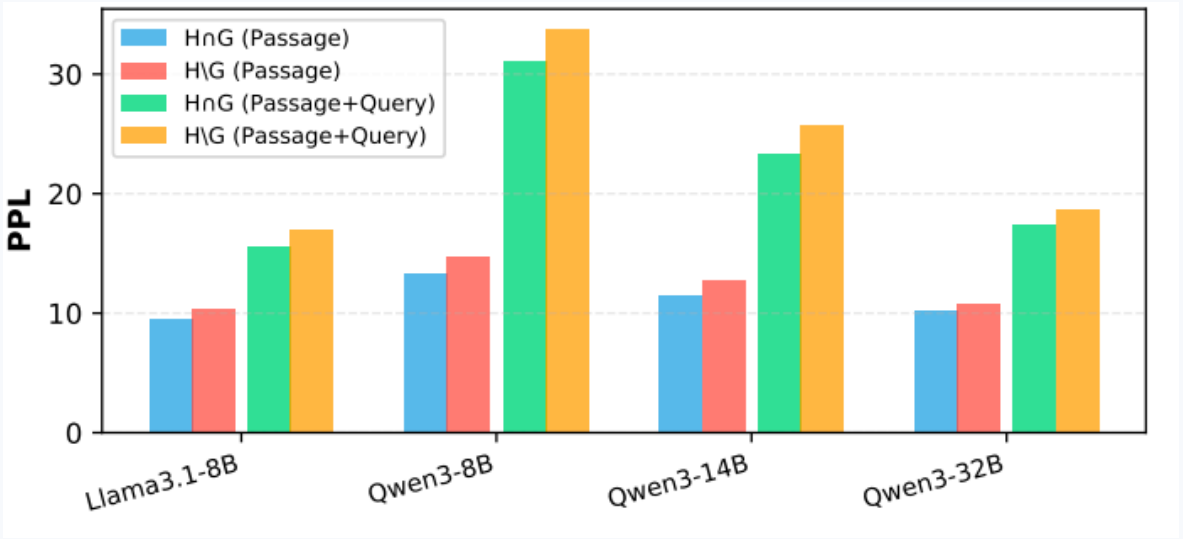
3rd **Passages picked for other LLMs**

Worst — especially across families.

Zhang et al. LLM-Specific Utility: A New Perspective for Retrieval-Augmented Generation. arXiv, 2025.

Human-Useful ≠ Model-Usable

Even among **human positives**, the LLM's gold passages stand apart.



READING THE PPL

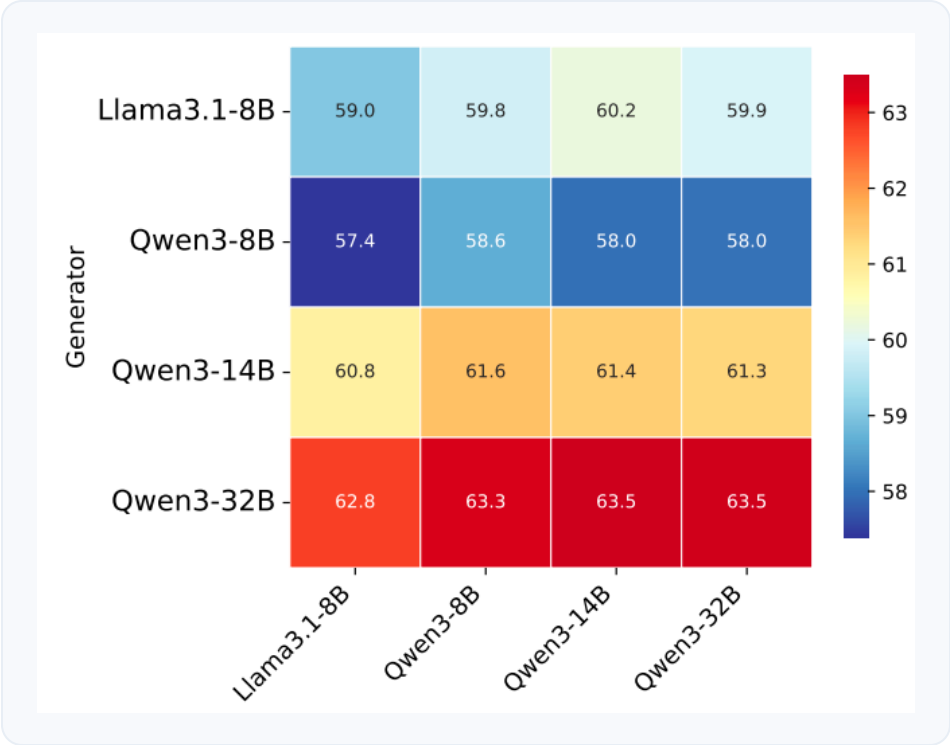
The LLM's own gold passages have **lower perplexity** than other human positives — the model reads them more easily.

> A passage humans find useful is **not always usable by the LLM**.
> PPL of gold passages < PPL of other positives.

Zhang et al. LLM-Specific Utility: A New Perspective for Retrieval-Augmented Generation. arXiv, 2025.

Judging LLM-Specific Utility

Existing judgment methods are **barely LLM-specific**.



Utility judged by different LLMs looks alike — columns barely differ.

WHAT THIS TELLS US

LLMs cannot reliably select their own **gold-utility passages**.

Passages picked by stronger models score better on every generator.

AN OPEN PROBLEM

Eliciting truly model-specific signals.

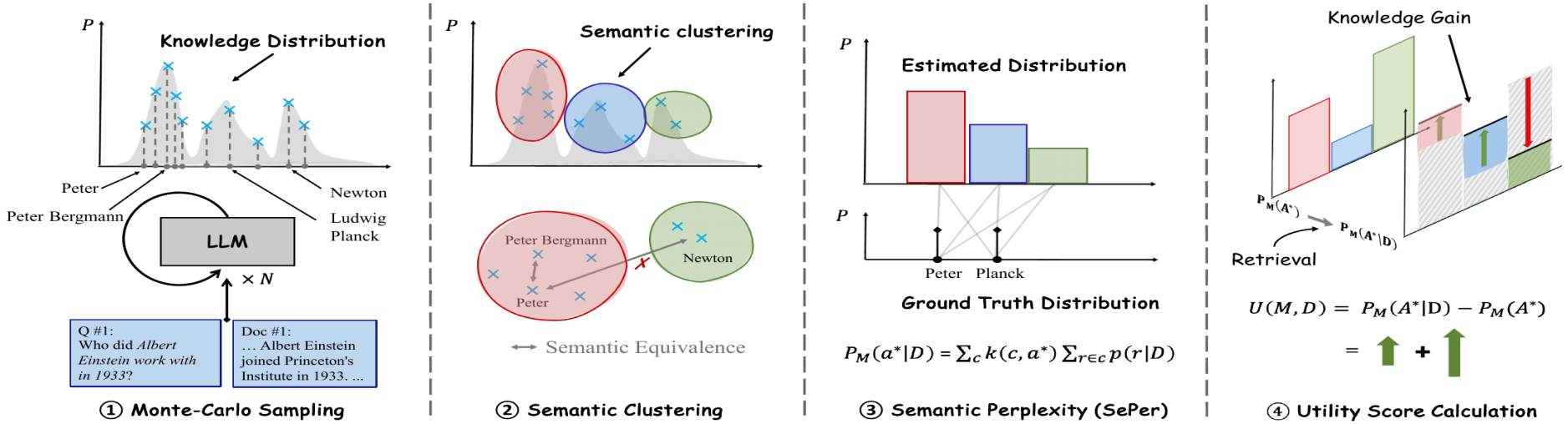
Zhang et al. LLM-Specific Utility: A New Perspective for Retrieval-Augmented Generation. arXiv, 2025.

SePer: Utility as Belief Revision

Semantic Perplexity Reduction (sample meanings, not tokens) of ground-truth answers after using d.

$$U(M, d) = \Delta\text{SePer} = P_M(a^* \mid q, d) - P_M(a^* \mid q)$$

belief after retrieval minus the model-specific baseline

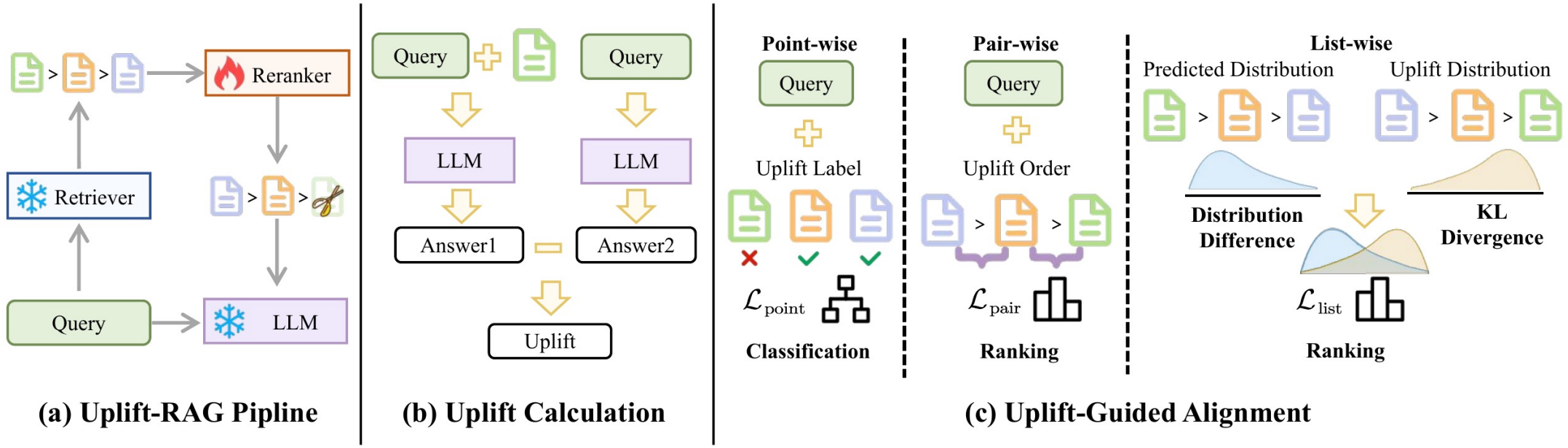


Monte-Carlo samples → semantic clusters → ground-truth belief → retrieval-induced knowledge gain.

Dai et al. SePer: Measure Retrieval Utility Through the Lens of Semantic Perplexity Reduction. ICLR, 2025.

Uplift-RAG: Align the Reranker to Uplift

Compute uplift labels. **Teach a lightweight reranker to predict them.**



POINT-WISE
classify: helpful or not

PAIR-WISE
order: which document helps more

LIST-WISE
match the full uplift distribution

Qu et al. Uplift-RAG: Uplift-Driven Knowledge Preference Alignment for Retrieval-Augmented Generation. Findings of EMNLP, 2025.

Signal Source vs. Target

Where the utility signal **comes from** decides who it serves.

Signal source	Utility type	Deployed to
Human relevance labels	**LLM-agnostic**	general retriever
LLM utility judgments by a strong LLM utility-focused annotation	**LLM-agnostic**	general retriever
Downstream performance of a strong LLM REPLUG	**LLM-agnostic**	same-family LLM
Per-model answer gain gold labels · SePer · uplift	**LLM-specific**	the target LLM

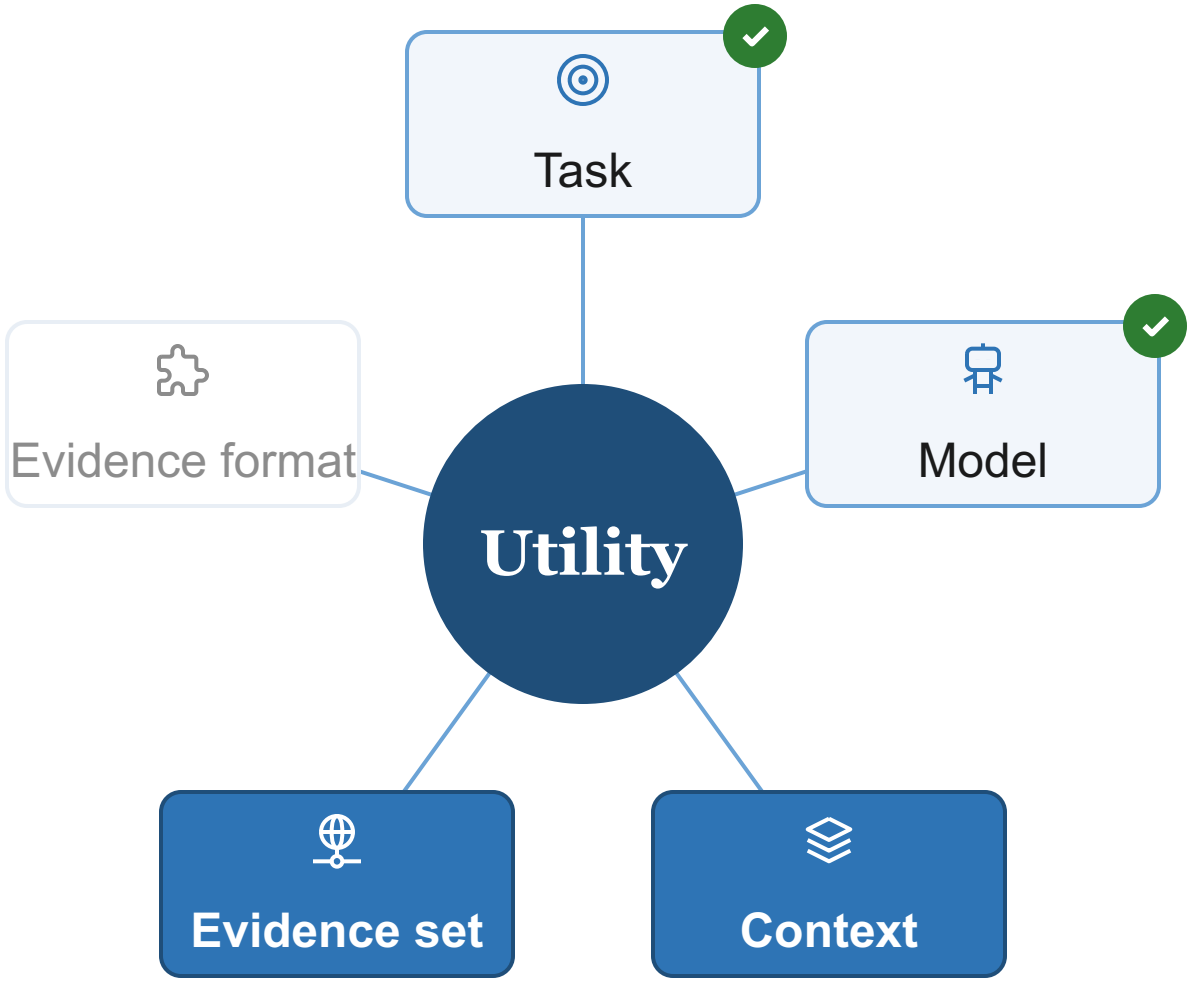
Top two rows: **scalable, reusable** supervision. Bottom row: **aligned to one model**.

Model Dimension: Takeaway

> **Agnostic** methods scale utility supervision;
> **specific** methods align with the model that will actually consume the evidence.

Choose by what you are training: a general retriever — or a component serving one known model.

2.4 · Context & Evidence Set

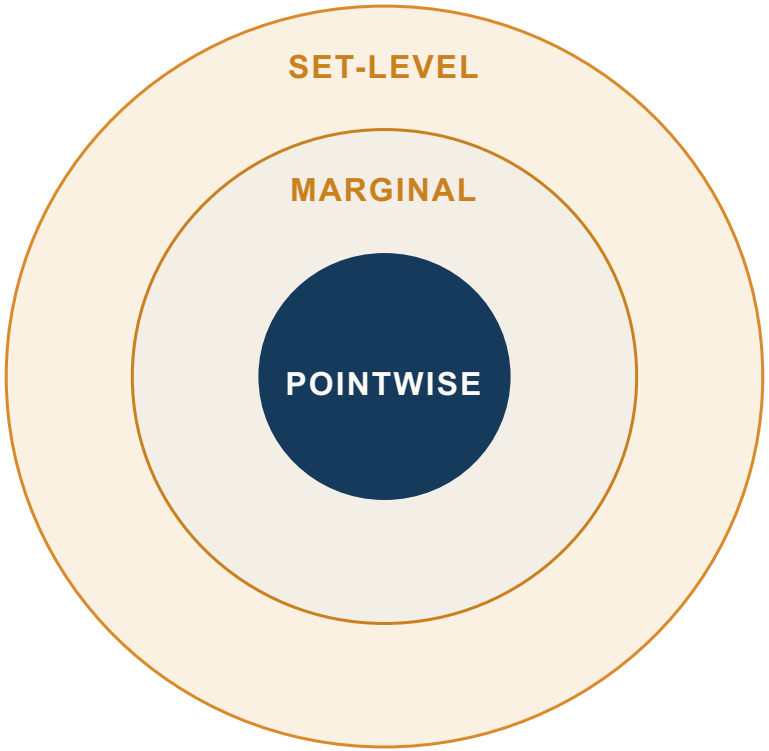


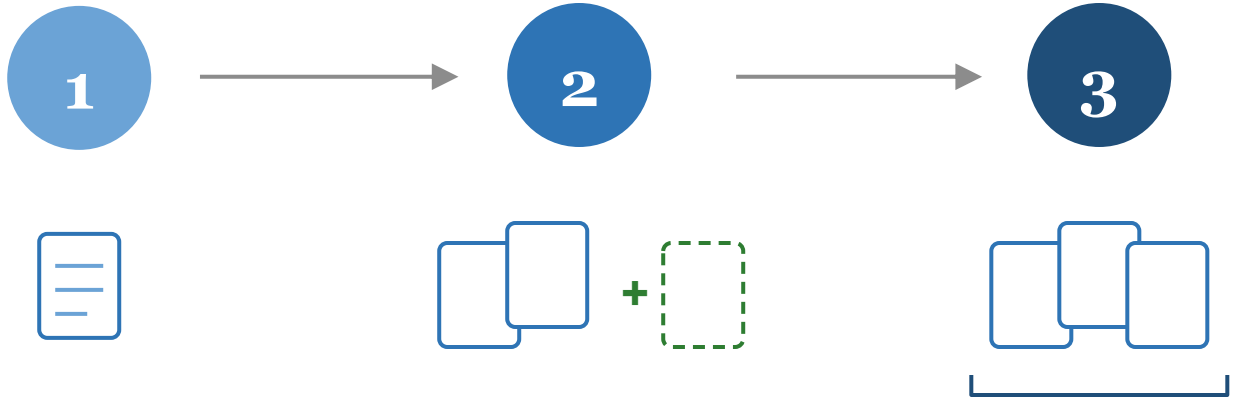
THIRD DIMENSION

A document is **not judged alone** — it is judged given what is already there.

Three lenses: pointwise, marginal, and set-level utility.

Three Granularities of Utility





Pointwise

Is this document useful by itself?

Marginal

How much does adding this document help?

Set-level

Does the set jointly support the output?

Pointwise Utility

One document, judged **by itself**.

Strengths

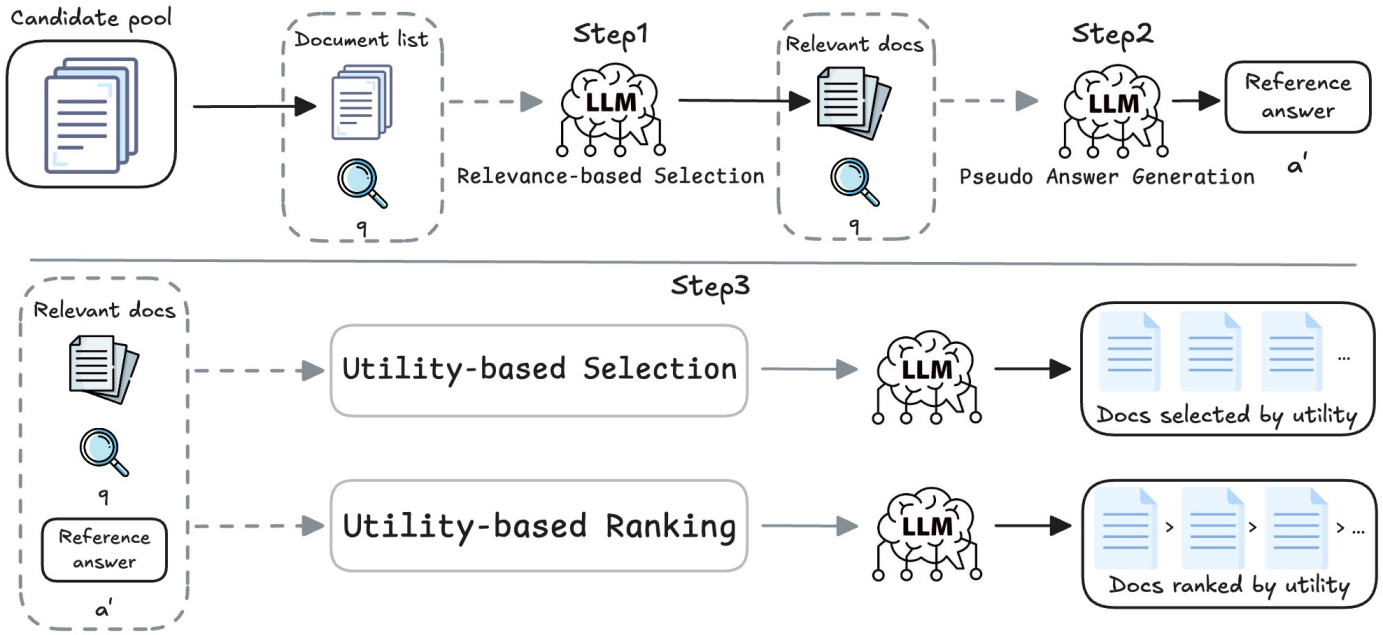
- simple to label
- easy to train and deploy
- fits one-at-a-time scoring
  selectors and rerankers

Blind to

- redundancy
- complementarity
- conflict between documents
- multi-hop coverage

Pointwise Labeling

Select/Rank documents based on their **utility.**



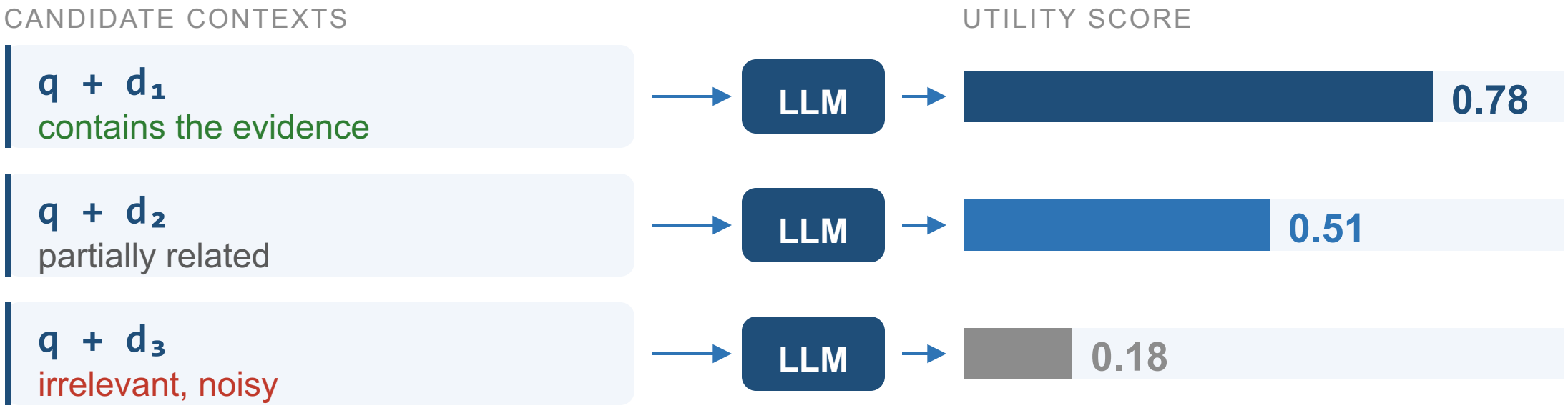
Per-document Utility Labels

Selected -> pos; otherwise -> neg
High-ranked -> pos; otherwise -> neg

Zhang et al. Utility-Focused LLM Annotation for Retrieval and Retrieval-Augmented Generation. EMNLP, 2025.

Pointwise Scoring

Downstream performance: how much does d help produce the answer a?

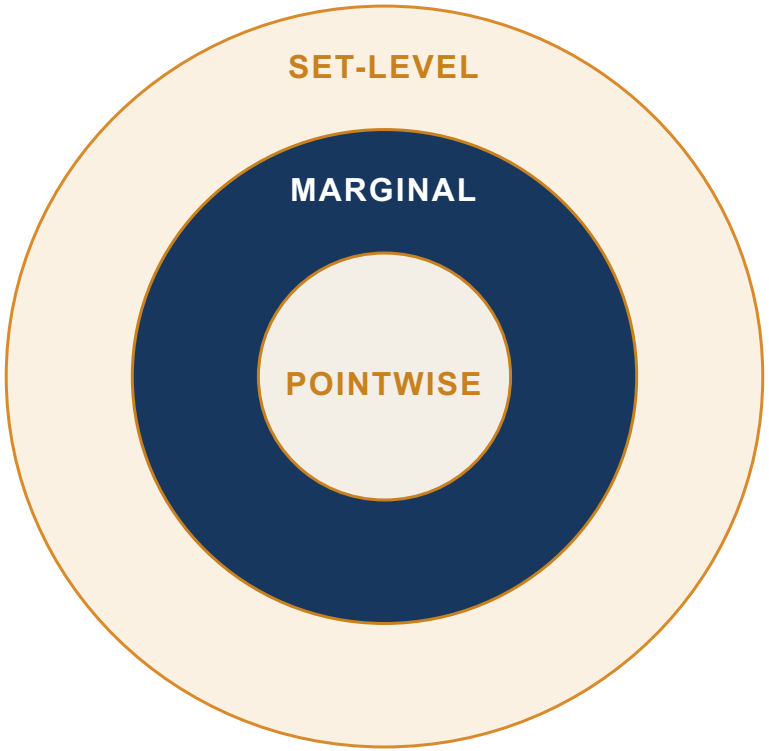


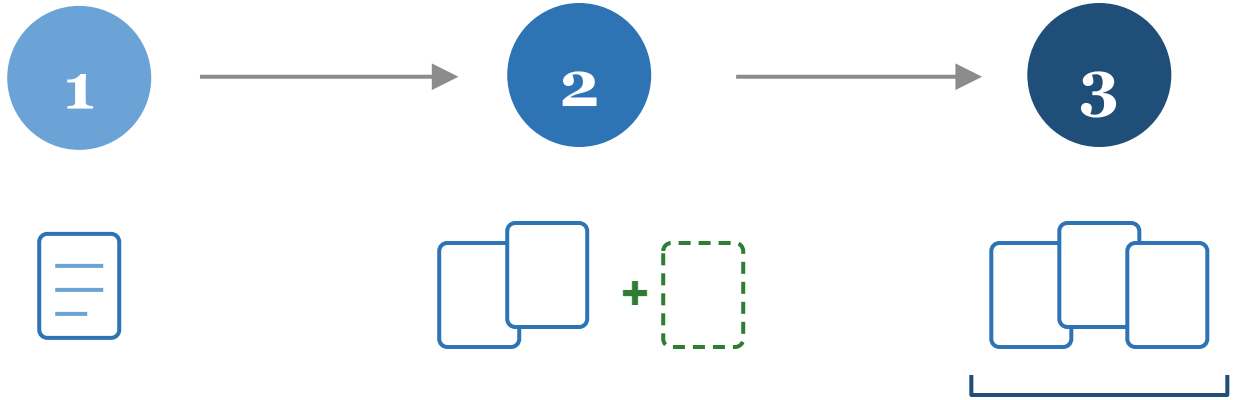
Practical signal: the model's **likelihood of the gold answer**

$U(d) \propto \log P(a \mid q, d)$ — REPLUG distills it into the retriever; Atlas ranks documents with it.

Shi et al. REPLUG: Retrieval-Augmented Black-Box Language Models. NAACL, 2024.
Izacard et al. Atlas: Few-shot Learning with Retrieval Augmented Language Models. JMLR, 2023.

Three Granularities of Utility





Pointwise

Is this document useful by itself?

Marginal

How much does adding this document help?

Set-level

Does the set jointly support the output?

Marginal Utility: James Webb

SECTION 1, REVISITED

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

P1 *"launched December 25, 2021, on an Ariane 5 from Kourou, French Guiana."*

≈

P2 *"lifted off on Christmas Day 2021, aboard Ariane 5 from Kourou..."*

same facts, different words

P2 is relevant — but once P1 is selected it adds nothing: **marginal utility ≈ 0**.

MARGINAL GAIN, GIVEN THE CURRENT SET S

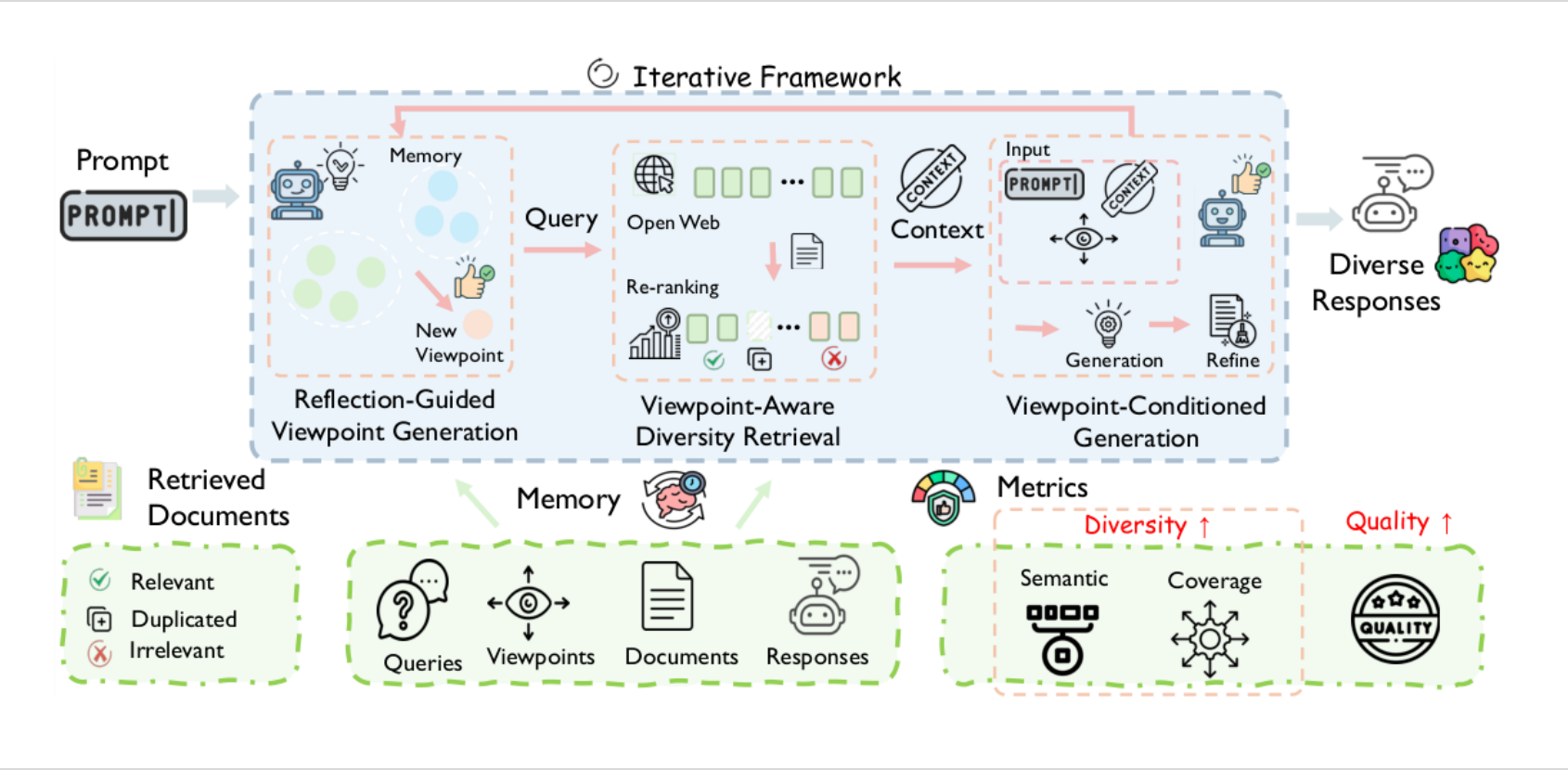
$$u(d \mid S, q, M) = \mathrm{Perf}(M, S \cup \{d\}) - \mathrm{Perf}(M, S)$$

WHY MARGINAL? EXACT SET OPTIMIZATION IS COMBINATORIAL — SO BUILD GREEDILY

score marginal gain → add the best document → repeat until budget

A document can be **pointwise useful** yet **marginally redundant**.

DIVERGE: Diversity Across Viewpoints



Within batch

Avoid redundancy with documents already in S_t

Across rounds

Global memory M<t penalizes evidence seen under earlier viewpoints

It iteratively explores different viewpoints and using an **extended MMR objective** to select documents that are both relevant and novel relative to current and previously retrieved evidence.

Hu et al. DIVERGE: Diversity-Enhanced Retrieval-Augmented Generation for Open-Ended Information Seeking. arXiv, 2026.

DIVERGE: Extended MMR Selection

Extended MMR objective

$$s_t(d) = \alpha \,\mathrm{Rel}(d, q_t) - \beta \max_{h \in M_{<t}} \mathrm{Sim}(d, h) - (1 - \alpha) \max_{s \in S_t} \mathrm{Sim}(d, s)$$

RELEVANCE **CROSS-ROUND diversity** **INTRA-BATCH diversity**

≈2× diversity with **under 2% average quality loss**

Best overall diversity–quality trade-off among the compared open-ended RAG systems.

Hu et al. DIVERGE: Diversity-Enhanced Retrieval-Augmented Generation for Open-Ended Information Seeking. arXiv, 2026.

Contextual Utility in Multihop QA

Question:
What nationality was James Henry Miller's wife?

Supporting Passages

Passage 1:

Ewan MacColl - James Henry Miller (25 January 1915 – 22 October 1989), better known by his stage name Ewan MacColl, was an English folk singer, songwriter, communist, labour activist, actor, poet, playwright and record producer.

Passage 2:

Peggy Seeger - Margaret "Peggy" Seeger (born June 17, 1935) is an **American** folksinger. She is also well known in Britain, where she has lived for more than 30 years, and was married to the singer and songwriter Ewan MacColl until his death in 1989.

Utility Scores

Independent Assessment

Passage	Score
Passage 1	4
Passage 2	1

Conditioned on Passage 1

Passage	Score
Passage 1	5
Passage 2	4

Passage 2 independently provides little utility for answering the question. However, when conditioned on Passage 1 (which establishes that James Henry Miller used the stage name Ewan MacColl), Passage 2 becomes highly useful as it identifies Peggy Seeger as Ewan MacColl's American wife.

WHY ORDER CHANGES UTILITY

Passage 1
introduces a bridge entity

↓

Passage 2
resolves the bridge into the answer

↓

Complete reasoning chain
the same passage can change role by position

Jain and Garimella. Modeling Contextual Passage Utility for Multihop Question Answering. IJCNLP-AACL, 2025.

Contextual Utility: Learned Scorer

$$U(p_i \mid P_{<i}, q) \in [1, 5], \qquad \mathcal{L} = \frac{1}{N} \sum_{j=1}^{N} \left(f_{\text{RoBERTa}}(p_{i,j}, q_j) - U_{i,j} \right)^2$$

1 · Reason

GPT-o1 produces a detailed trace and cites support passages

2 · Label

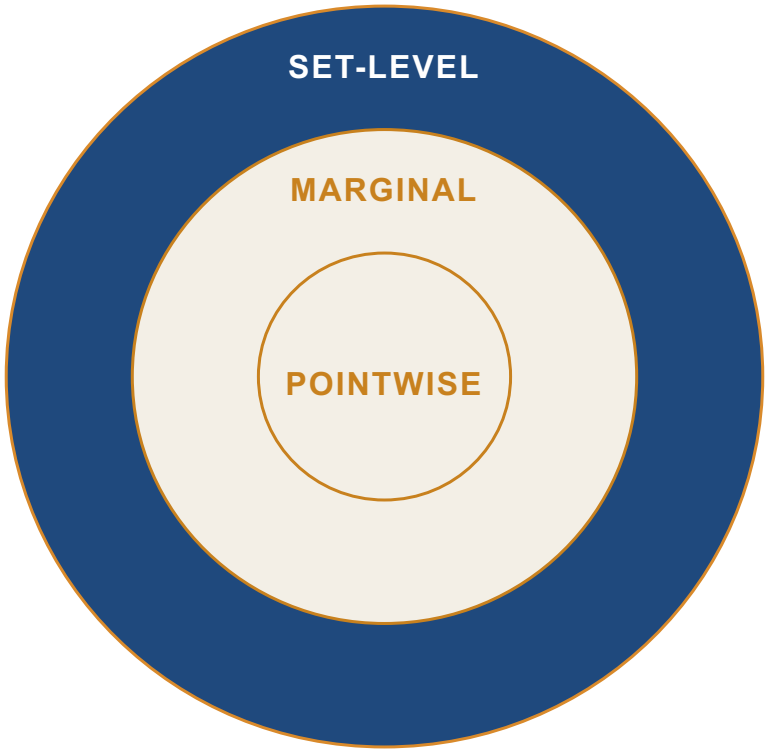
GPT-4o assigns utility 1–5 from role and criticality

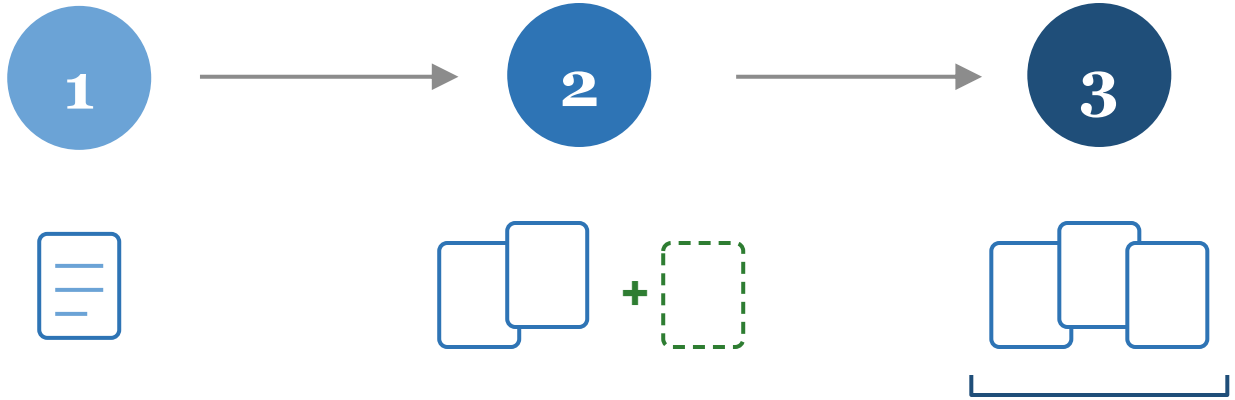
3 · Distill

RoBERTa-large regression learns a deployable scorer

Jain and Garimella. Modeling Contextual Passage Utility for Multihop Question Answering. IJCNLP-AACL, 2025.

Three Granularities of Utility





Pointwise

Is this document useful by itself?

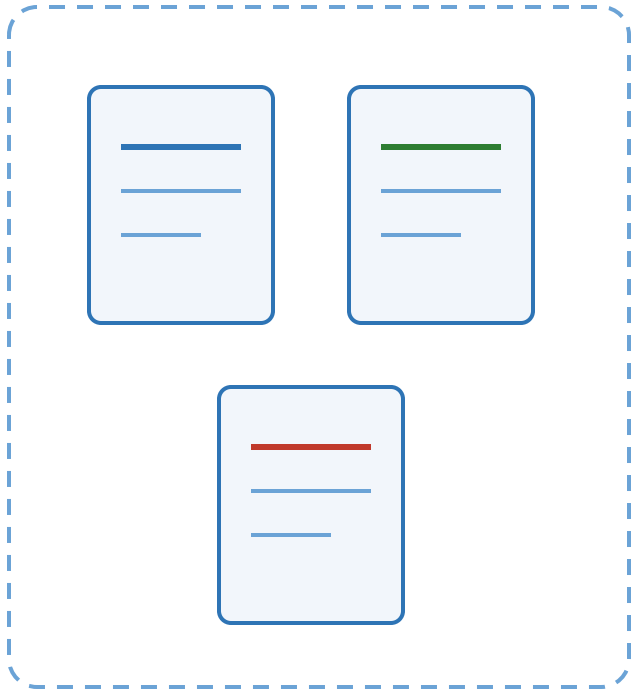
Marginal

How much does adding this document help?

Set-level

Does the set jointly support the output?

What Makes a Good Evidence Set



the selected set, judged jointly

Coverage

all necessary aspects or reasoning hops are present

Complementarity

each document adds different useful information

Low redundancy

repeated evidence must not crowd out missing facts

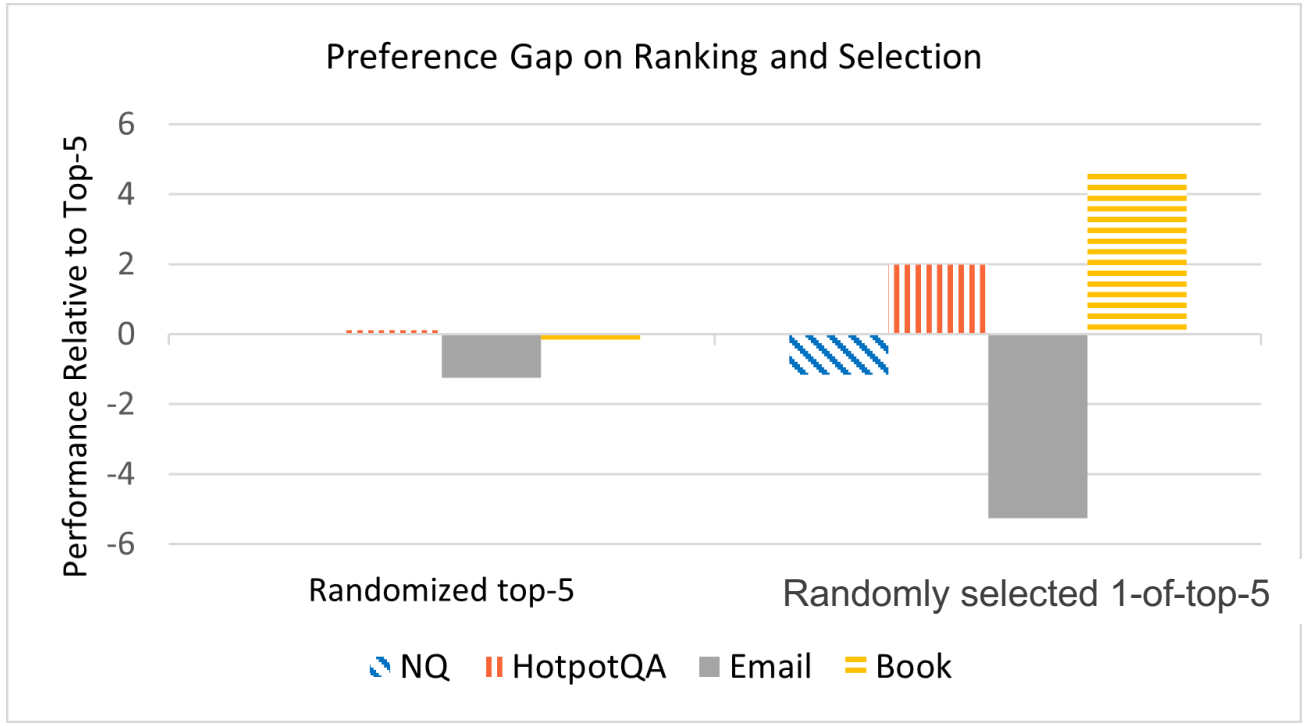
Conflict management

contradictions noticed and resolved — recall the Nine Planets page

Provenance

evidence is traceable and reliable

BGM: Bridging the Preference Gap



Effect of changing selection vs changing ranking on a frozen LLM

Ranking barely matters

Randomizing the order of the Top-5 has little impact. Humans read top→bottom; LLMs do not share that preference.

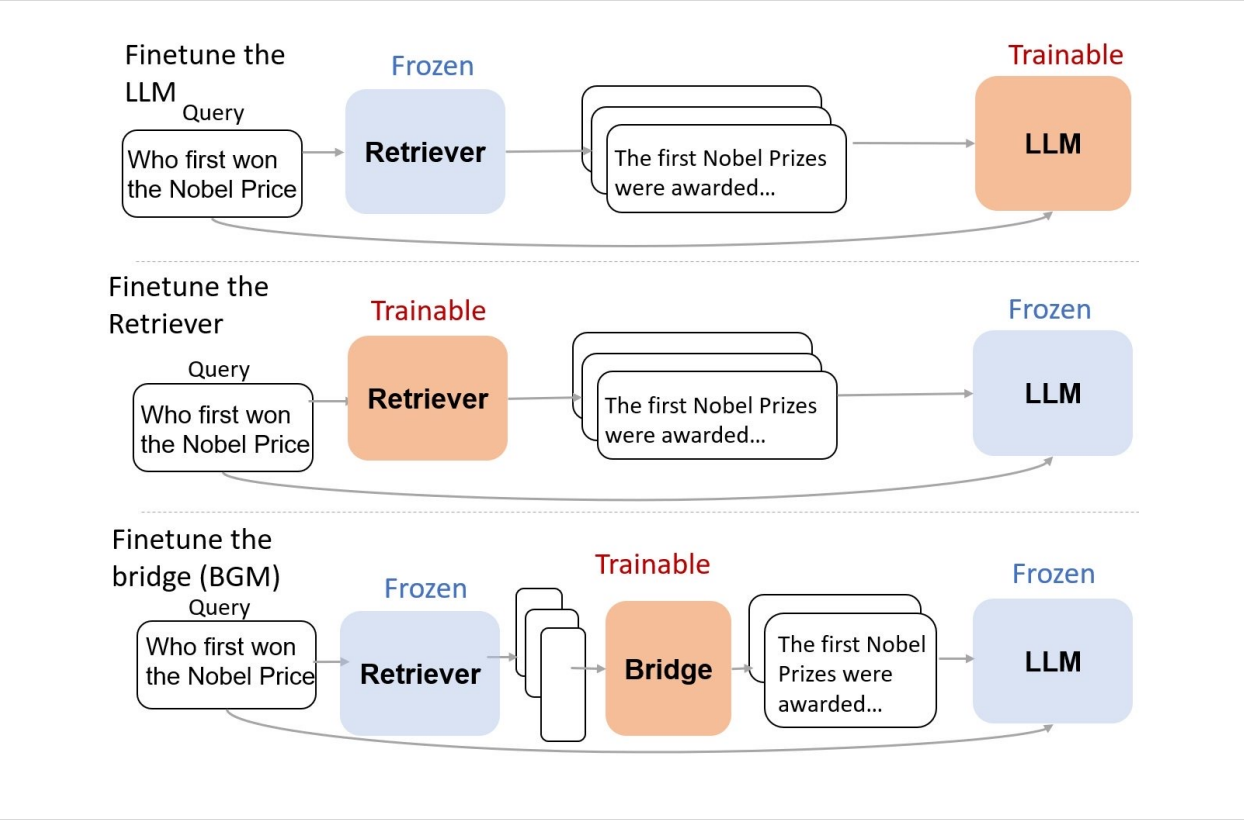
Selection does

Only feeding 1-of-top-5 passage shifts LLM performance by **more than 5%** — information selection is important

Preference gap: retrievers optimize human-friendly ranking; LLMs need an LLM-friendly *selection*.

Ke et al. Bridging the Preference Gap between Retrievers and LLMs. ACL, 2024.

BGM: A Bridge Trained by LLM Feedback



1 Silver sequences

Greedy search over retrieved passages builds target passage-ID sequences.

2 Supervised learning

A seq2seq bridge maps query + passages to the passage IDs to keep and reorder.

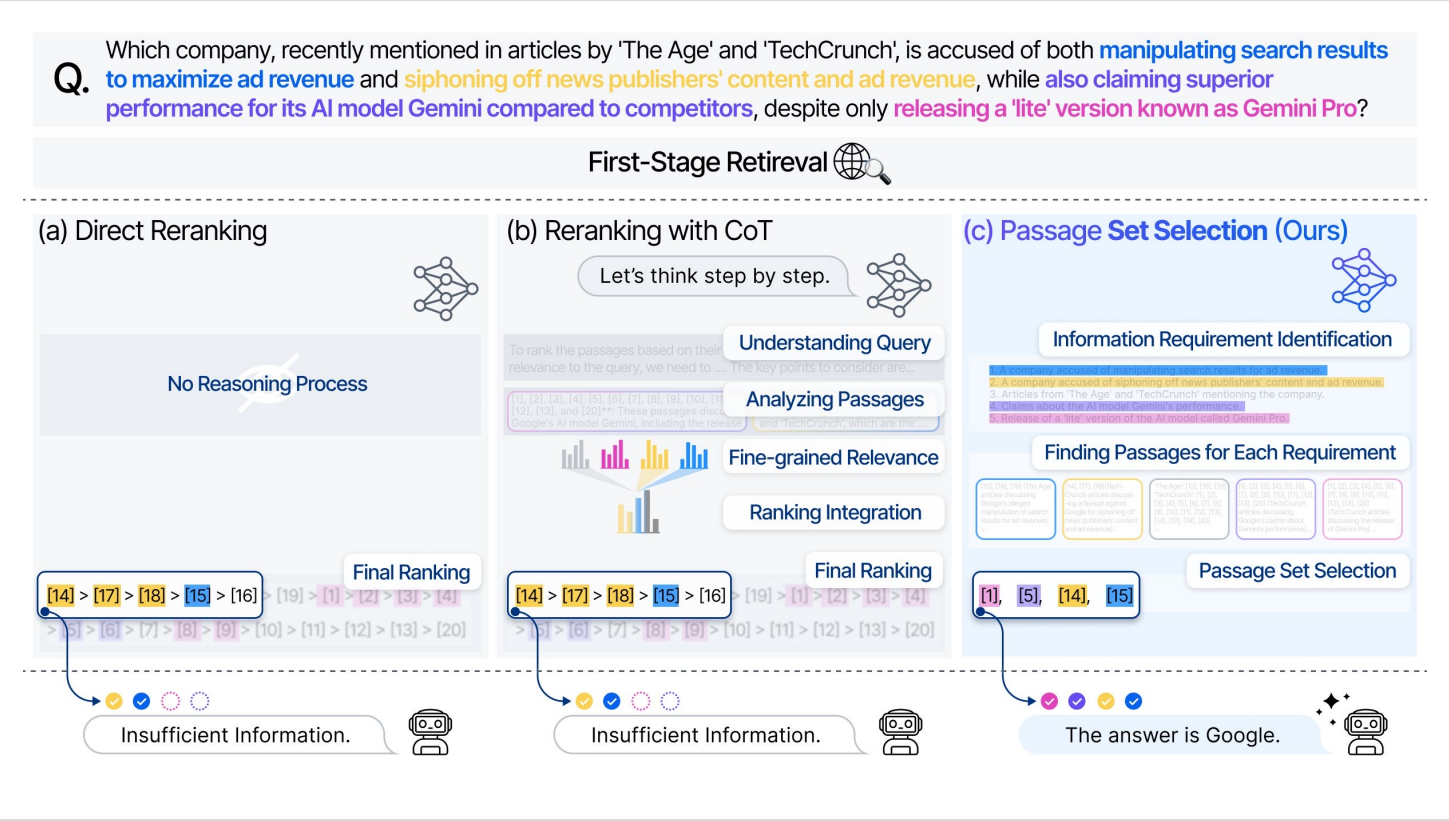
3 Reinforcement learning

Bridge = policy; reward = the frozen LLM's answer quality. Retriever & LLM stay fixed.

The sequence-to-sequence bridge selects and reorders an LLM-friendly subset

Ke et al. Bridging the Preference Gap between Retrievers and LLMs. ACL, 2024.

SetR: From Reranking to Set Selection



1 · Identify needs (IRI)

Chain-of-Thought makes the query's information requirements explicit — every aspect a complete answer must cover.

2 · Select the set

Map each information requirement to supporting candidate passages. Select an unordered subset that **jointly** provides comprehensive, diverse, and concise coverage.

GPT-4o reasoning and selection, distilled to a Llama-3.1-8B selector; fewer input tokens, better results

Lee et al. Shifting from Ranking to Set Selection for Retrieval Augmented Generation. ACL, 2025.

Vendi-RAG: Adaptive Trade-Off

SET-LEVEL DIVERSITY

$$\mathrm{VS}_k(D) = \exp\left(-\sum_{i=1}^{n} \lambda_i \log \lambda_i\right), \qquad \lambda_i = \mathrm{eig}_i(K/n)$$

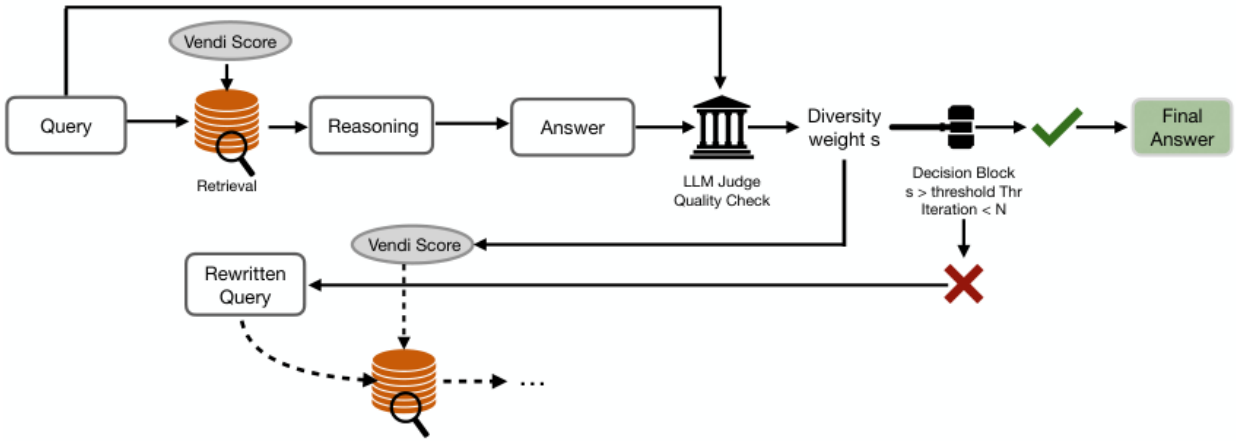
Vendi Score measures diversity from the eigenvalue distribution of the set's similarity matrix.

Judge says quality is low

increase diversity weight s

Judge says quality is high

shift weight back toward relevance



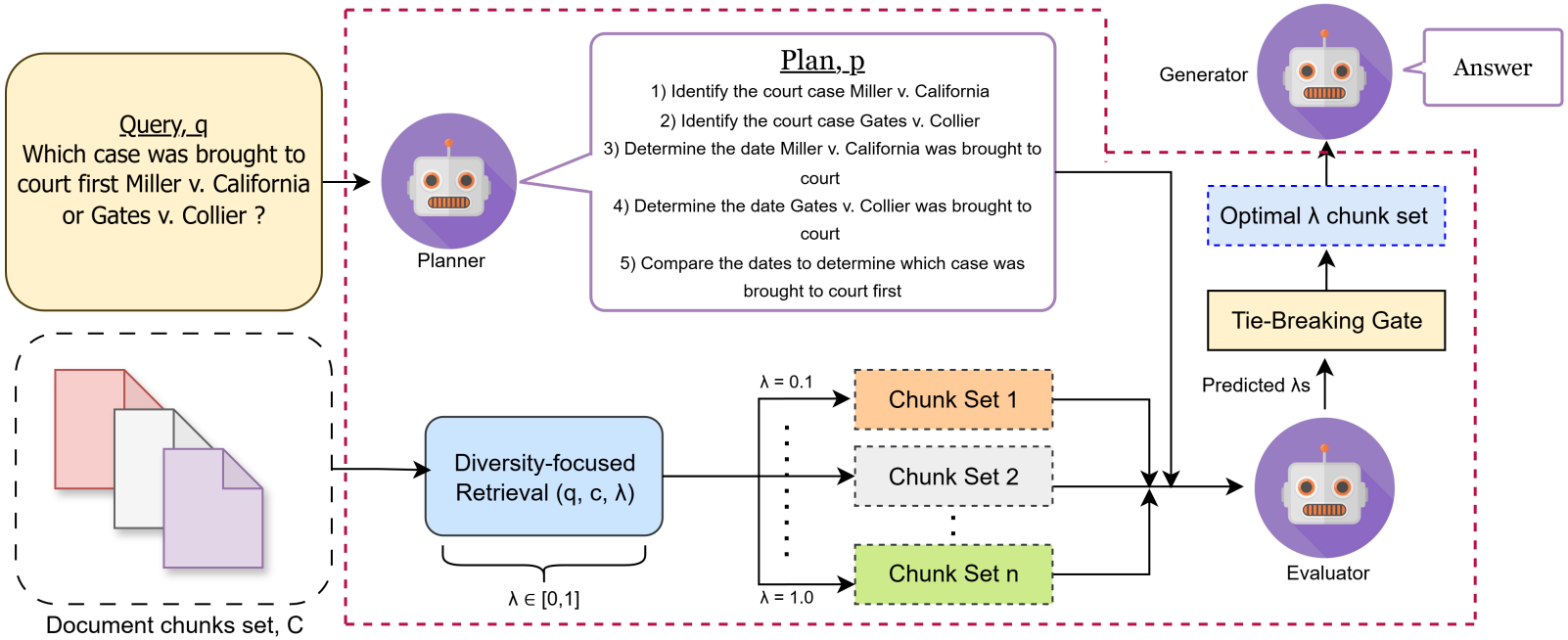
Utility: s · Vendi Score + (1−s) · relevance

Repeat greedy selection until the judge-guided weight converges.

Useful diversity changes with the **candidate pool, document budget, and answer quality**.

Rezaei and Dieng. Vendi-RAG: Adaptively Trading-Off Diversity and Quality Significantly Improves Retrieval Augmented Generation with LLMs. arXiv, 2025.

DF-RAG: Query-Aware Diversity



$$\mathrm{gMMR}(c) = \lambda \cos(\vec{q}, \vec{c}) + (1-\lambda)\sqrt{2 - 2\cos(\vec{c}, \vec{c_S})}$$

c_S is the centroid of the already selected chunks.

QUERY-AWARE CONTROL

λ → 1: relevance · **λ → 0:** diversity

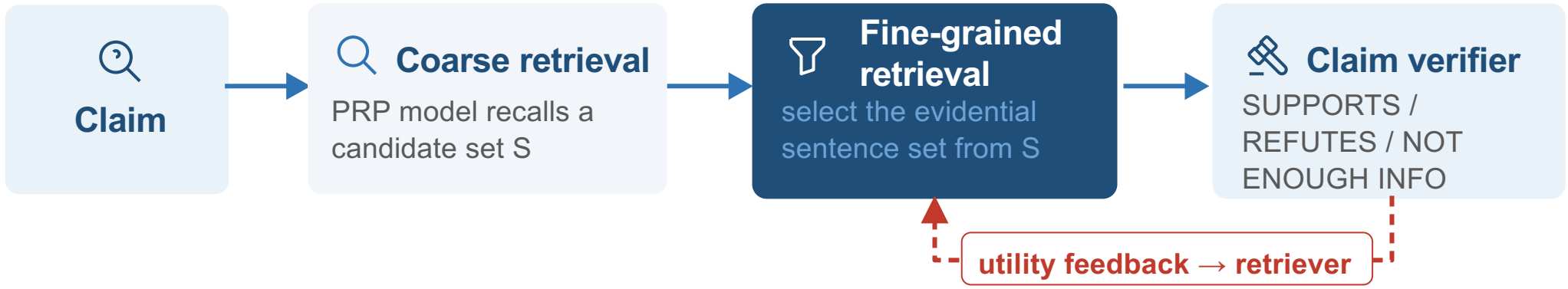
Evaluator selects among multiple λ-conditioned sets.

Training-free framework that dynamically selects the right level of diversity for different queries

Khan et al. DF-RAG: Query-Aware Diversity for Retrieval-Augmented Generation. Findings of EACL, 2026.

FER: From Relevance to Utility

Optimize the evidence set by how much utility the verifier gains — the gap to gold evidence.



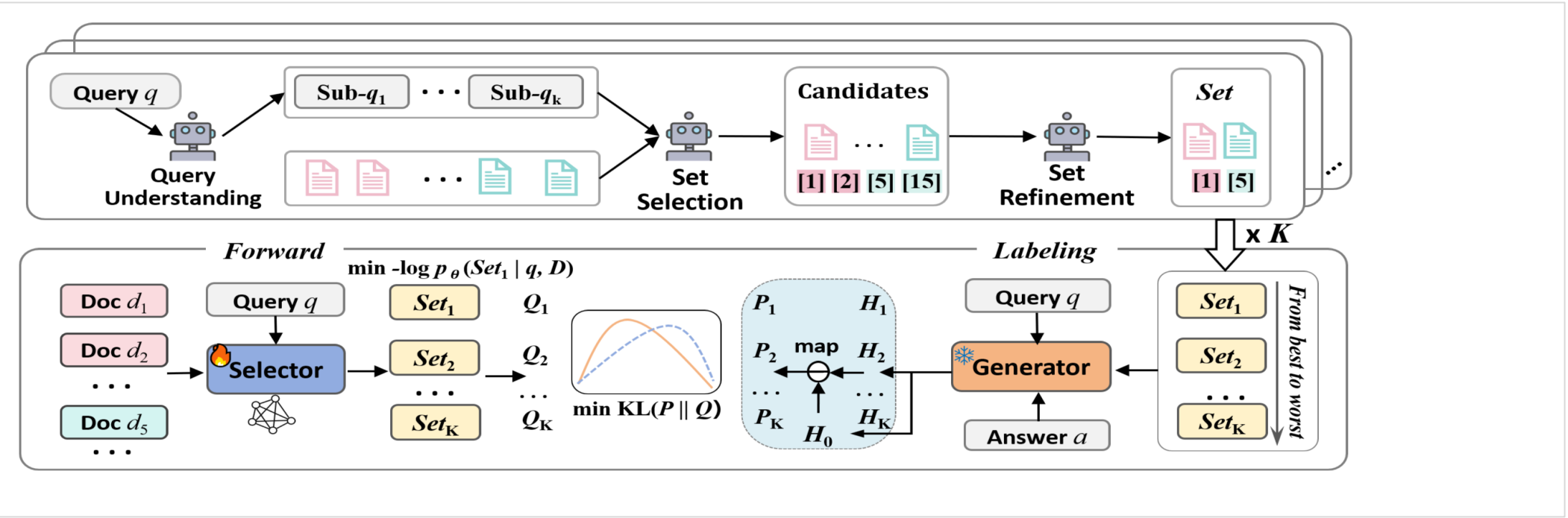
A CLS token before each sentence; use gumbel-softmax to learn which sentence should be kept

FEEDBACK SIGNAL · UTILITY DIVERGENCE

$$\mathcal{L}_{uti} = \underbrace{y^* D_\phi(c, E^*)}_{\text{utility of gold evidence}} - \underbrace{y^* D_\phi\big(c, R_\theta(c, S)\big)}_{\text{utility of retrieved set}}$$

Zhang et al. From Relevance to Utility: Evidence Retrieval with Feedback for Fact Verification. Findings of EMNLP, 2023.

OptiSet: Expand-then-Refine



Set-wise retrieval: Selects passages jointly, rather than ranking them independently.
Expand–refine: Retrieves diverse evidence, then removes redundancy. Run multiple times with stochastic sampling, producing alternative sets.

Generator-based utility: Scores each set by its reduction in answer perplexity.
Set-listwise training: Combines imitation of the best sampled set with listwise preference learning over multiple candidate sets.

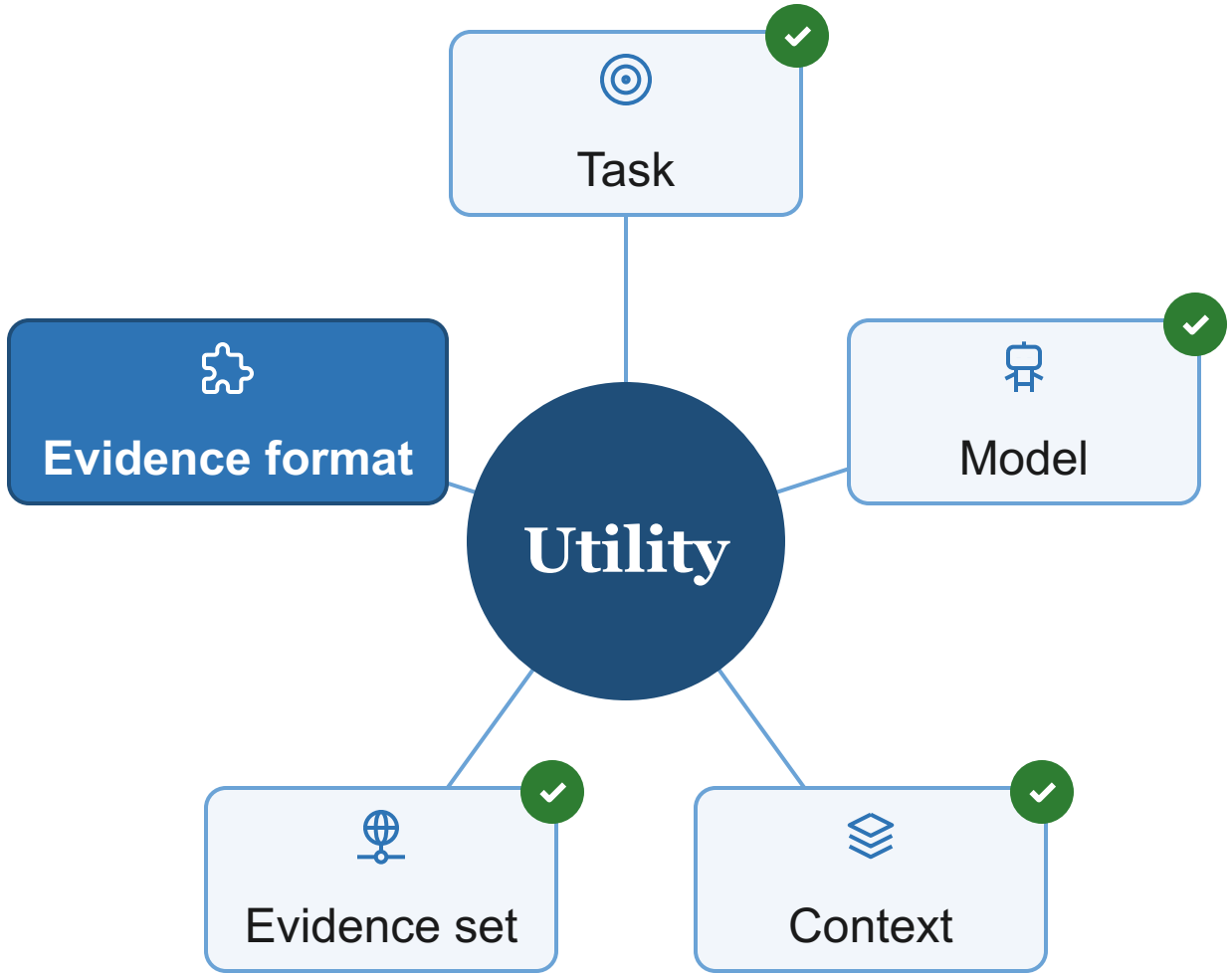
Jiang et al. OptiSet: Unified Optimizing Set Selection and Ranking for Retrieval-Augmented Generation. arXiv, 2026.

Context & Set: Recap

Granularity	Question asked	Limitation
Pointwise	Is this document useful by itself?	blind to interactions
Marginal	Is this worth adding, given the current set?	greedy misses global optima
Set-level	Does the set jointly support the output?	exact optimization is hard

The unit of utility is shifting — from **the document** to **the context it joins**.

2.5 · Evidence Format



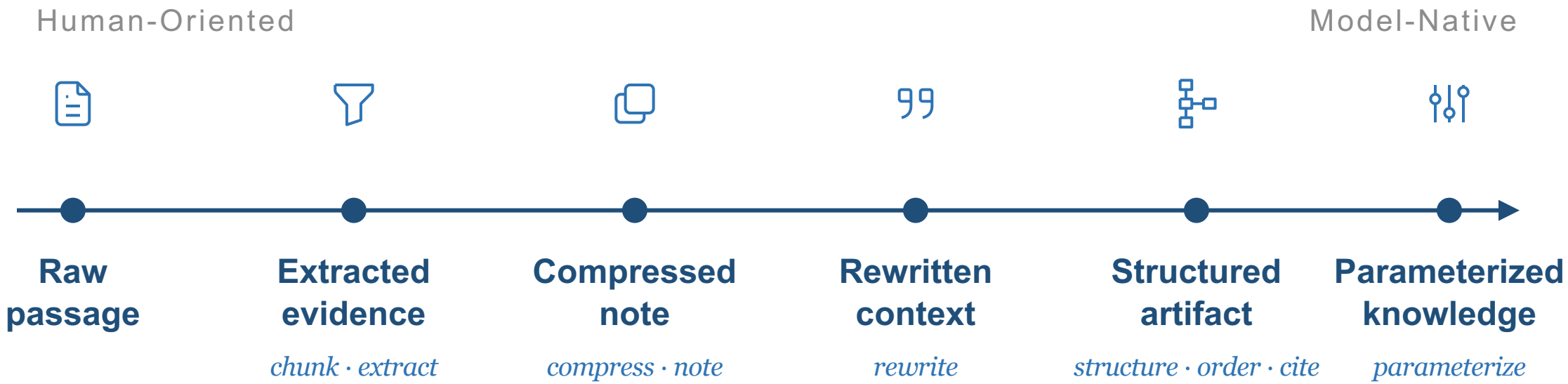
LAST DIMENSION

Utility also depends on **how** evidence reaches the model — not only **which** document.

The same fact, in a different form, has different utility.

The Format Spectrum

Evidence can be **transformed** — from raw text to model-native forms.



R2U: Rewriting for Utility

RELEVANCE → UTILITY

Trains a rewriter that transforms retrieved documents into versions more useful for the generator.

1 · Rewrite

A rewriter transforms each retrieved passage into a more answer-oriented version

2 · Score

A fixed downstream generator computes $\Delta\ell = \ell(d') - \ell(d)$ on the gold answer

3 · Train

train the rewriter via SFT or DPO on rewrite pairs

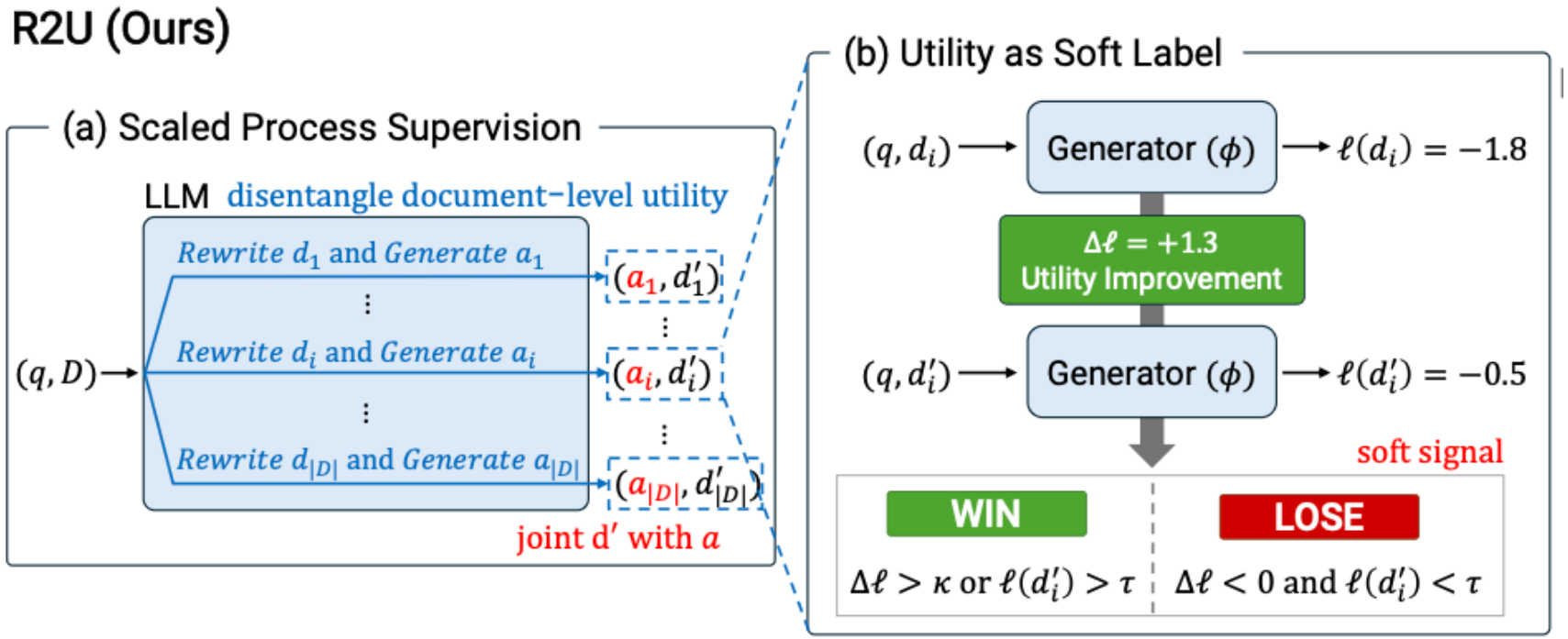
Relevance-trained rewriters keep fluent but useless spans.

Utility-trained rewriters push every edit **toward the correct answer**.

Kim et al. Relevance to Utility: Process-Supervised Rewrite for RAG. Findings of ACL, 2026.

R2U: Rewriting for Utility

Keep a rewrite only if **the gold answer becomes more likely.**



$\Delta\ell = \ell(d') - \ell(d)$ $\Delta\ell > 0 \Rightarrow$ win pair; else lose pair — used for preference optimization.

Kim et al. Relevance to Utility: Process-Supervised Rewrite for RAG. Findings of ACL, 2026.

Chain-of-Note

Standard RAG

Retrieve docs → Answer

One noisy passage can flip the answer.

Chain-of-Note

Retrieve docs → Reading notes → Answer

Each passage is assessed before answering.

Three Reading-Note Types

TYPE 1
Relevant, sufficient

The passage answers the question directly — read the answer out of it.

TYPE 2
Relevant, partial

On-topic but incomplete — add the model's own knowledge to the note.

TYPE 3
Irrelevant, unknown

No passage helps — use parametric knowledge or reply 'unknown'.

Yu et al. Chain-of-Note: Enhancing Robustness in Retrieval-Augmented Language Models. EMNLP, 2024.

Passage Injection

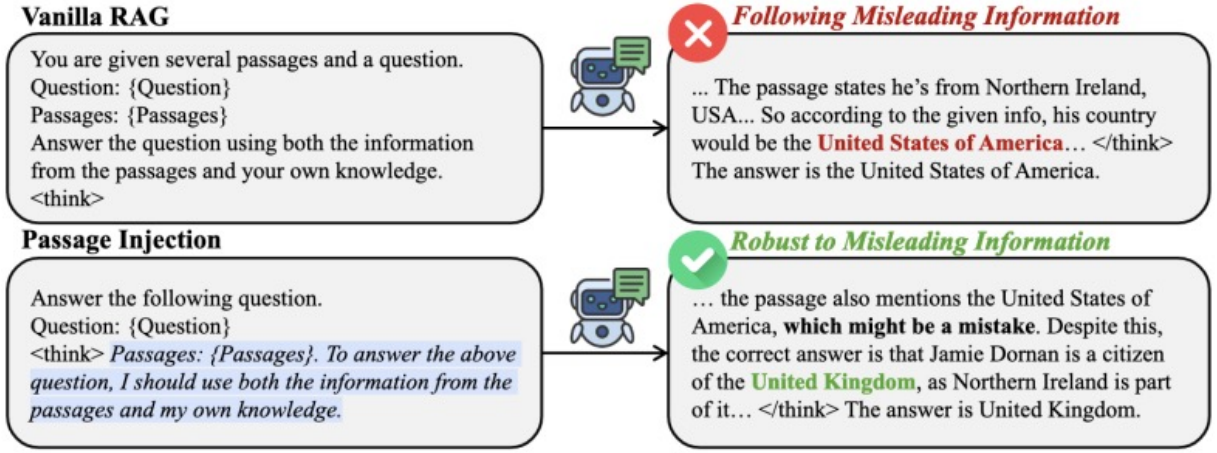
Put retrieved passages **inside the reasoning**, not the input prompt.

➢ **Vanilla RAG** — passages sit in the input prompt

Input([Instruction] + **[Passages]** + [Question]) → **<think>** … **</think>** → Response

➢ **Passage Injection** — passages sit inside the reasoning

Input([Question]) → **<think>** [Instruction] + **[Passages]** … **</think>** → Response



*The model can then **verify and correct the passages** before producing the final answer.*

Tang et al. Injecting External Knowledge into the Reasoning Process Enhances Retrieval-Augmented Generation. SIGIR-AP, 2025.

Agent-Native Artifacts

For AI agents, the most useful evidence **may not be a PDF**.

For human readers

PDF

prose, figures, layout — interpreted by people

For AI agents: a research package

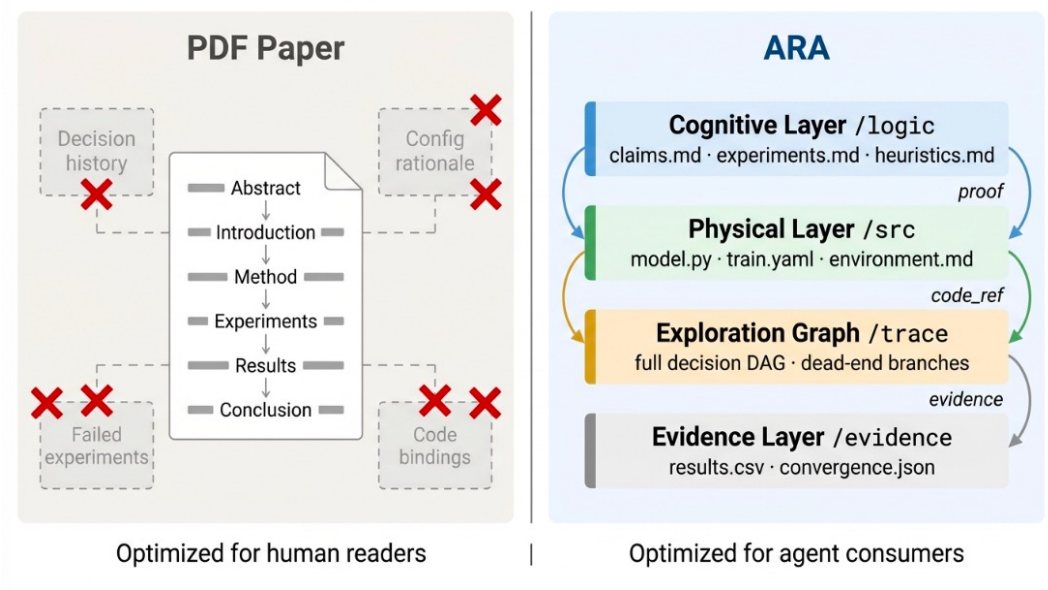
- scientific logic
- executable code
- exploration graphs
- evidence grounding

Future retrieval targets: structured, executable, directly inspectable.

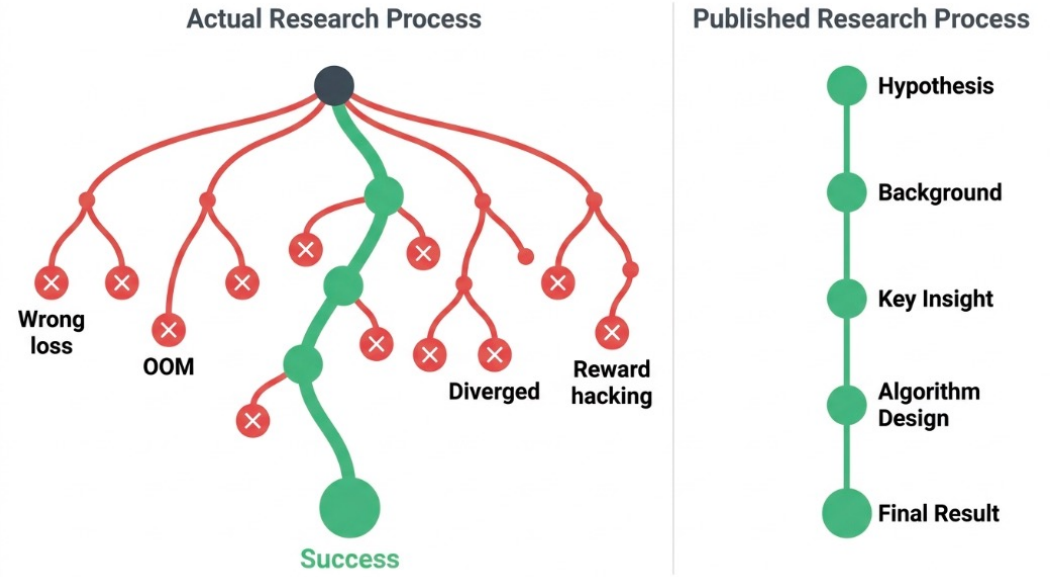
Liu et al. The Last Human-Written Paper: Agent-Native Research Artifacts. arXiv, 2026.

Inside an Agent-Native Artifact

Four layers replace the narrative — **logic, code, trace, evidence.**



The narrative PDF vs. the four-layer package.



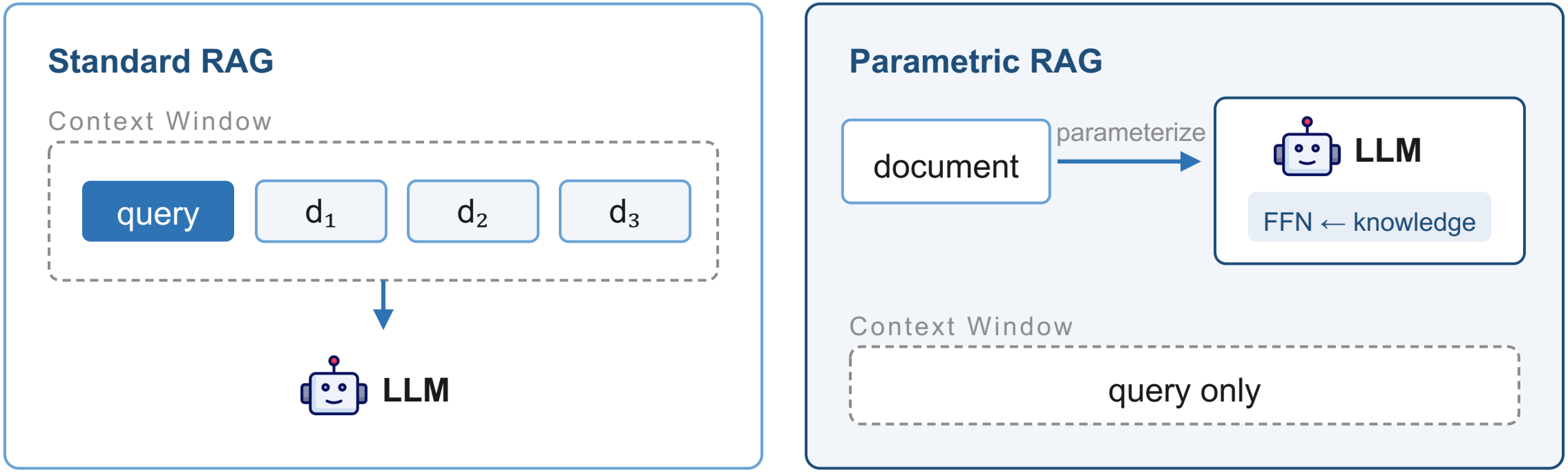
The exploration a published paper compiles away.

Agents load only the layer they need —
no prose parsing, and failed branches survive as first-class evidence.

Liu et al. The Last Human-Written Paper: Agent-Native Research Artifacts. arXiv, 2026.

Parametric RAG

Attention-based RAG may lack sufficient interactions between external evidence and internal knowledge.
-> injected into the model through the **parameters**.

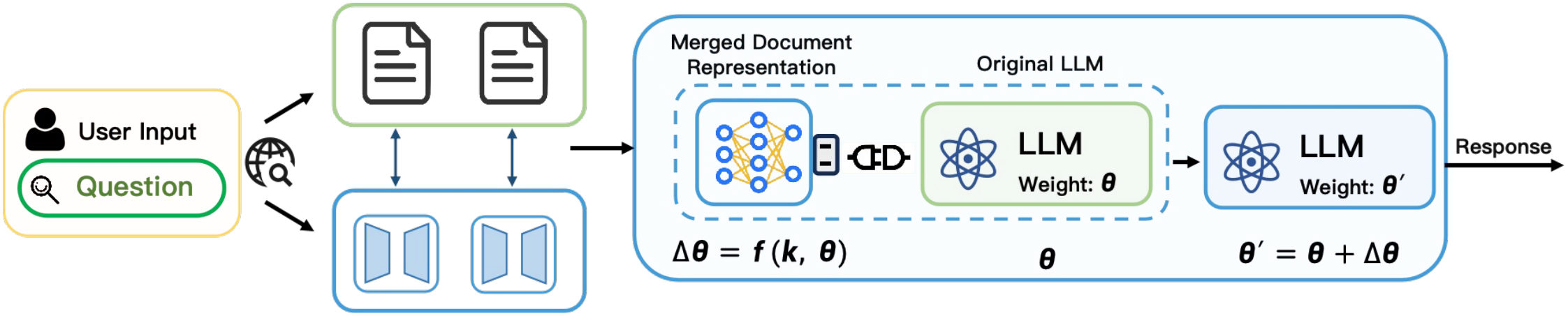


in-context text → parameterized knowledge

Su et al. Parametric Retrieval Augmented Generation. SIGIR, 2025.

Parametric RAG

Encode documents into adapters offline — **plug them in the inference time.**



Documents become parameter updates Δθ; the prompt carries only the query.

Offline

Each document → a LoRA adapter trained on the FFN layers

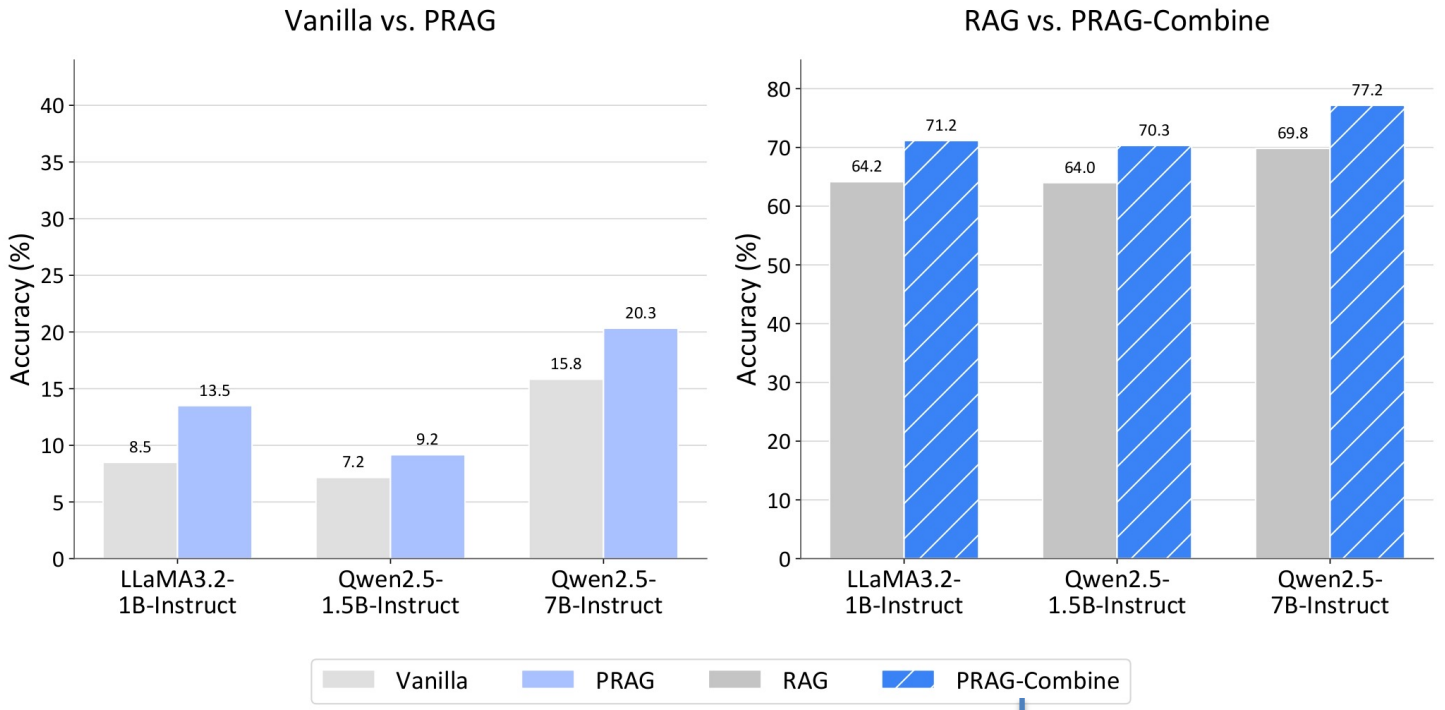
Online

Retrieve → plug in the document's LoRA → generate with **a query-only prompt**

Su et al. Parametric Retrieval Augmented Generation. SIGIR, 2025.

Parametric RAG

However, parameterized documents can lose **fine-grained details**.



- On unseen knowledge outside the LLM's training cutoff, PRAG alone is far weaker than RAG.
- Combining parametric injection with token context outperforms both — semantic guidance from parameters plus precise recall from context.

Both parametric and token based RAG

Tang et al. Understanding Parametric Knowledge Injection in Retrieval-Augmented Generation. Arxiv, 2025.

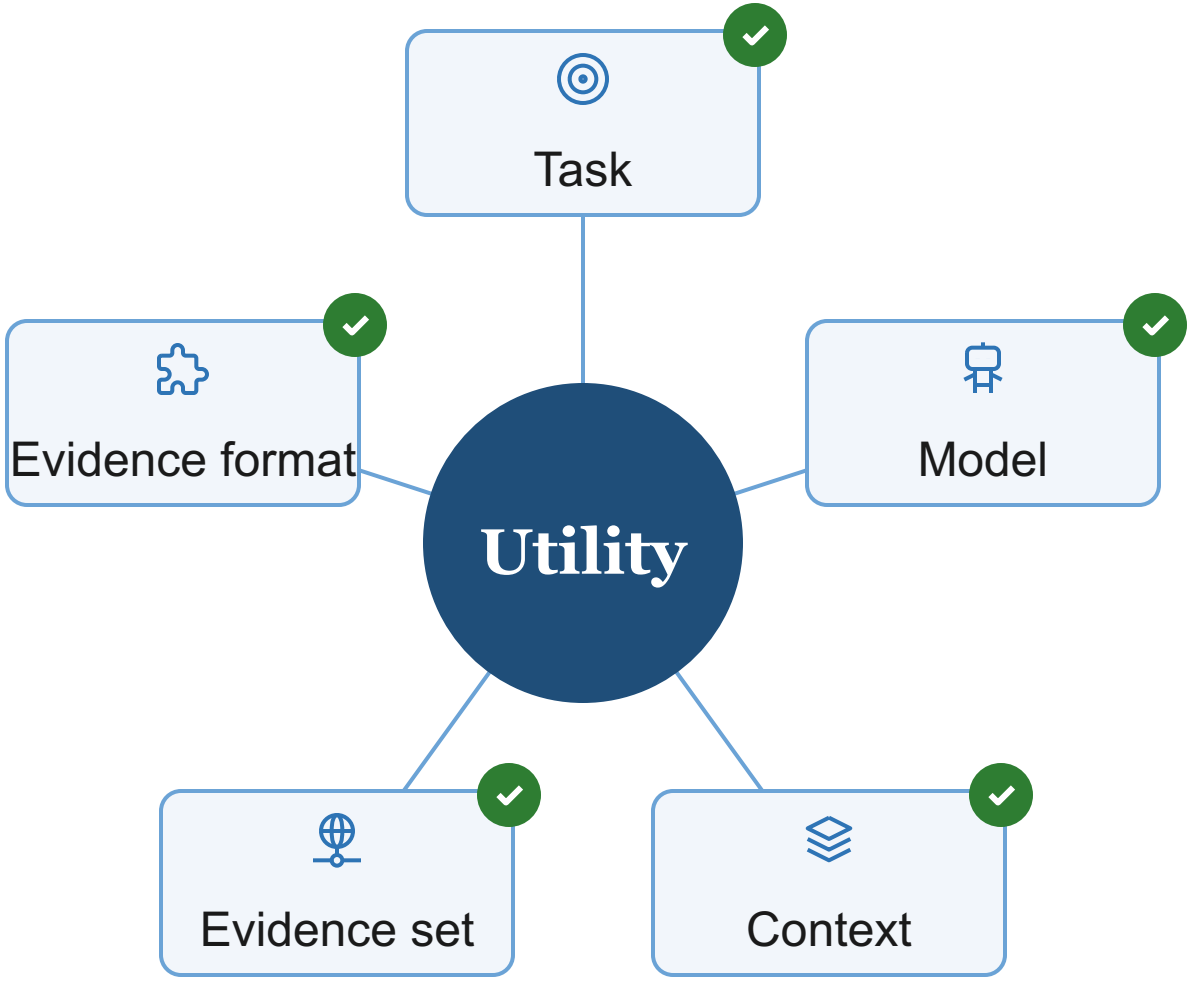
Evidence Format as Utility Transformation 2 — LLM-Centric Utility

Maximize evidence utility for the LLM

Method	Transformation	Utility
R2U	**Rewrite**	Log-likelihood shift toward the correct answer after rewriting
Chain-of-Note	**Note**	Per-document assessment that filters noise before generation
Passage Injection	**Reposition**	Reflecting retrieved passages within the reasoning process
Agent-Native Artifact	**Restructure**	Machine-executable structure that eliminates narrative compilation loss
Parametric RAG	**Parameterize**	Knowledge internalized as parameter updates

Different transforms, one test: **does the evidence help the model answer?**

The Wheel, Filled In



Task

"useful" is defined by the task

Model

agnostic scales; specific aligns

Context

judged given what is already selected

Evidence set

coverage · complementarity ·
low redundancy · conflict · provenance

Format

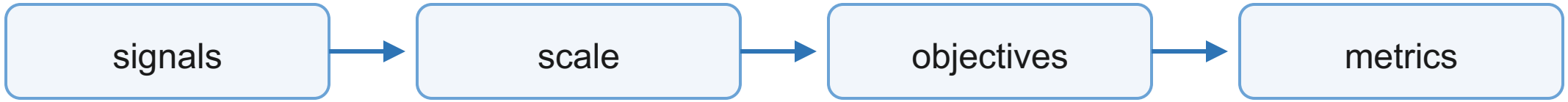
presentation changes utility

Section 2 — Takeaways

1. Utility **changes across tasks** — each task redefines "useful".
2. Utility can be **LLM-agnostic or LLM-specific**.
3. Utility is **pointwise, marginal, or set-level** within set construction.
4. Utility depends on **evidence format and presentation**, not only document identity.

NEXT — SECTION 3

We defined utility.
Can we optimize it?



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 3

Utility Modeling & Optimization

Measurement, training, and evaluation — utility made operational.

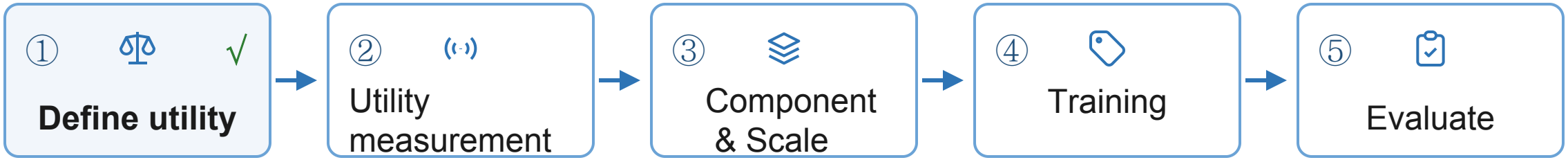
The Core Question

PICKING UP FROM SECTION 2

We defined utility. Can we optimize it?

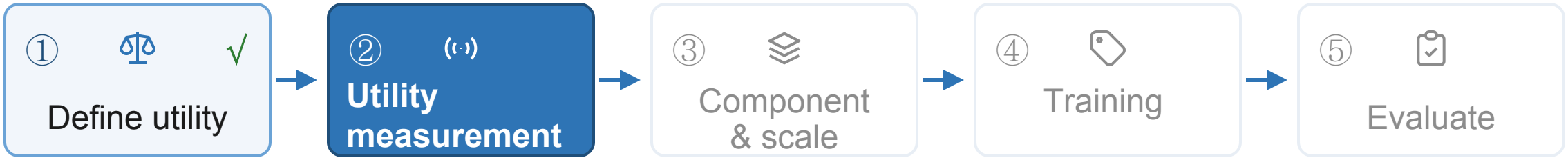
> How can we **modeling** utility — and **optimize** utility-focused retrievers/selectors/rerankers, and **evalute** utility-focused RAG?

The Utility Modeling Pipeline



done — Section 2

3.2 · Utility Measure Taxonomy



STAGE ② — SELECT A UTILITY MEASUREMENT

How to **measure** document **utility**?

Utility Measure Families

Verbalized LLM utility judgment ①

The LLMs **say** which document has utility or
the **rationale / reasoning** that explains what information is useful.

Attention-based method ②

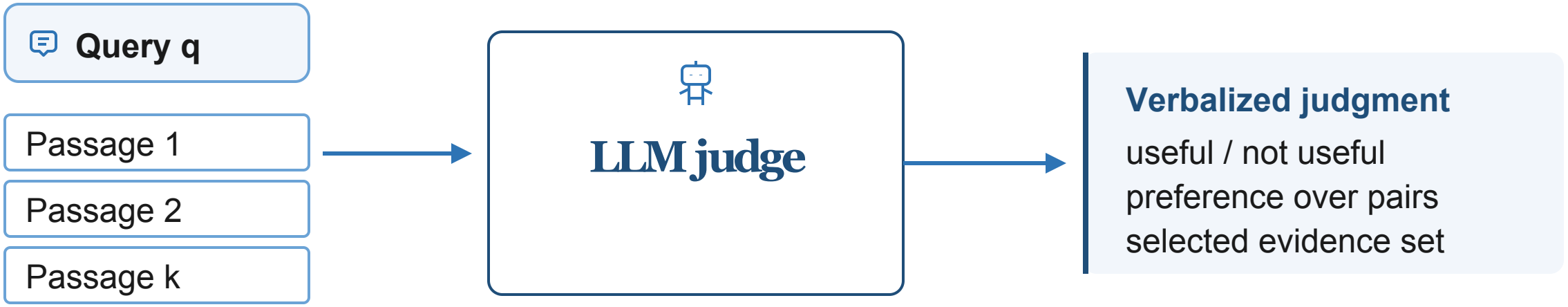
Use attention weights as soft utility to measure how the reader model **actually used** the passages.

Performance, likelihood ③

What the target model's downstream performance or whether the **response improves to measure utility**?

Family ① · Verbalized Judgment

Prompt a model to **say** whether evidence is useful — for a passage, a pair, a list, or a set.



Direct
output whether the passage has utility or not

CoT / Reason
reason first, then judge

Pseudo-answer
draft an answer, then judge document

Zhang et al. Are Large Language Models Good at Utility Judgments? SIGIR, 2024.

Pointwise, Pairwise & Listwise

The same judge, **three ways** to present the evidence.

Pointwise

score each passage on its own

q + one passage d

LLM judge

useful?

parallel & scalable · no comparison

Pairwise

compare two at a time

q + doc A + doc B

LLM judge

more useful: A ≻ B

reliable order · $O(k^2)$ calls

Listwise

judge the whole list together

q + all k passages at once

LLM judge

ranking / selection

context-limited

Zhang et al. Are Large Language Models Good at Utility Judgments? SIGIR, 2024.

How to Judge Utility via LLMs?

Different utility-based selection performance

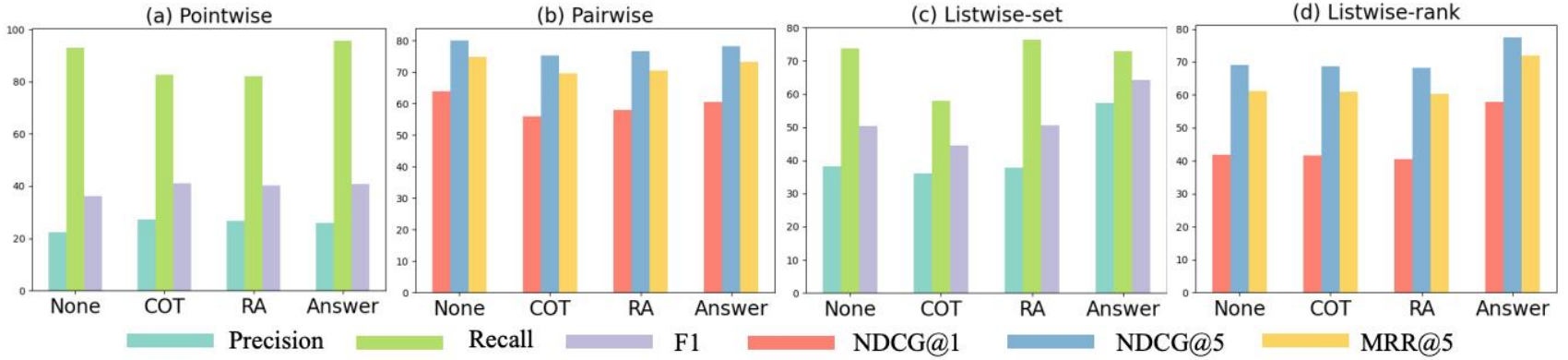


Figure 6: The performance (%) of utility judgments using ChatGPT under different forms, i.e., pointwise, pairwise, listwise-rank and listwise-set, on the NQ dataset using the prompts with different requirements.

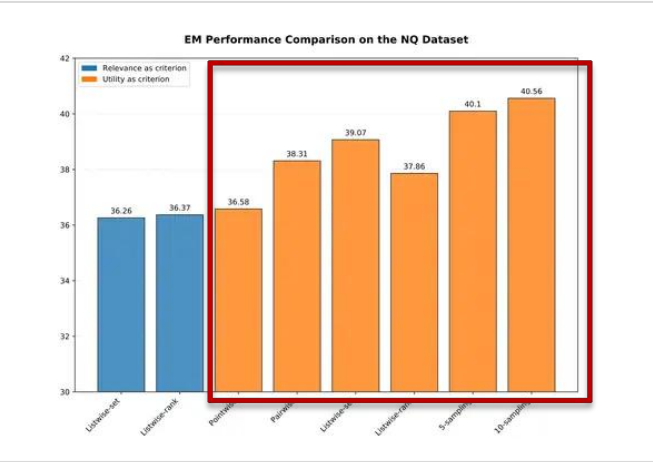
Reasoning and **pseudo-answer generation** markedly improve the LLM's utility judgments performance.

Zhang et al. Are Large Language Models Good at Utility Judgments? SIGIR, 2024.

Utility Beats Relevance

PAPER · VERBALIZED UTILITY JUDGMENT — FINDINGS

Utility- vs relevance-based selection



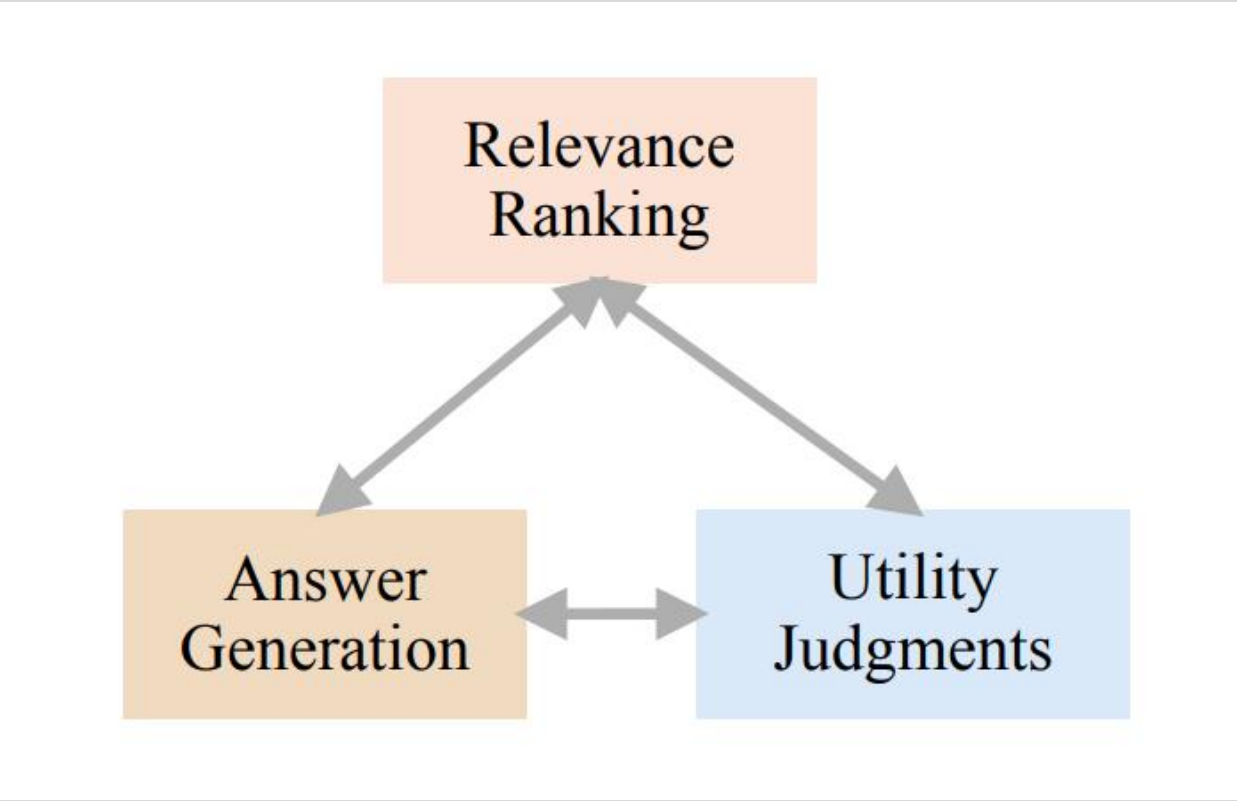
- ✓ **Utility-based selection** is better aligned with generation than relevance-based selection.
- ✓ **Listwise** selection achieves the best RAG performance compared to other judgment methods.

Zhang et al. Are Large Language Models Good at Utility Judgments? SIGIR, 2024.

Iterative Utility Judgment

PAPER · VERBALIZED UTILITY JUDGMENT

Judgment and answering **reinforce each other**, iteratively.



Inspired by relevance in philosophy

Human understanding grows through the dynamic interaction of topical, interpretive, and motivational relevance.

Answer from current selected evidence,

re-**judge** utility given that answer,

then answer generation again.

Zhang et al. Iterative Utility Judgment Framework via LLMs Inspired by Relevance in Philosophy. ACL Findings, 2026.

Family ① · Strengths & Risks

Strengths

- ✓ Flexible across tasks
- ✓ No generator logits needed
- ✓ Work with black-box systems
- ✓ No need ground truth answer
- ✓ Annotate unlabeled data at scale

Risks

- × prompt sensitivity
- × judge-model bias
- × relevance confused with utility

Family ① · Rationale & Process

PAPER · RATIONALE & PROCESS SUPERVISION

Score passages by how well they **match the answer's reasoning**.

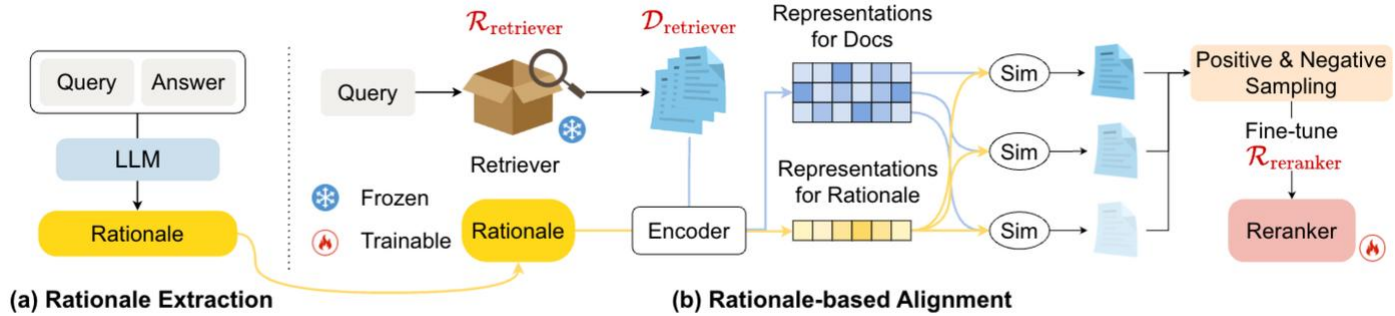


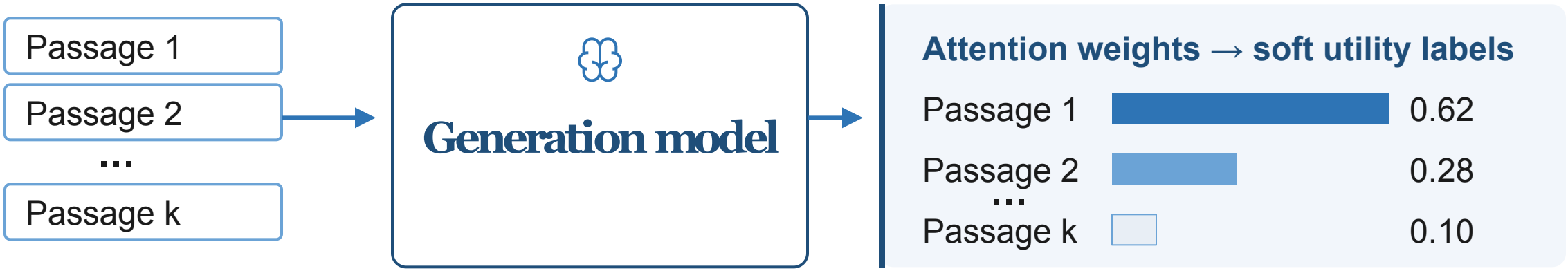
Figure 1: Overview of RADIO.

You are a professional QA assistant. Given a question and the ground truth answer, you can output the rationale why the ground truth answer is correct. Question: {question}. Answer: {answer}. Rationale:

1. Prompt the LLM for the **rationale** of the gold answer.
2. Score each passage using **cosine similarity** between the passage embedding and the rationale embedding

Jia et al. Bridging Relevance and Reasoning: Rationale Distillation in Retrieval-Augmented Generation. ACL Findings, 2025.

Family ② · Attention



FiD Attention

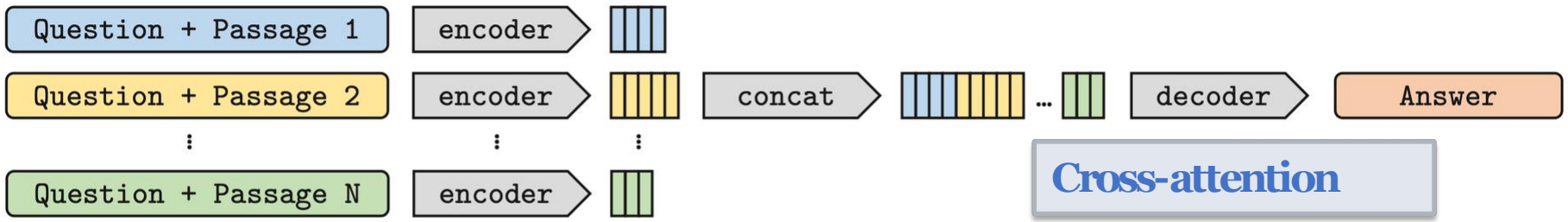


Figure 2: Architecture of the Fusion-in-Decoder method.

Fusion-in-Decoder (FiD) is a retrieval-augmented language model that independently encodes multiple retrieved passages and then fuses them during the decoding stage to generate an answer.

How the utility score of the passage is formed?

Average the cross-attention weights over all input tokens belonging to that passage, across all layers and all attention heads, to get a utility score.

Izacard and Grave. Distilling Knowledge from Reader to Retriever for Question Answering. arXiv, 2020.

Attention Needs the Value Norm

PAPER · ATTENTION / READER-DERIVED

Attention weight **alone** may be a noisy utility proxy.

Score = attention weight × norm of the value vector

$$\text{Utility}(\mathbf{d}_k) = \frac{1}{L \times H \times T} \sum_{l=1}^{L} \sum_{h=1}^{H} \sum_{t=1}^{T} \left(\sum_{n \in \mathcal{N}_k} \alpha_{l,h,t,n} \cdot \|\mathbf{v}_{l,h,t,n}\|_2 \right)$$

A token can receive high attention yet carry **little information**.

Weighting attention by the **value-vector norm** better reflects how much a passage actually contributes to the reader's output.

Izacard et al. Atlas: Few-shot Learning with Retrieval Augmented Language Models. JMLR, 2023.

Attention Is Architecture-Specific

PAPER · ATTENTION / READER-DERIVED

Attention-based method has **worst performance** in **decoder-only LLM**!

Source	NQ		HQ		TQ		MQ		FEVER		2Wiki		Overall
	Unknown	Known	Unknown	Known	Unknown	Known	Unknown	Known	Unknown	Known	Unknown	Known	
Llama3.1-8B-Instruct (Utility-Based Selection)													
Initial Retrieval (Top-20)	42.62	85.41	25.92	80.82	70.05	94.22	31.19	79.58	69.26	92.61	24.97	**62.85**	58.63
PVS (*w/* A)	43.45	**85.89**	27.09	**82.79**‡	**72.26**‡	**95.21**‡	31.23	80.23	56.72	90.75	23.96	60.63	57.52
LVS (*w/* A)	**44.06**‡	85.70	**28.62**†	**82.79**‡	71.11	94.86‡	**31.43**	**81.37**	**71.77**†‡	**93.29**†‡	**25.19**	62.11	**59.46**†‡
Llama3.1-8B-Instruct (Utility-Based Ranking)													
Retrieval (Top-5)	43.56	81.03	25.66	79.16	65.93	92.51	30.34	75.16	71.16	92.83	22.85	60.79	57.53
AttR (Top-5)	32.01	74.03	17.77	67.25	51.45	87.01	23.81	69.28	61.91	83.80	18.65	54.71	50.19
Likelihood (Top-5)	42.40	83.17	26.73‡	**80.43**	**70.58**‡	**94.63**‡	**30.73**	76.96	69.65	92.37	24.78‡	61.29	58.40‡
LVR (*w/* A) (Top-5)	**44.11***	**84.05**‡	**27.99**†‡*	79.48	69.82	94.29	30.65	**78.59**‡	**72.72***	**93.71**†‡*	**24.91**‡	**62.27**	**59.14**†‡*

All passages are concatenated with the query into a single sequence, which **spreads attention too thinly**. FiD, however, pairs each passage separately with the query, enabling clearer and more discriminative attention per document.

Zhang et al. LLM-Specific Utility: A New Perspective for Retrieval-Augmented Generation. arXiv, 2025.

Family ③ · Performance

Direct

Passage has utility if it **help LLMs achieve better task performance**

Answer likelihood

does p(correct answer) rise with the evidence?

Metric

EM / F1 / pass rate

Shi, et al. Replug: Retrieval-augmented black-box language models. NAACL 2024
Zamani, et al. Stochastic rag: End-to-end retrieval-augmented generation through expected utility maximization. SIGIR 2024

Family ③ · Performance

Information Gain

Passage has utility if it **improves the target model's ability** to produce the correct output compared to without it.

Answer likelihood change

Does p(correct answer) raise with the passage compared to without it?

Metric delta

EM / F1 / pass rate, with vs without passage

Cross-Mutual Info

Information gain about the answer beyond the query alone

Uncertainty reduction

The degree of moving the model response toward the correct cluster

Izacard et al. Atlas: Few-shot Learning with Retrieval Augmented Language Models. JMLR, 2023.
Qu et al. Uplift-rag: Uplift-driven knowledge preference alignment for retrieval-augmented generation. Findings of EMNLP 2025.

Family ③ · Performance

Information Gain

Passage has utility if it **improves the target model's ability** to produce the correct output compared to without it.

Answer likelihood change
Does p(correct answer) raise with the passage compared to without it?

Metric delta
EM / F1 / pass rate, with vs without passage

LLM-specific utility

Cross-Mutual Info
Information gain about the answer beyond the query alone

Uncertainty reduction
The degree of moving the model response toward the correct cluster

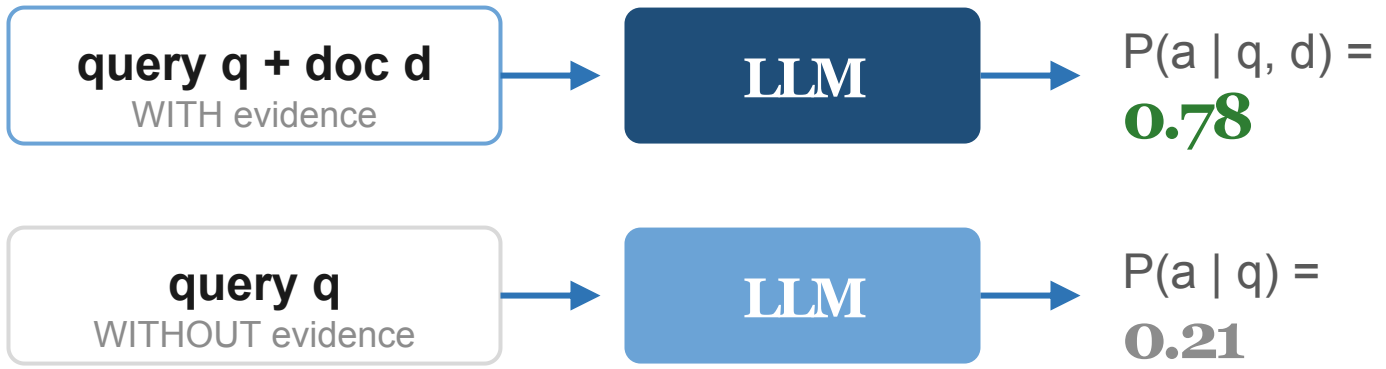
SePer

Izacard et al. Atlas: Few-shot Learning with Retrieval Augmented Language Models. JMLR, 2023.
Qu et al. Uplift-rag: Uplift-driven knowledge preference alignment for retrieval-augmented generation. Findings of EMNLP 2025.

Answer-Likelihood Divergence

PAPER · PERFORMANCE · LIKELIHOOD & BELIEF CHANGE

Utility = how much a document **raises the answer's likelihood** compared to without the document.



Utility score

$$u(d) = P(a|q,d) - P(a|q)$$

$$= 0.57$$

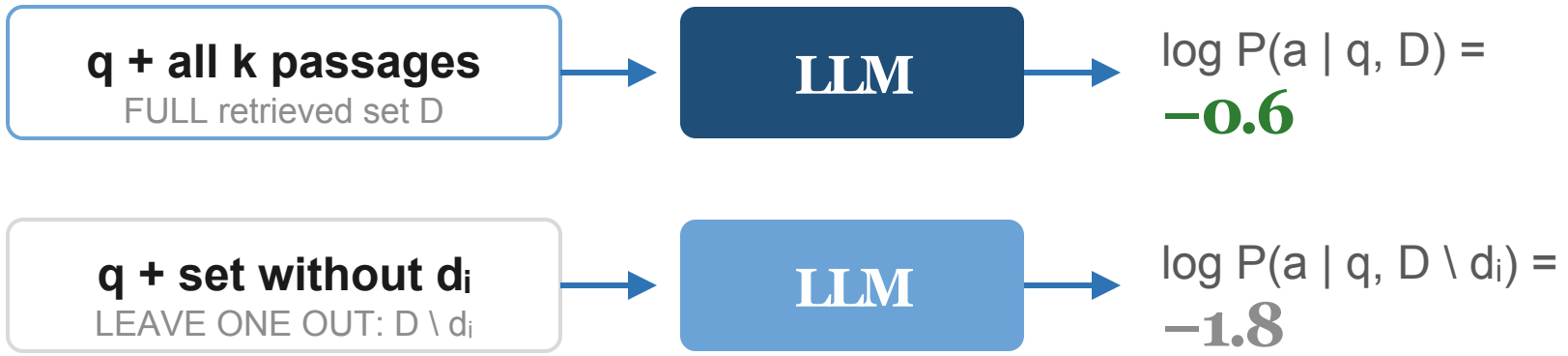
Practical scoring: the normalized log-probability of the ground-truth answer under the prompt containing q and d — a dense, differentiable, model-specific utility signal.

Qu et al. Uplift-rag: Uplift-driven knowledge preference alignment for retrieval-augmented generation. Findings of EMNLP 2025.

Atlas: Leave-One-Out Utility

PAPER · PERFORMANCE · LIKELIHOOD & BELIEF CHANGE

A passage is useful if **removing it hurts** the answer's likelihood.



Utility of d_i

= likelihood drop

= 1.2

LOOP (Leave-One-Out Perplexity distillation): score every passage by how much the answer's log-likelihood falls when it is dropped, normalize the drops into a target distribution over passages.

Izacard et al. Atlas: Few-shot Learning with Retrieval Augmented Language Models. JMLR, 2023.

FILCO: Cross-Mutual Info

PAPER · PERFORMANCE · LIKELIHOOD & BELIEF CHANGE

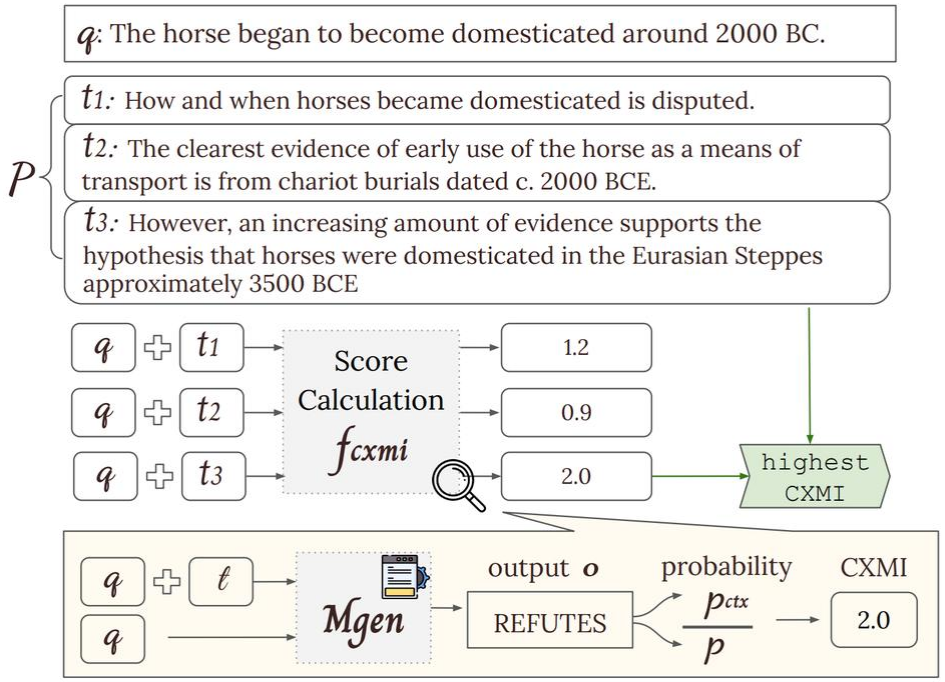


Figure 3: An example illustration of context filtering with the CXMI strategy.

Keep context that **changes** the model's answer.

Conditional Cross-Mutual Information

$$U(d|q,d) = \frac{M_{gen}(a^*|q+d)}{M_{gen}(a^*|q)}$$

the answer's log-probability lift from the context

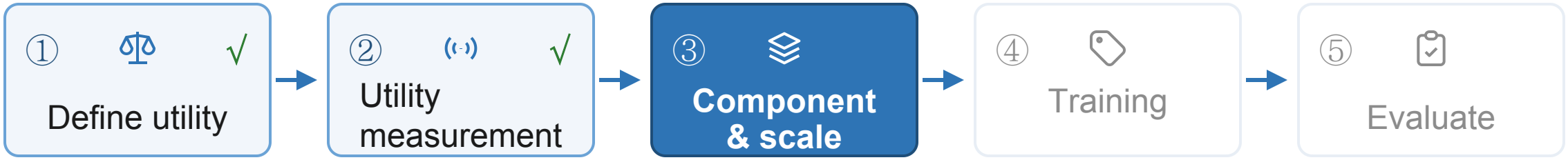
Filtering noisy context improves the generator's accuracy and robustness.

Wang et al. Learning to Filter Context for Retrieval-Augmented Generation. arXiv, 2023.

Families: Summary

Signal family	Advantage	Disadvantage	Best use
Verbalized judgment	• Flexible across tasks • Works with black-box systems • No generator logits needed	• Prompt sensitivity • Judge-model bias	High fidelity judgement and training labels
Attention-based method	• can work without the ground truth answer	• Need LLMs's attention weights	Distilling training labelds
Performance & Likelihood change	• LLM-specific utility • Task-specific utility	• Need ground truth answer • Need generator logits	Training labelds

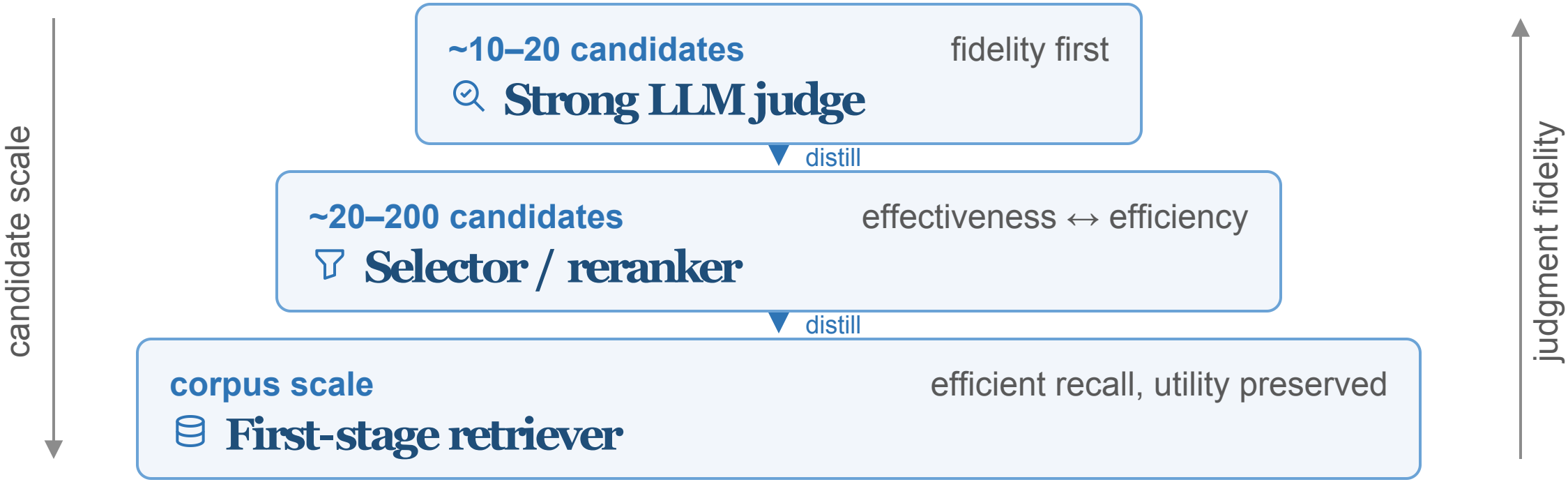
3.4 · The Efficiency Ladder



STAGE ③ — COMPONENT & SCALE

Candidate scale determines which **component** you train — and how to train using the **utility measurement results**.

The Efficiency Ladder



Level 1 · Small Sets: Judge Deeply

LLM judge

selector / reranker

first-stage retriever

~10–20 candidates
every passage affords deep assessment

- **Deep judgment per candidate** — can use strong LLM to judge and collect high-quality utility labels

Level 1 · Anchors

High-fidelity judgment powers **evaluation and annotation**.

LLMs Utility Judge

Ask stronger LLMs output whether the passage has utility under **small candidate list.**

Distillation Selector

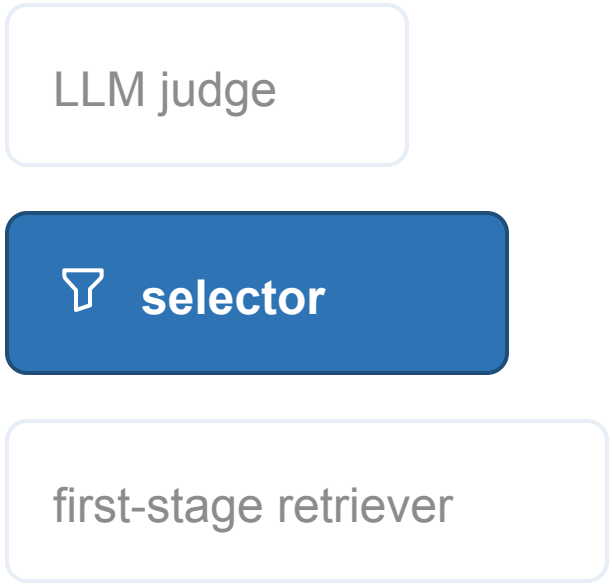
Using stronger LLMs in utility judgments for **utility-focused annotation**

SePer

Using change in the model's belief about the correct answer for better **evaluation** metric

Are Large Language Models Good at Utility Judgments. 2024.
Distilling a Small Utility-Based Passage Selector to Enhance Retrieval-Augmented Generation. 2025
SePer: Measure Retrieval Utility Through the Lens of Semantic Perplexity Reduction. 2025

Level 2 · Medium Sets



- Directly applying LLM judgments becomes too expensive
- Train a small selector/ranker to judge utility

Goal: **preserve utility while filtering efficiently**.

Level 2 · Anchors

Selectors/rankers carry utility into **medium-scale after training**.

SECTION 2, REVISITED

Distilled Utility Selector

After **distillation from LLMs**, small model selects the passages have utility among the large candidate list

SECTION 2, REVISITED

SETR

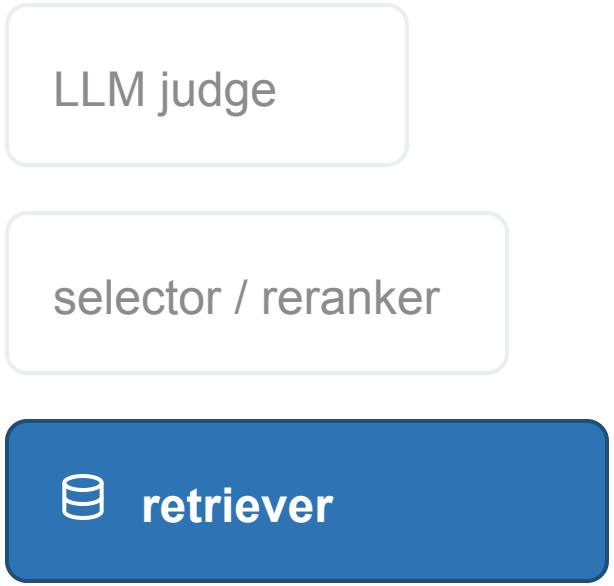
After **distillation from LLMs**, small model selects the **set** that meets the query's information needs.

RAG-DDR

After **RL training**, small model give the passage utility score.

Distilling a Small Utility-Based Passage Selector to Enhance Retrieval-Augmented Generation. 2025.
Shifting from Ranking to Set Selection for Retrieval-Augmented Generation. 2025.
RAG-DDR: Optimizing Retrieval-Augmented Generation Using Differentiable Data Rewards. 2024.

Level 3 · Corpus Scale



corpus scale
no LLM can score
the whole corpus

- A massive number of candidates make full judgement infeasible.
- Utility must already live inside the retriever.

> Train a **utility-focused dense retriever** to achive this goal.

Level 3 · Anchors

Four routes carry utility into **corpus-scale retrievers after training**.

Utility-Focused Annotation 2025

LLM utility labels become **contrastive training data** for general retrievers.

REPLUG 2023

The frozen LM's likelihood distribution is the **KL distillation** target.

EMDR² 2021

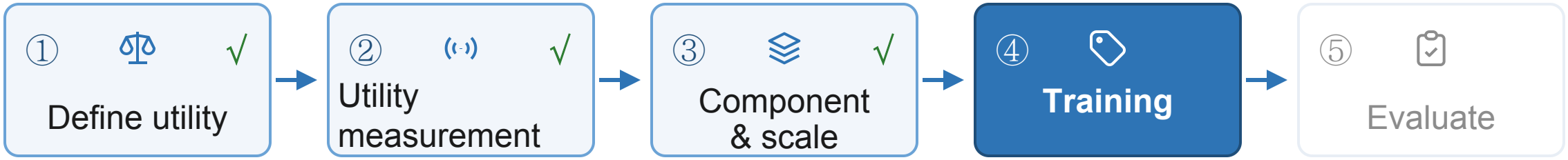
Documents as **latent variables** — EM-style joint training from answers alone.

Stochastic RAG 2024

Expected utility maximization over sampled sets via differentiable top-k.

Leveraging LLMs for Utility-Focused Annotation. 2025. · REPLUG: Retrieval-Augmented Black-Box Language Models. 2023.
End-to-End Training of Multi-Document Reader and Retriever for Open-Domain Question Answering. 2021.
Stochastic RAG: End-to-End Retrieval-Augmented Generation through Expected Utility Maximization. 2024.

3.5 · Label ↔ Objective Pairing



STAGE ④ — PAIR LABEL WITH OBJECTIVE

Training should consider
both the utility measurement method and the candidate scale.

Training the Selector & Reranker

Medium scale (~20–200):
six ways to turn a utility label into a **trained selector/re-ranker**.

Sequence Distillation

copy the teacher LLM's selected-set decisions

Regression

predict a continuous utility / gain score

Classification

useful / not-useful, context-aware

Contrastive Learning

pull useful passages in, push useless ones away

Downstream Feedback

selected set should match gold-set task performance

Reinforcement Learning

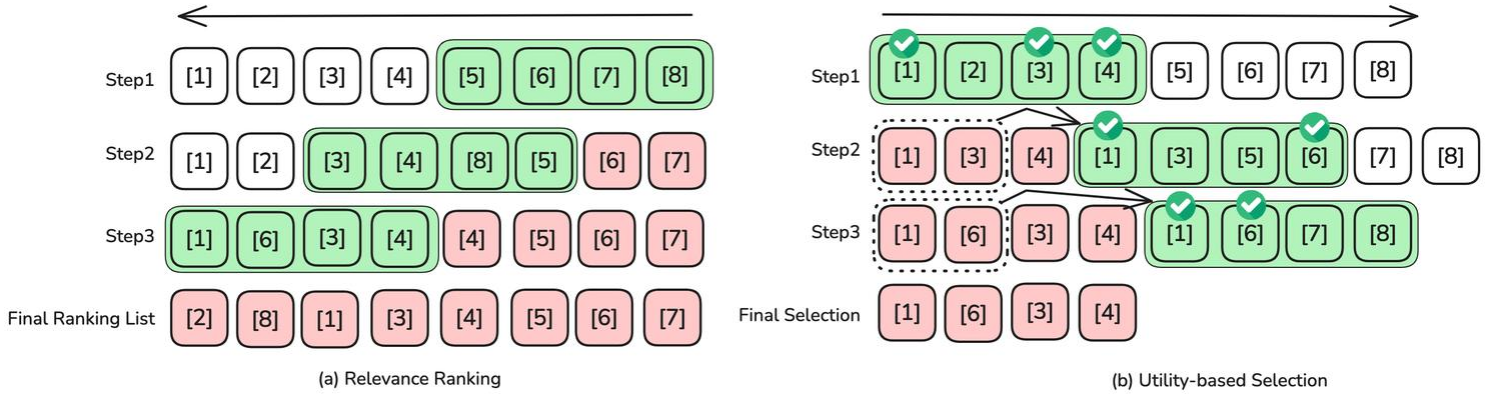
reward the selector when it helps the generator

The choice depends on the available **utility labels** and the **interaction between retrieval and generation**.

Distilled Utility Selector

PAPER · SEQUENCE DISTILLATION

Distill a stronger LLM's **utility selection** into a small model.



Teacher → student

32B LLM annotates utility-based selection → distill into 1.7B Utility models.

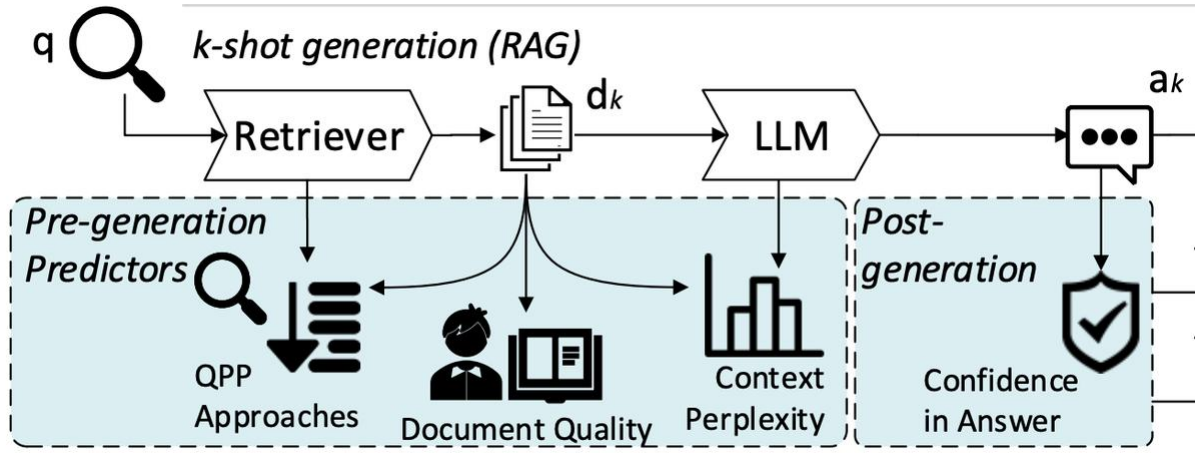
Utility measurement method

LLM Verbalized Judgements

Zhang et al. Distilling a Small Utility-Based Passage Selector to Enhance Retrieval-Augmented Generation. SIGIR-AP, 2025.

Predicting Retrieval Utility

PAPER · REGRESSION



Utility measurement method

Performance & Likelihood change

Regress a **calibrated score** from query + candidate features.

Tian et al. Predicting Retrieval Utility and Answer Quality in Retrieval-Augmented Generation. ECIR, 2026.

Contextual Passage Utility

Question:
What nationality was James Henry Miller's wife?

Supporting Passages

Passage 1:

Ewan MacColl - James Henry Miller (25 January 1915 – 22 October 1989), better known by his stage name Ewan MacColl, was an English folk singer, songwriter, communist, labour activist, actor, poet, playwright and record producer.

Passage 2:

Peggy Seeger - Margaret "Peggy" Seeger (born June 17, 1935) is an **American** folksinger. She is also well known in Britain, where she has lived for more than 30 years, and was married to the singer and songwriter Ewan MacColl until his death in 1989.

Utility Scores

Independent Assessment

Passage	Score
Passage 1	4
Passage 2	1

Conditioned on Passage 1

Passage	Score
Passage 1	5
Passage 2	4

Passage 2 independently provides little utility for answering the question. However, when conditioned on Passage 1 (which establishes that James Henry Miller used the stage name Ewan MacColl), Passage 2 becomes highly useful as it identifies Peggy Seeger as Ewan MacColl's American wife.

Utility measurement method

LLM Verbalized Judgements

Synthetic supervision

Use reasoning process traces passage order by o1;
Use GPT-4o assigns 1–5 utility score.

RoBERTa classifier

Fine-tuned to predict contextual utility,
learning inter-passage dependencies for multi-hop QA.

Jain and Garimella. Modeling Contextual Passage Utility for Multihop Question Answering. IJCNLP-AACL, 2025.

Downstream Feedback

PAPER · DOWNSTREAM PERFORMANCE FEEDBACK

- **FER** uses verifier feedback to train a selector performing like the gold set.

$$\mathcal{L}_{uti} = y^* D_\phi(c, E^*) - y^* D_\phi(c, R_\theta(c, S)),$$

- **Read It Twice** minimizes the downstream performance of the removed set's complement.

$$\mathcal{L}_{full} = \mathcal{L}_{CE}(\mathcal{F}_{ver}(S_{emb}), y^*) - \mathcal{L}_{CE}(\mathcal{F}_{ver}(S_{emb} \backslash E_{emb}, y^*).$$

Utility measurement method

Performance & Likelihood change

Making discrete selection trainable

selection is discrete, so gradients cannot naturally flow through the selection process. Methods such as **Gumbel-Softmax** provide differentiable approximations.

Zhang et al. From Relevance to Utility: Evidence Retrieval with Feedback for Fact Verification. EMNLP Findings, 2023.
Hu et al. Read It Twice: Towards Faithfully Interpretable Fact Verification by Revisiting Evidence. SIGIR, 2023.

RAG-DDR

PAPER · REINFORCEMENT LEARNING

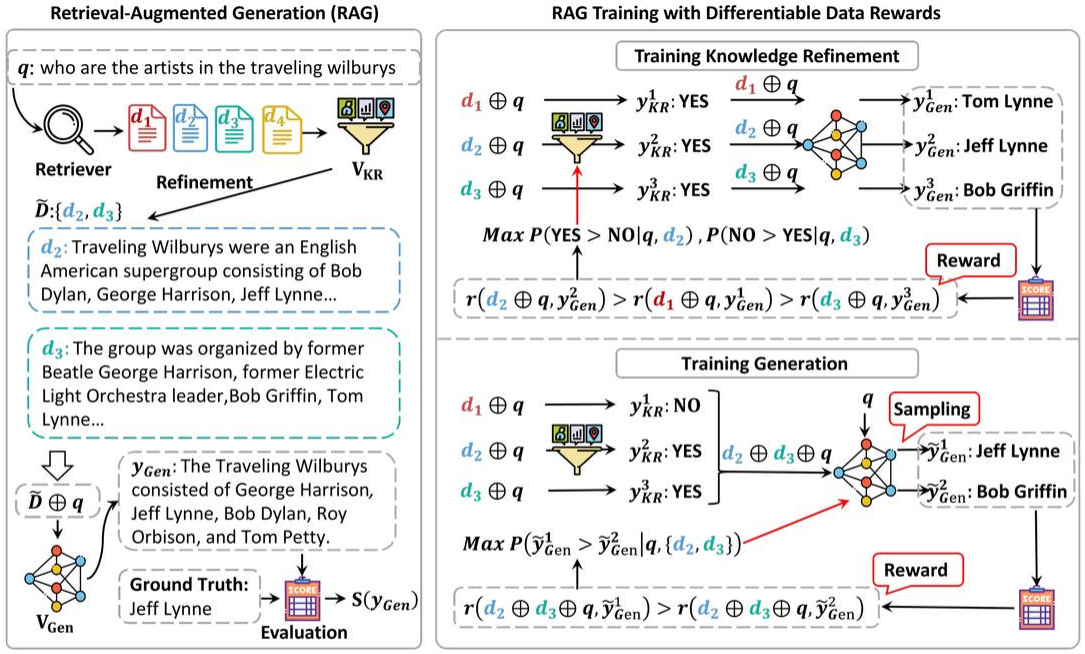


Figure 1: The Illustration of End-to-End Retrieval-Augmented Generation (RAG) Training with Our Differentiable Data Reward (DDR) Method. During training, we iteratively optimize the Generation module (V_{Gen}) and Knowledge Refinement module (V_{KR}).

Utility measurement method

Direct performance, like EM, F1

Iterative DPO

- **Iteratively tune the knowledge refinement and generation modules.**
  - Typically optimize the generation module first, followed by the knowledge refinement module.

Li et al. RAG-DDR: Optimizing Retrieval-Augmented Generation Using Differentiable Data Rewards. ICLR, 2025.

Training the Selector & Reranker

Four label forms cover **medium-scale** training.

Training style	Label / reward form	Typical use
Classification	useful / not useful · multi-class	fast filtering
Regression	utility score · likelihood / metric gain	calibrated utility scoring
DPR / RL	downstream performance reward	when discrete selection blocks gradients
Sequence / set distillation	teacher-selected list or evidence set	scaling LLM selection

All four fit the ladder's Level 2 — the selector or reranker scoring 20–200 candidates.

Utility-Focused Retriever

Corpus scale:
no LLM can score millions of docs-utility must live in the **retriever**.

Contrastive

Utility positives pulled in, utility negatives pushed away.

Utility-Focused Annotation
SCARLet

KL Distillation

Match the retriever's scores to a reader / LM distribution.

FiD-KD
REPLUG

End-to-End

Marginal likelihood / expected utility, jointly.

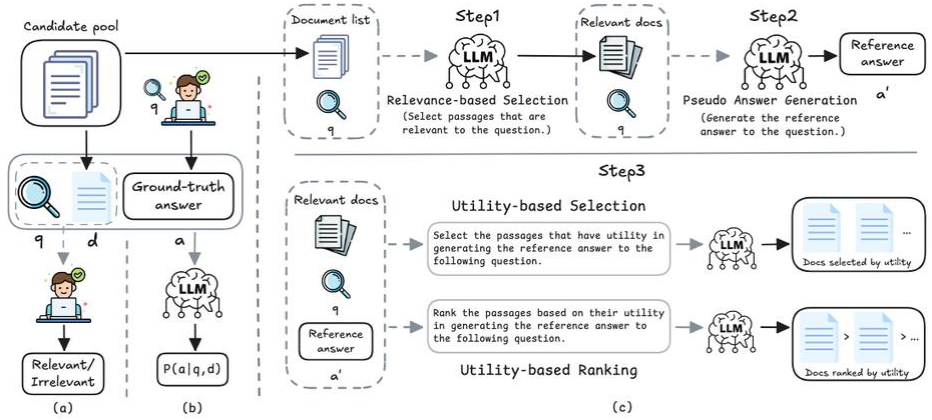
RAG
Stochastic RAG

All three **compress the reader's utility judgment** into a fast vector-space retriever that preserves it across the whole corpus.

Utility-Focused Annotation

PAPER · CONTRASTIVE RETRIEVER TRAINING

Let an LLM label utility annotation — then **train the retriever** on them.



Utility measurement method

LLM Verbalized Judgements

Three annotation sources compared

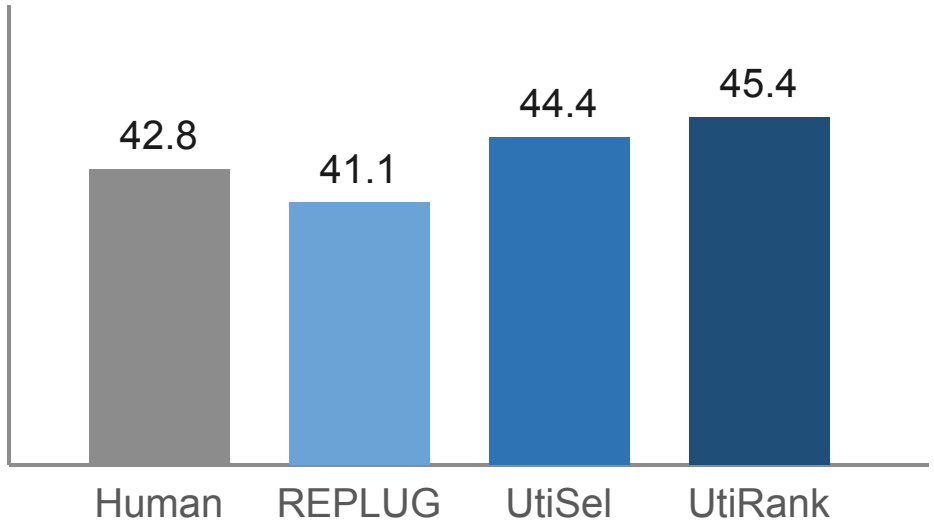
① Human annotation ② Ground-truth answer likelihood ③ **LLM utility judgment**

Zhang et al. Leveraging LLMs for Utility-Focused Annotation: Reducing Manual Effort for Retrieval and RAG. EMNLP, 2025.

Utility Labels Generalize Better

PAPER · CONTRASTIVE RETRIEVER TRAINING — FINDINGS

Retriever trained on utility vs other labels



Handling multiple LLM positives

Component	Variants	MRR@10	R@1000
Training Loss	Rand1LH	34.5	**97.9**
	JointLH	34.0	97.5
	SumMargLH	**35.3**	97.7

The LLM marks several positives per query.

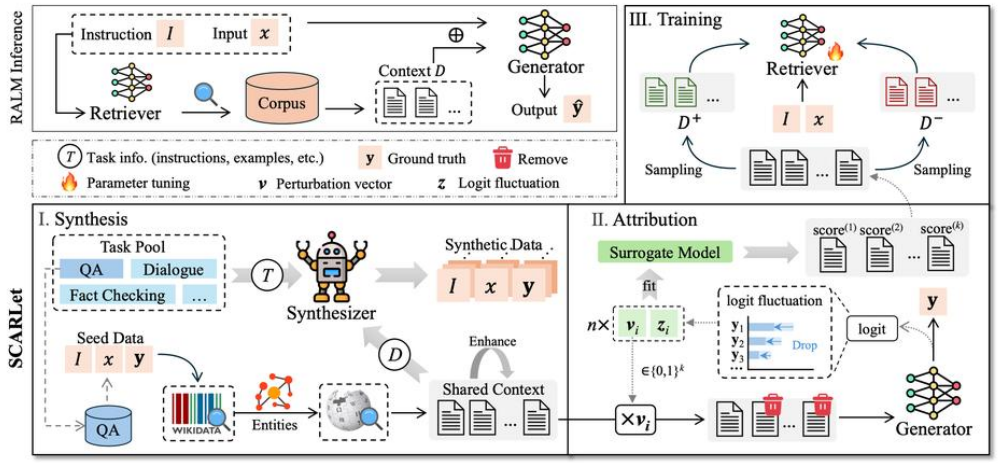
Objectives: Single / Joint / Sum-of-marginal likelihood — SumMargLH is most robust to noise.

> Utility-focused annotation **generalizes better** in retrieval and RAG compared to other annotations.

Zhang et al. Leveraging LLMs for Utility-Focused Annotation: Reducing Manual Effort for Retrieval and RAG. EMNLP, 2025.

PAPER · CONTRASTIVE RETRIEVER TRAINING

Turn continuous utility score from likelihood change into **positives** and **negatives**.



Utility measurement method

Performance & Likelihood change

High-utility passages become **positives**, low-utility passages become **negatives**, and uncertain examples are discarded.

Xu et al. Training a Utility-Based Retriever Through Shared Context Attribution for RALMs. EMNLP, 2025.

KL-Distilling Attention

PAPER · KL DISTILLATION

KL distillation transfers **reader knowledge** into the retriever.

$$\mathcal{L}_{\mathrm{KL}}(\theta, \mathcal{Q}) = \sum_{q \in \mathcal{Q}, p \in \mathcal{D}_q} \tilde{G}_{q,p}(\log \tilde{G}_{q,p} - \log \tilde{S}_\theta(q,p)),$$

Utility measurement method

Attention-based method

1. The FiD's cross-attention provides a utility distribution over passages, and the retriever learns to approximate this distribution.

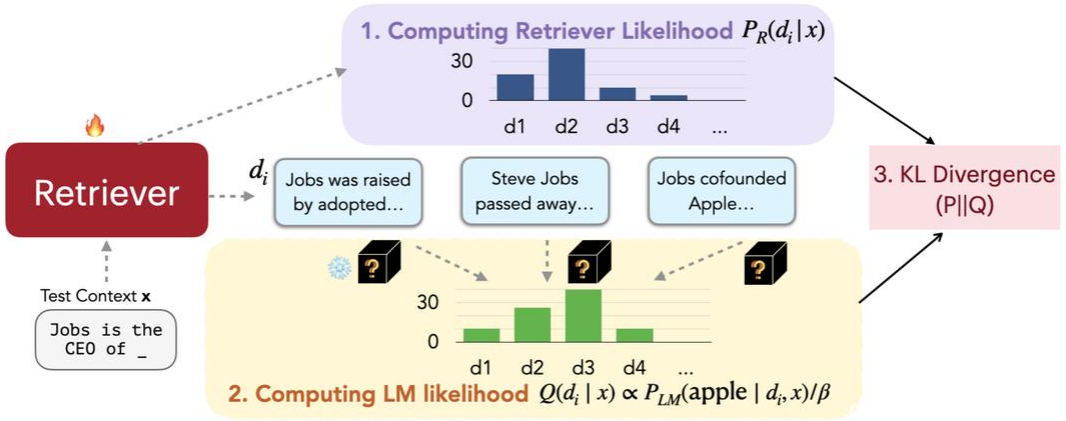
2. repeated retrieval and distillation, the retriever gradually becomes utility-aware..

Izacard and Grave. Distilling Knowledge from Reader to Retriever for Question Answering. arXiv, 2020.

KL-Distilling Likelihood

PAPER · KL DISTILLATION

The retriever learns from a **frozen** LM.



Utility measurement method

Direct likelihood

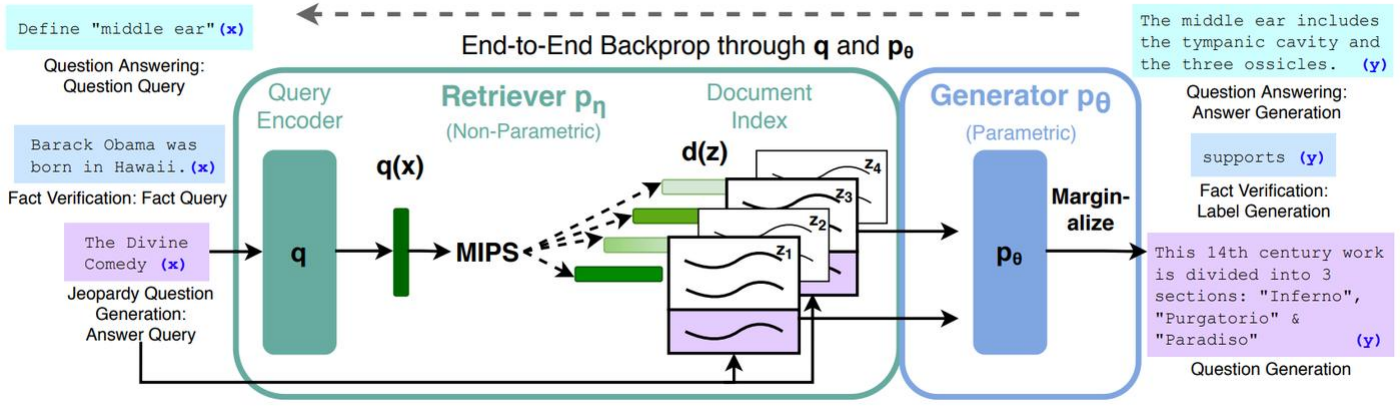
Likelihood as the teacher

KL-match the retriever's similarity distribution and utility distribution.

Shi et al. REPLUG: Retrieval-Augmented Black-Box Language Models. NAACL, 2024.

RAG: End-to-End Training

Utility measurement method **Direct likelihood**



Retriever is a latent variable; treat top-k docs as a mixture

gradients flow to the retriever through the marginal likelihood

RAG-Sequence

One document conditions the whole answer. Strong on extractive QA; more diverse.

RAG-Token

A different document per token. Better on generative tasks; more factual.

Table 5: Ratio of distinct to total tri-grams for generation tasks.

	MSMARCO	Jeopardy QGen
Gold	89.6%	90.0%
BART	70.7%	32.4%
RAG-Token	77.8%	46.8%
RAG-Seq.	83.5%	53.8%

Jointly training retriever + generator on the end task sets strong open-domain QA results

the retriever is shaped purely by answer usefulness.

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

Stochastic RAG: Expected Utility

PAPER · END-TO-END

Optimize the **expected utility** of a sampled evidence set.

$$
\begin{aligned}
p(\hat{y}|x; G_\theta, R_\phi) &= \sum_{\mathbf{d} \in \pi_k(C)} p(\hat{y}, \mathbf{d}|x; G_\theta, R_\phi) \\
&= \sum_{\mathbf{d} \in \pi_k(C)} p(\hat{y}|x, \mathbf{d}; G_\theta) p(\mathbf{d}|x; G_\theta, R_\phi) \\
&= \sum_{\mathbf{d} \in \pi_k(C)} p(\hat{y}|x, \mathbf{d}; G_\theta) p(\mathbf{d}|x; R_\phi)
\end{aligned}
$$

Utility measurement method

Direct performance, like EM, F1

The Gumbel-top-k techniques make discrete retrieval decisions trainable, allowing the system to directly optimize downstream utility.

Zamani and Bendersky. Stochastic RAG: End-to-End Retrieval-Augmented Generation Through Expected Utility Maximization. SIGIR, 2024.

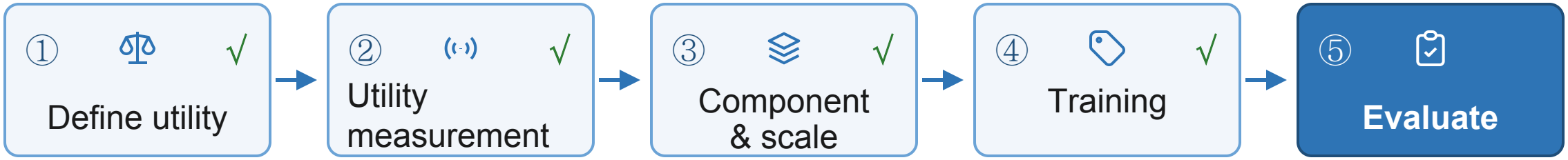
Training the Retriever & System

Corpus- and system-level objectives pair with **four more forms**.

Training style	Label / reward form	Typical use
Contrastive learning	utility positives & negatives	corpus-scale retrieval
KL distillation	soft distribution from reader / LM	distilling model behavior
End-to-end differentiable	marginal likelihood · expected utility	joint retriever–generator learning

The **key principle** is to choose objectives based on the component being optimized

3.6 · Evaluating Utility



STAGE ⑤ — EVALUATE

How to evalute the **whole utility-focused RAG system**?

Anchor: eRAG

Instead of relying on human relevance labels, each document is evaluated by asking the LLM to answer using that document alone.

eRAG

- correlates with end-to-end RAG quality **far better than relevance labels**
- These per-document scores are then aggregated by standard IR metrics (MAP / MRR / NDCG@k) **into a retriever-level utility score.**

Salemi and Zamani. Evaluating Retrieval Quality in Retrieval-Augmented Generation. SIGIR, 2024.

PAPER · EVALUATION

Coverage@K — Measures partial success. It calculates the fraction of required facts (atomic units) that are retrieved in the topK passages.

PerfRecall@K — Measures strict success. It is a binary metric: does the retriever find all required facts within the top-K

Retrieval vs parametric	Hit?	Parametric?	What it reveals
Intersection	Yes	Yes	Both sources agree (redundant signals)
Retrieval-only	Yes	No	Context utilization — using facts it lacks internally
Parametric-only	No	Yes	Parametric retention — model relies on its own knowledge
Complementary	No	No	Signal fusion — neither alone suffices, together correct

Cho and Lee. RARE: Redundancy-Aware Retrieval Evaluation Framework for High-Similarity Corpora. arXiv, 2026.

Section 3: Takeaways

1. Utility signals come from **LLM judgments, reader attribution, performance & belief change, and process traces**.

2. **Candidate scale** determines the right component and training paradigm.

3. Labels map to objectives **many-to-many** — the same signal can feed different training objectives.

4. Evaluation must measure **downstream usefulness**, not only document relevance.

NEXT — SECTION 4

We optimized one retrieval.
What about a whole trajectory?

agentic RAG → search agents → action rewards

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 4

Utility in Agentic RAG

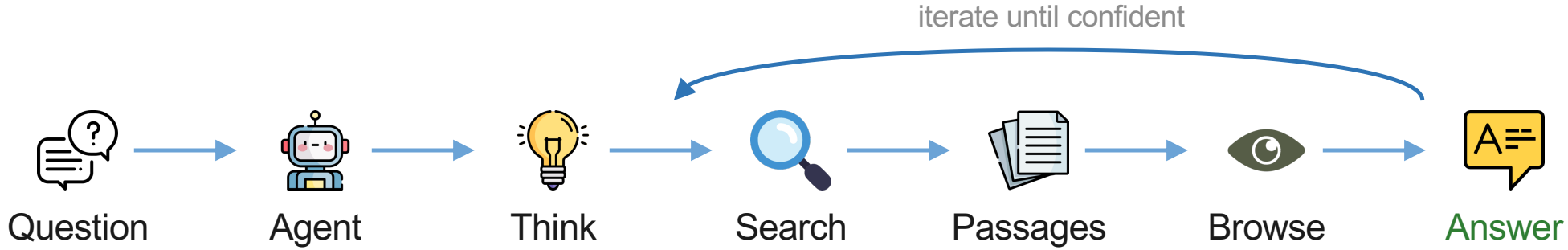
When utility becomes state-dependent — query, document, and action utility across an agent's trajectory.

4

Static RAG vs. Agentic RAG

STATIC RAG — retrieve once, then generate an answer

AGENTIC RAG — reason, search, and verify iteratively across steps

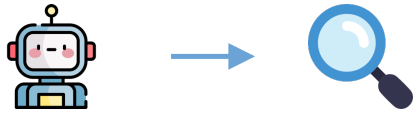


Utility becomes **dynamic** —
the information need changes over steps, so what counts as useful changes too.

Three Utilities in Agentic RAG

① Query-side Utility

What query should it issue for the current state?

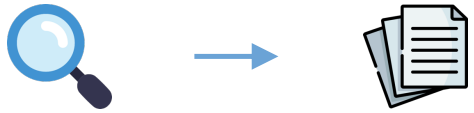


`[ query: ? ]`

a state-relevant query

② Document-side Utility

Does the retrieved evidence **actually help**?



`useful? ✓ / ✗`

local now & global answer

③ Action-side Utility

Which **action next** — incl. when to retrieve, or stop?

`[ Search / Browse / Find / Bash / Answer … ]`

A fourth, **trajectory-level utility/reward** binds the three — it supplies their training signal.

The Three Utilities — Map

LLM Information Need

The query-side question —
what query should the model issue for its current state?

The Three Utilities — Map

LLM Information Need

The query-side question — what query should the model issue for its current state?

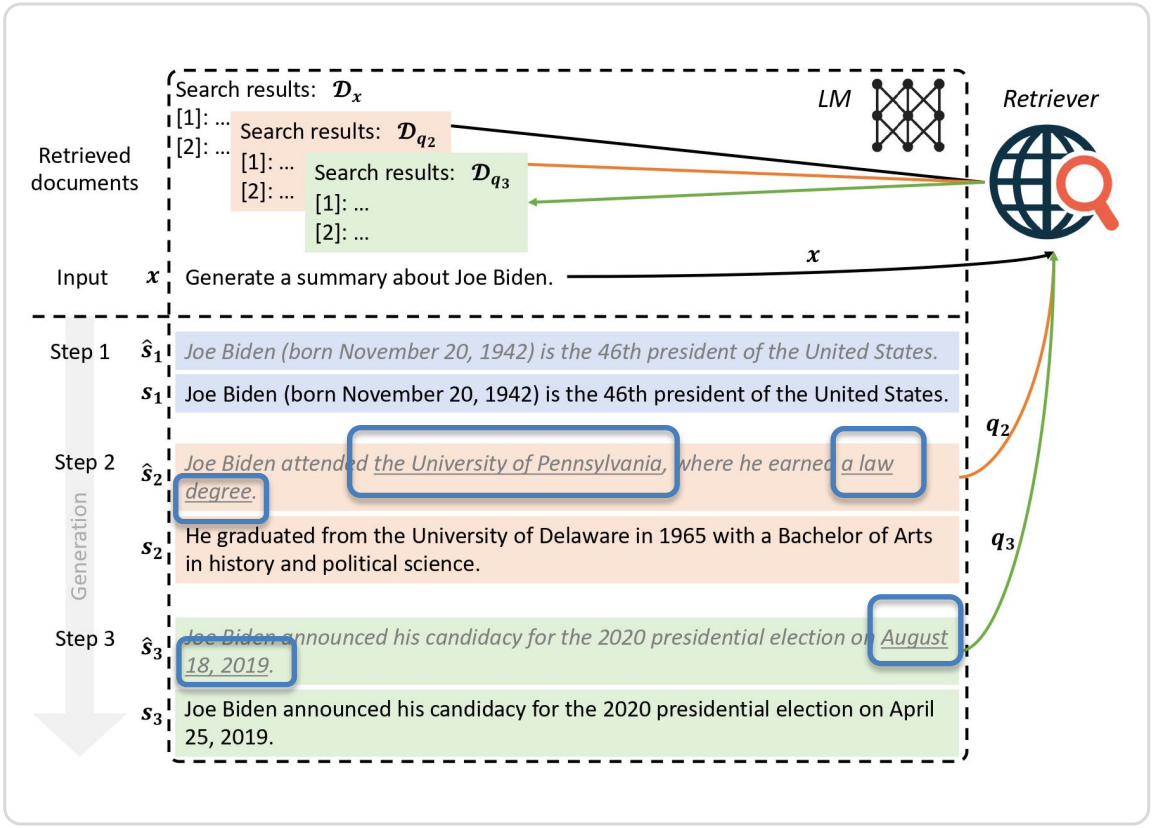
> Its companion — **when to retrieve** (retrieve vs. keep reasoning) — is an **action-side** decision; we give it its own treatment in later. They stay coupled — the same internal-state signal (FLARE, DRAGIN, Search-R1) informs both.

FLARE

QUERY-SIDE — WHERE KNOWLEDGE IS MISSING

> LLMs are fairly **well-calibrated**: low token probability is a proxy for missing knowledge.

Low-probability tokens in the next generated sentence pinpoint **where the model's knowledge is missing** — the gap a retrieval should go fill.



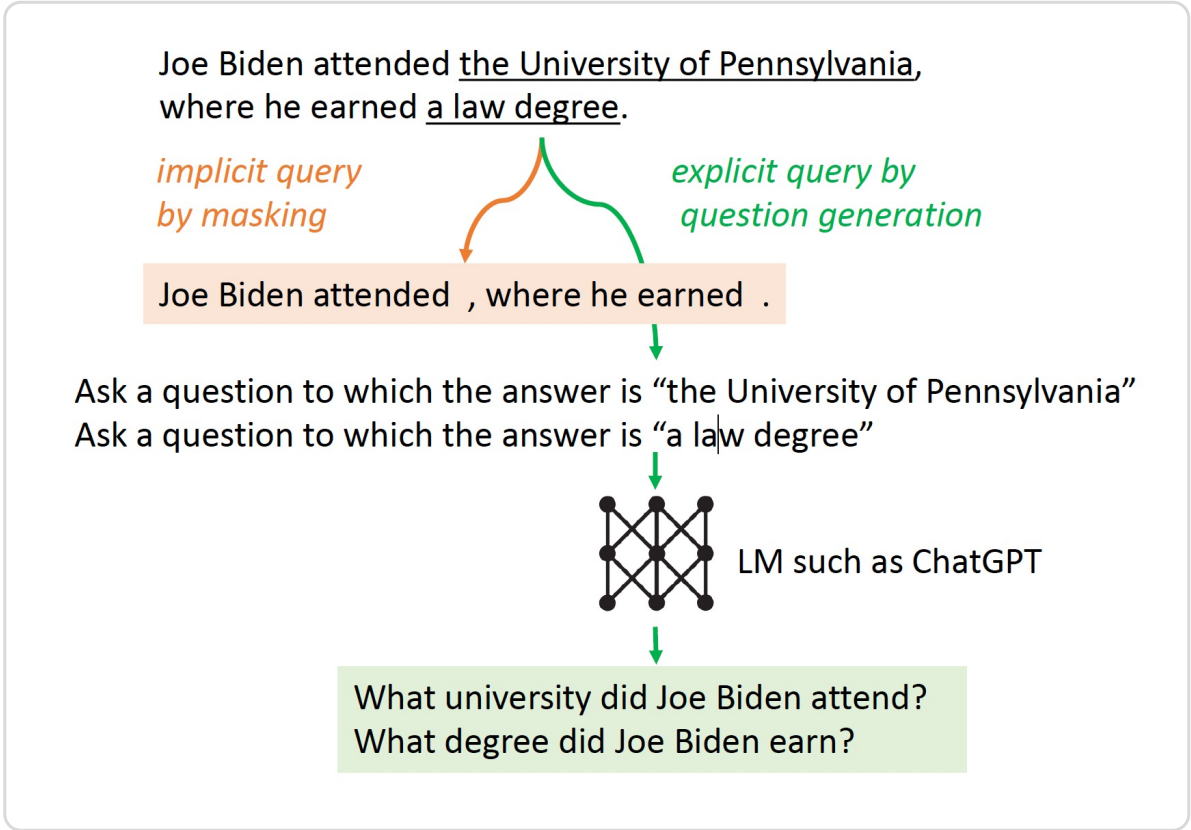
Jiang et al. Active Retrieval Augmented Generation. EMNLP, 2023.

FLARE — What Query to Issue

WHAT QUERY?

- **Implicit query** — mask the low-confidence tokens in the predicted next sentence.
- **Explicit query** — prompt the LLM to ask a question that targets the low-confidence span.

> Query utility = **filling the model's uncertainty** at this generation step.



Jiang et al. Active Retrieval Augmented Generation. EMNLP, 2023.

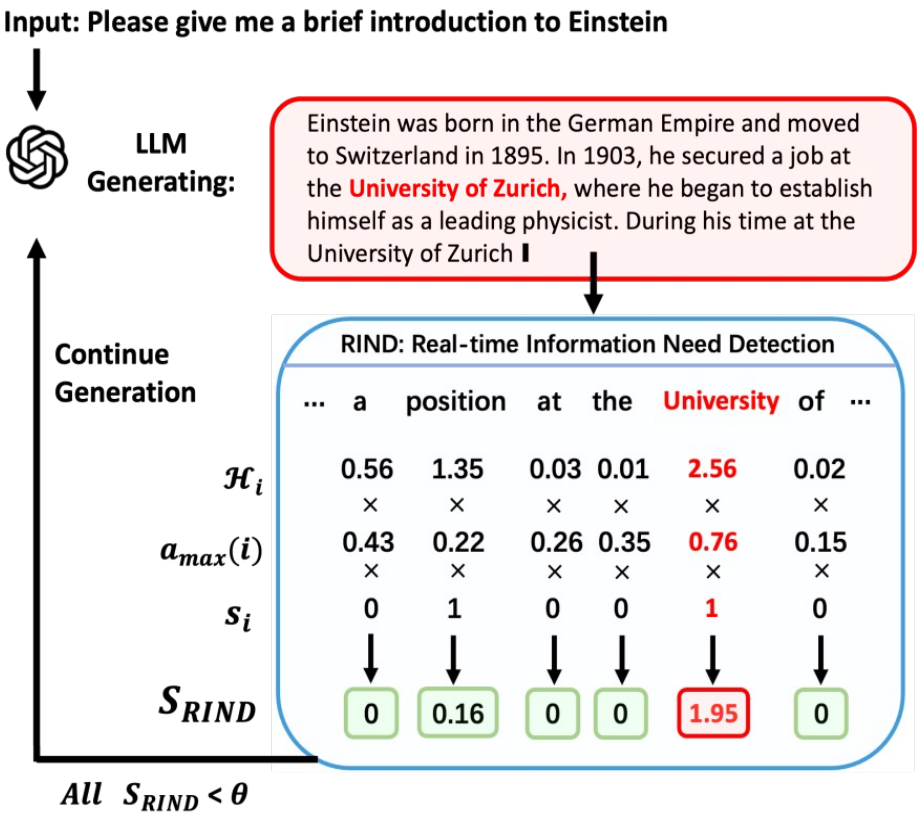
DRAGIN

QUERY-SIDE — FILTERING THE REAL NEED

> Not every uncertain token is a real need — stop words and pronouns are uncertain but meaningless.

What worth querying?

- High entropy
- High attention
- Significant (not a stopword/pronoun)



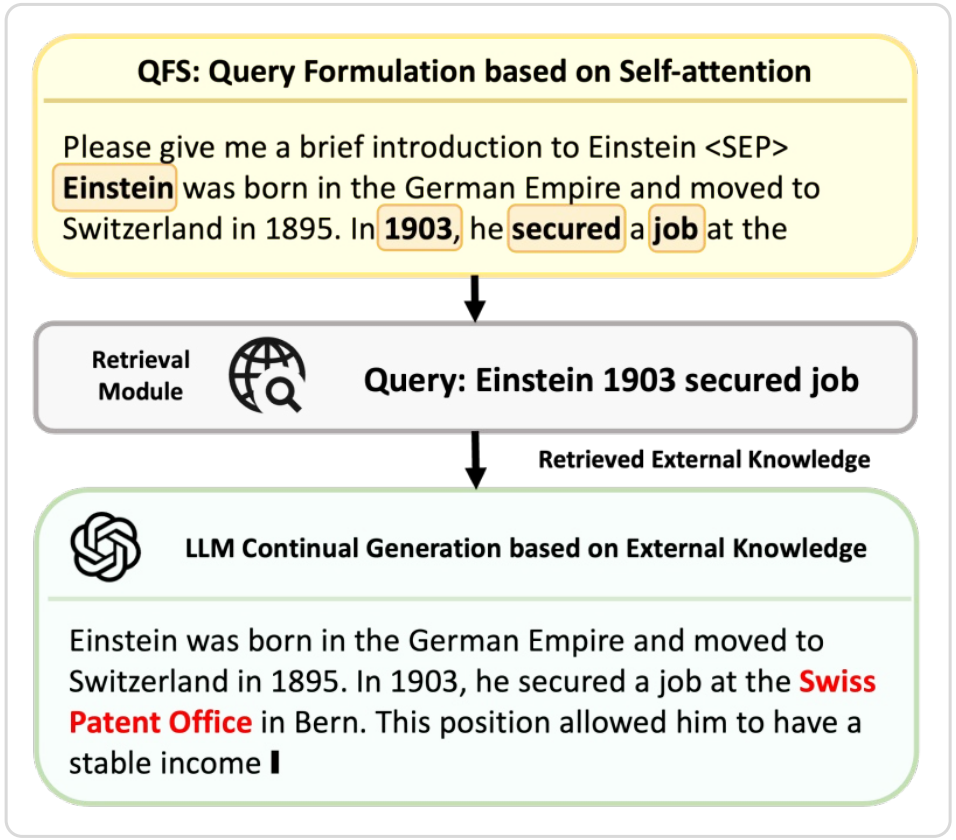
Su et al. DRAGIN: Dynamic Retrieval Augmented Generation based on the Real-time Information Needs of LLMs. ACL, 2024.

DRAGIN — What Query to Issue

WHAT QUERY?

Take the last layer‘s **attention** scores, keep the **top-n** most-attended tokens as the query

> The query reflects **what the model is actually attending to** — not the whole generated sentence.



Su et al. DRAGIN: Dynamic Retrieval Augmented Generation based on the Real-time Information Needs of LLMs. ACL, 2024.

Search-R1 — Learned Search Actions

The LLM learns to **write its own query** — no hand-designed rule.

What (query-side)

- it composes the query in <search>.

When (action-side)

- it also emits <search> itself.

THE ROLLOUT FORMAT

```
<think> reasoning </think>
<search> query </search>
<information> docs </information>
<answer> final answer </answer>
```

Jin et al. Search-R1: Training LLMs to Reason and Leverage Search Engines with Reinforcement Learning. COLM, 2025.

Search-R1 — Learned Search Actions

The LLM learns to **write its own query** — no hand-designed rule.

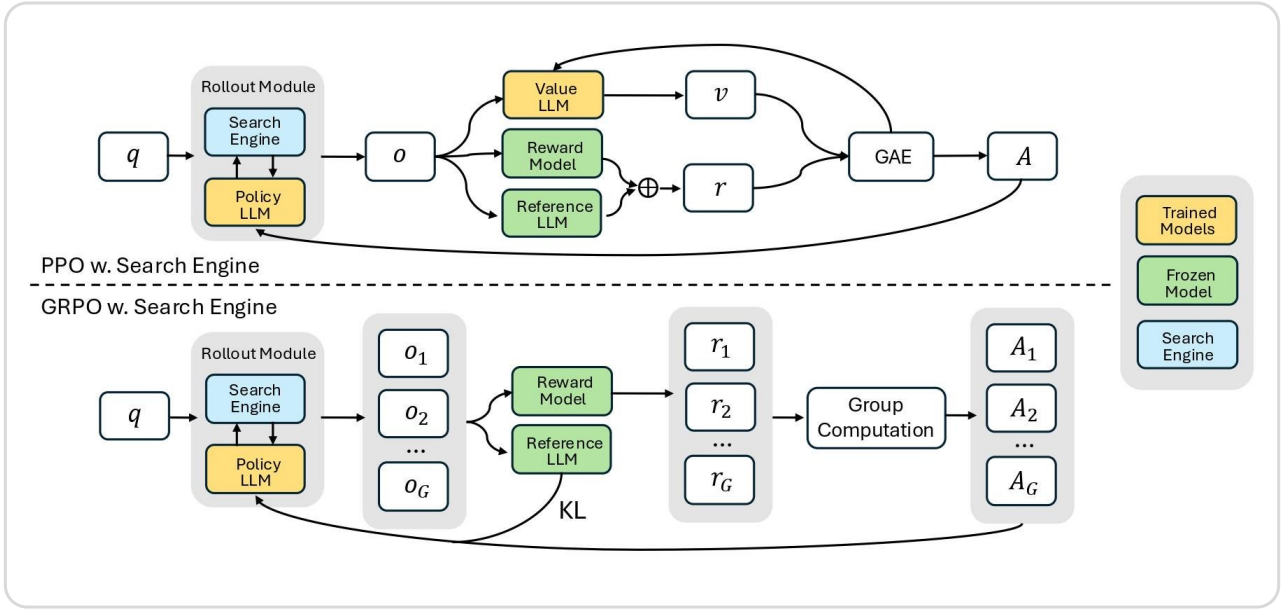
→ **learned through an end-to-end RL policy** (PPO / GRPO) (trajectory reward)

What (query-side)

- it composes the query in <search>.

When (action-side)

- it also emits <search> itself.



Jin et al. Search-R1: Training LLMs to Reason and Leverage Search Engines with Reinforcement Learning. COLM, 2025.

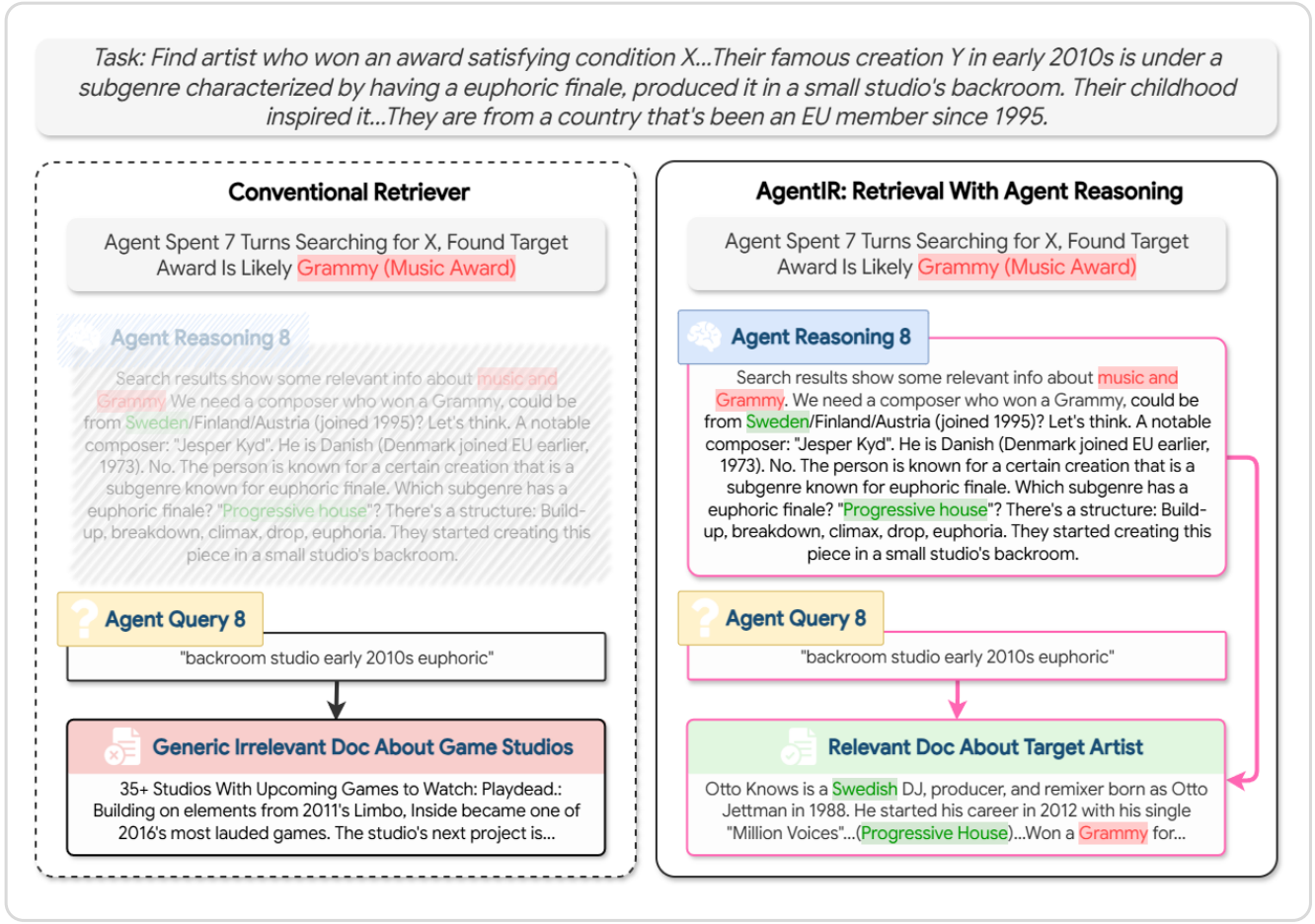
AgentIR

THE PROBLEM

An agent query is often **under-specified**. — the reasoning chain that produced it carries the **real search intent**

> Jointly embed the query and the agent's **reasoning trace alongside its query**

A type of query expansion with more specified context



Chen et al. AgentIR: Reasoning-Aware Retrieval for Deep Research Agents. arXiv, 2026.

Query-side Utility — Recap

A query is useful if it retrieves evidence that:

- fills current **knowledge gaps**
- supports the **next reasoning step**
- resolves **ambiguity or contradictions**
- **verifies** a claim
- improves the **final answer**

> **Query rewriting = utility optimization —** beyond lexical match, toward evidence acquisition.
> And the query to issue can be learned by RL over the trajectory.

The Three Utilities — Map

DOCUMENT-SIDE UTILITY

Query-side and document-side utility are coupled —
a query is useful because of the documents it brings back.

good queries → **relevant documents** → better reasoning

Document-side Utility

THE RETRIEVER OBJECTIVE SHIFTS

Document relevance ≠ document utility in agentic search.

Revisiting Text Ranking

A relevant document may pack 8K tokens — but only **~200** serve the current step.

> **Passage-level units** consistently beat document-level — **higher utility density**, more rounds, no information loss.

Retriever	Passage corpus				Document corpus			
	#Search	Recall	Acc.	Comp.	#Search	Recall	Acc.	Comp.
BM25	30.97	**0.616**	**0.572**	0.968	32.22	0.366	0.259	0.805
SPLADE-v3	32.75	0.545	0.516	0.980	28.95	<u>0.628</u>	<u>0.476</u>	0.824
RepLLaMA	36.13	0.449	0.406	0.961	30.69	0.514	0.363	0.786
Qwen3-Embed-8B	37.83	0.470	0.417	0.963	30.14	0.570	0.421	0.821
ColBERTv2	32.88	<u>0.552</u>	<u>0.521</u>	0.978	27.13	**0.633**	**0.481**	0.855

Meng et al. Revisiting Text Ranking in Deep Research. SIGIR, 2026.

Revisiting Text Ranking

A re-ranker lifts **recall** and **accuracy** while cutting the number of search calls.

> Re-ranking restores per-token **utility density** before the passages reach the agent.

Re-ranker	BM25				SPLADE-v3				Qwen3-Embed-8B			
	#Search	Recall	Acc.	Comp.	#Search	Recall	Acc.	Comp.	#Search	Recall	Acc.	Comp.
no re-ranking	30.97	0.616	0.572	0.980	32.75	0.545	0.516	0.980	37.83	0.470	0.417	0.963
monoT5 ($d = 10$)	29.89	0.660	0.631	0.971	31.58	0.630	0.599	0.965	35.64	0.524	0.471	0.969
monoT5 ($d = 20$)	28.96	0.701	0.674	0.960	30.07	0.647	0.598	0.974	34.16	0.570	0.541	0.959
monoT5 ($d = 50$)	27.57	**0.716**	**0.689**	0.974	29.72	0.689	0.646	0.971	32.94	0.614	0.559	0.952
RankLLaMA ($d = 10$)	29.85	0.657	0.617	0.980	31.30	0.632	0.595	0.964	36.07	0.508	0.465	0.964
RankLLaMA ($d = 20$)	29.03	0.681	0.655	0.971	30.82	0.646	0.605	0.981	34.25	0.553	0.494	0.961
RankLLaMA ($d = 50$)	27.10	0.710	0.678	0.977	28.96	0.691	**0.663**	0.959	32.85	0.613	**0.568**	0.964

A 20B model + a 3B re-ranker matches a full GPT-5 agent — at a fraction of the cost.

Meng et al. Revisiting Text Ranking in Deep Research. SIGIR, 2026.

Revisiting Text Ranking

Agent queries follow **web-search style**: keywords, quoted phrases, exact strings.

```
"90'+7" attendance 61700
```

```
"90'+4" football match attendance
```

BM25 outperforms neural retrievers:

Retriever	Passage corpus				Document corpus			
	#Search	Recall	Acc.	Comp.	#Search	Recall	Acc.	Comp.
BM25	30.97	**0.616**	**0.572**	0.968	32.22	0.366	0.259	0.805
SPLADE-v3	32.75	0.545	0.516	0.980	28.95	<u>0.628</u>	<u>0.476</u>	0.824
RepLLaMA	36.13	0.449	0.406	0.961	30.69	0.514	0.363	0.786
Qwen3-Embed-8B	37.83	0.470	0.417	0.963	30.14	0.570	0.421	0.821
ColBERTv2	32.88	<u>0.552</u>	<u>0.521</u>	0.978	27.13	**0.633**	**0.481**	0.855

Meng et al. Revisiting Text Ranking in Deep Research. SIGIR, 2026.

Revisiting Text Ranking

FIXING THE MISMATCH FOR DENSE

Keyword queries starve dense encoders. **Q2Q** rewrites them into the **questions** they trained on.

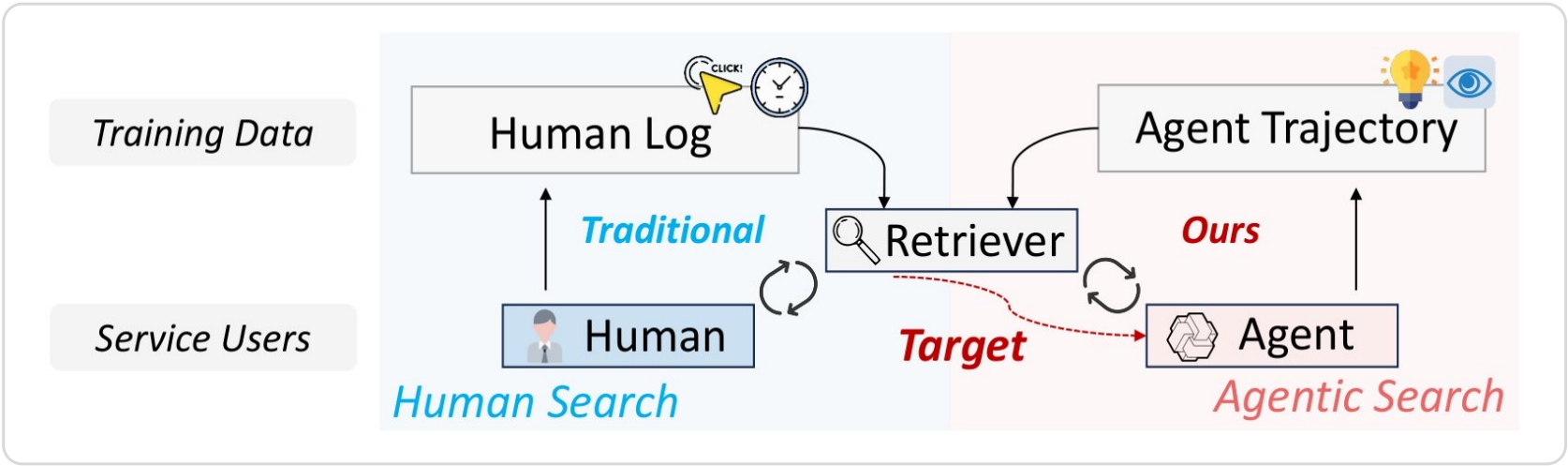
raw: "61,880" football attendance
Q2Q: What match drew an attendance of 61,880?

Retriever	Query	#Search	Recall	Acc.	Comp.
BM25	Raw	30.97	**0.616**	0.572	0.968
	Q2Q (Q)	31.88	0.593	**0.578**	0.977
	Q2Q (Q+R)	32.15	0.583	0.557	0.974
SPLADE-v3	Raw	32.75	0.545	0.516	0.980
	Q2Q (Q)	32.88	0.550	0.510	0.976
	Q2Q (Q+R)	31.70	**0.585***	**0.557***	0.983
Qwen3 -Embed-8B	Raw	37.83	0.470	0.417	0.963
	Q2Q (Q)	37.42	0.457	0.404	0.969
	Q2Q (Q+R)	35.82	**0.507***	**0.459***	0.965

Meng et al. Revisiting Text Ranking in Deep Research. SIGIR, 2026.

LRAT

Human search logs give retrievers one cheap signal — a **click = a positive**. Agents don't click; they issue a **[Browse]** action — LRAT reads it as the agent's click.



Browsed → positive

the agent opened it — a naive positive.

Unbrowsed → negative

left unopened in the same set — a naive negative.

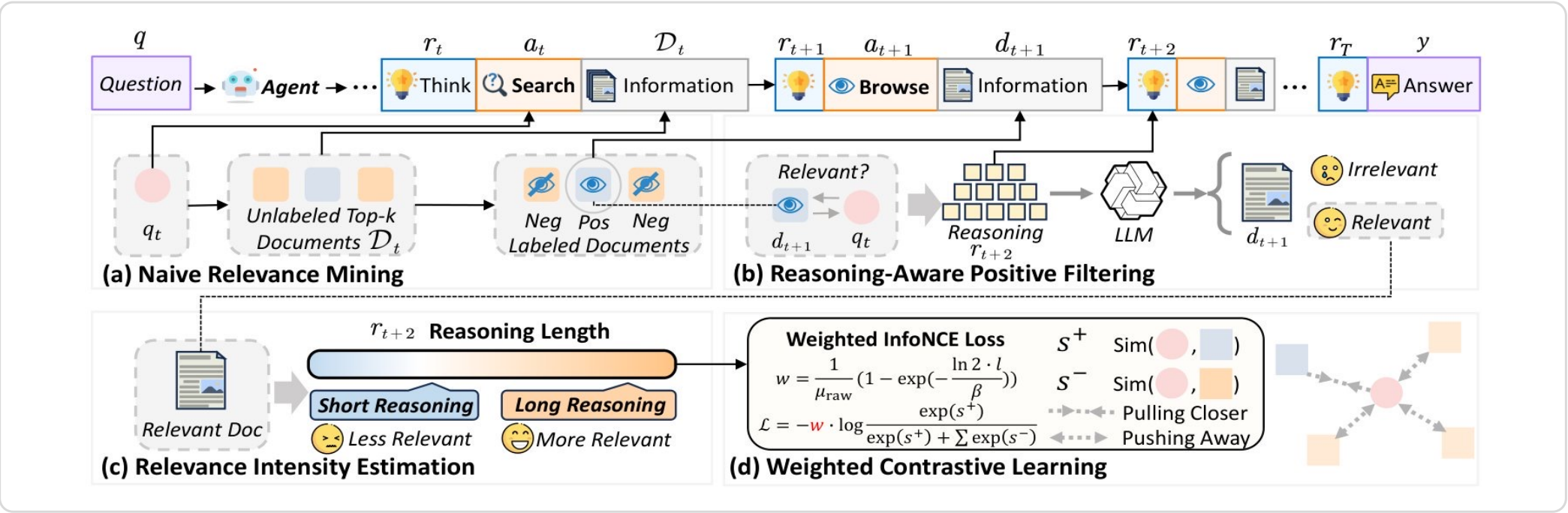
Zhou et al. Learning to Retrieve from Agent Trajectories. SIGIR, 2026.

LRAT

Naive browse labels are noisy — **browsed ≠ actually useful**.

LRAT refines them two ways:

① **Verifier** — the reasoning must **use** the doc. ② **Intensity** — **longer reasoning ⇒ more useful**.



Zhou et al. Learning to Retrieve from Agent Trajectories. SIGIR, 2026.

LRAT

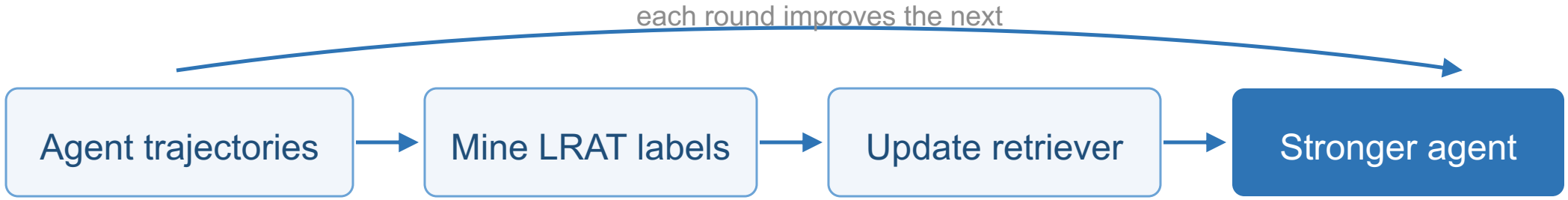
Agent Backbone	Retriever	InfoSeek-Eval (ID)		BrowseComp-Plus (OOD)		
		SR (↑)	Avg. Steps (↓)	SR (↑)	Recall (↑)	Avg. Steps (↓)
I. Task-Optimized Search Agents						
AgentCPM-Explore (4B)	Qwen3-Emb	40.3	38.0	13.5	23.2	40.7
	+ LRAT (Ours)	**55.7** (+38.2%)	**34.4**	**15.8** (+17.0%)	**32.0** (+37.9%)	**40.4**
	E5-Large	47.3	38.9	15.9	26.5	40.7
	+ LRAT (Ours)	**49.7** (+5.1%)	**35.5**	**15.9** (+0.0%)	**32.1** (+21.1%)	**40.1**
WebExplore (8B)	Qwen3-Emb	52.0	24.1	21.0	47.7	40.7
	+ LRAT (Ours)	**68.7** (+32.1%)	**19.0**	**27.2** (+29.5%)	**55.9** (+17.2%)	**38.7**
	E5-Large	60.0	23.8	25.4	50.4	40.1
	+ LRAT (Ours)	**63.3** (+5.5%)	**20.2**	**29.0** (+14.2%)	**56.1** (+11.3%)	**39.1**
Tongyi-DeepResearch (30B)	Qwen3-Emb	52.7	26.7	17.8	49.2	42.9
	+ LRAT (Ours)	**68.0** (+29.0%)	**20.7**	**23.7** (+33.1%)	**60.7** (+23.4%)	**41.0**
	E5-Large	56.7	25.1	20.7	54.8	42.4
	+ LRAT (Ours)	**68.0** (+19.9%)	**21.5**	**23.9** (+15.5%)	**61.8** (+12.8%)	**41.4**
II. Generalist Agentic Foundation Models						
GPT-OSS (120B)	Qwen3-Emb	40.0	34.9	9.0	43.7	45.4
	+ LRAT (Ours)	**47.0** (+17.5%)	**30.5**	**12.1** (+34.4%)	**56.4** (+29.1%)	**45.2**
	E5-Large	41.7	33.9	10.8	50.1	44.8
	+ LRAT (Ours)	**50.7** (+21.6%)	**29.7**	**13.1** (+21.3%)	**56.0** (+11.8%)	**44.6**

Higher **success rate** & recall, fewer search steps — ID & OOD.

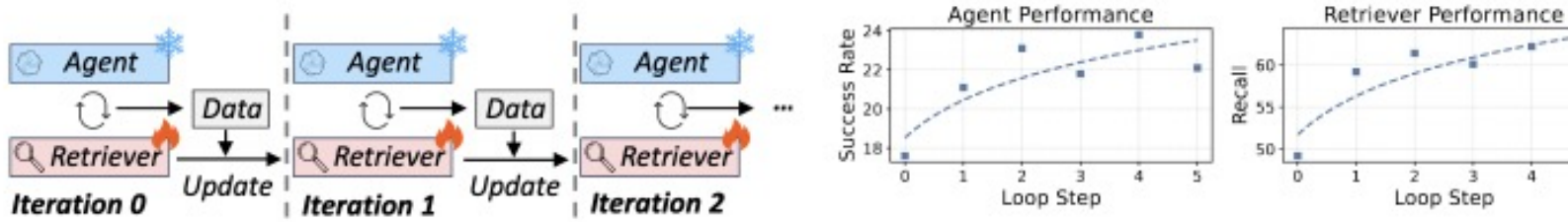
Zhou et al. Learning to Retrieve from Agent Trajectories. SIGIR, 2026.

LRAT — The Data Flywheel

Labels → training → a stronger agent → better labels — a **positive data flywheel**.



FLYWHEEL SIMULATION

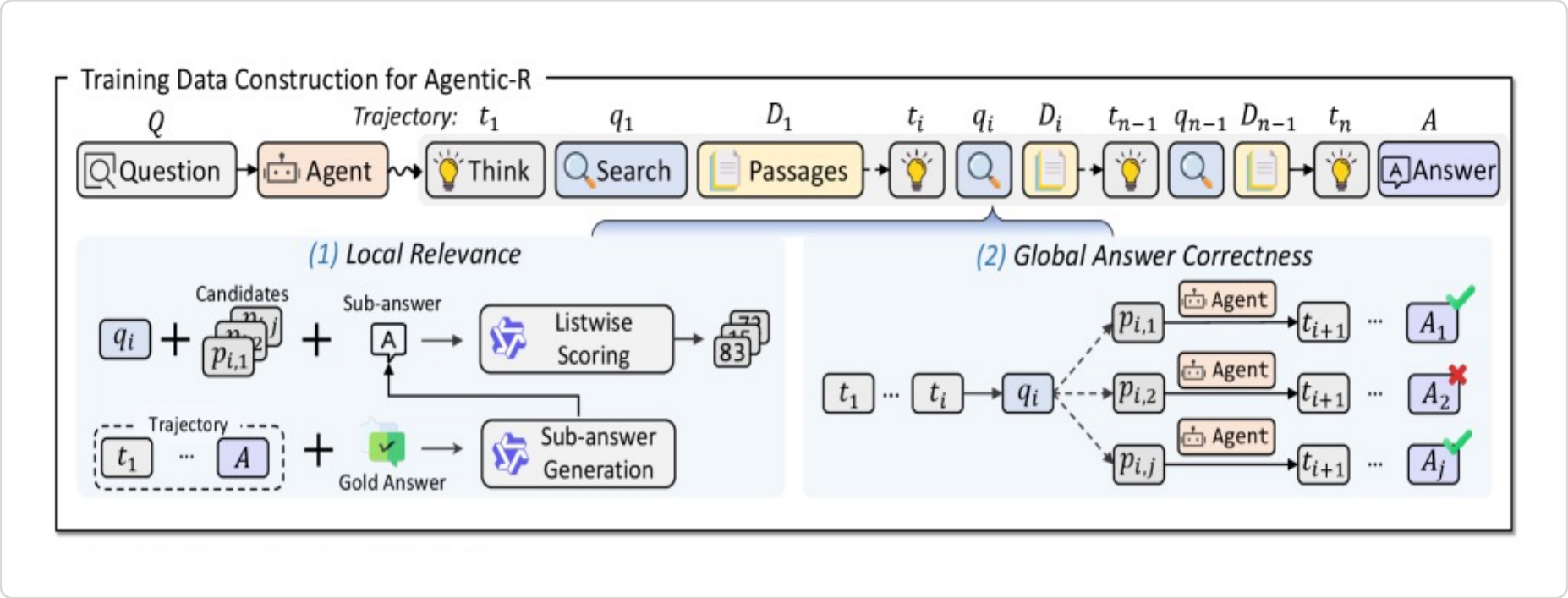


(a) Illustration of data flywheel simulation setting. (b) Performance of the data flywheel simulation.

Zhou et al. Learning to Retrieve from Agent Trajectories. SIGIR, 2026.

Agentic-R

Intermediate queries have **no gold answers** — single-turn utility signals don‘t apply, and relevance could be insufficient to serve the agent reasoning towards the final answer. -> utility label for passages



Liu et al. Agentic-R: Learning to Retrieve for Agentic Search. ACL Findings, 2026.

Agentic-R — Dual-Signal Utility

With no gold answer per turn, one signal isn't enough — Agentic-R scores on **two**.

Local Relevance

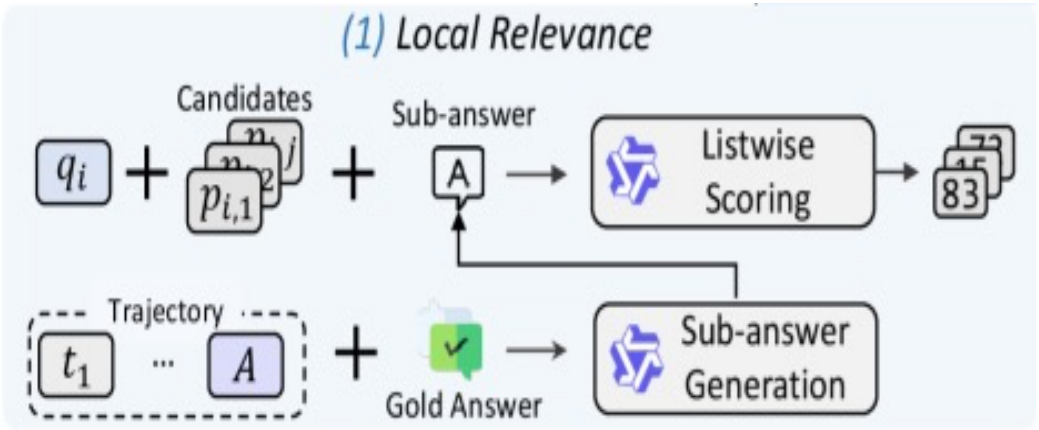
Does the passage address the **current sub-query**?

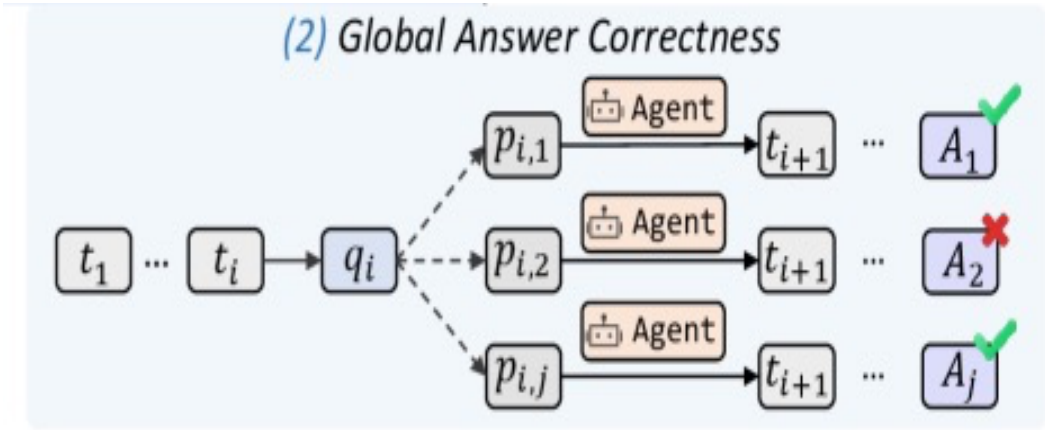
scored **0–100** by a strong LLM judge

Global Answer Correctness

Roll the agent forward on this passage — does it reach the **gold answer**?

checked against the ground truth **(0 / 1)**





Liu et al. Agentic-R: Learning to Retrieve for Agentic Search. ACL Findings, 2026.

Agentic-R — Iterative Optimization

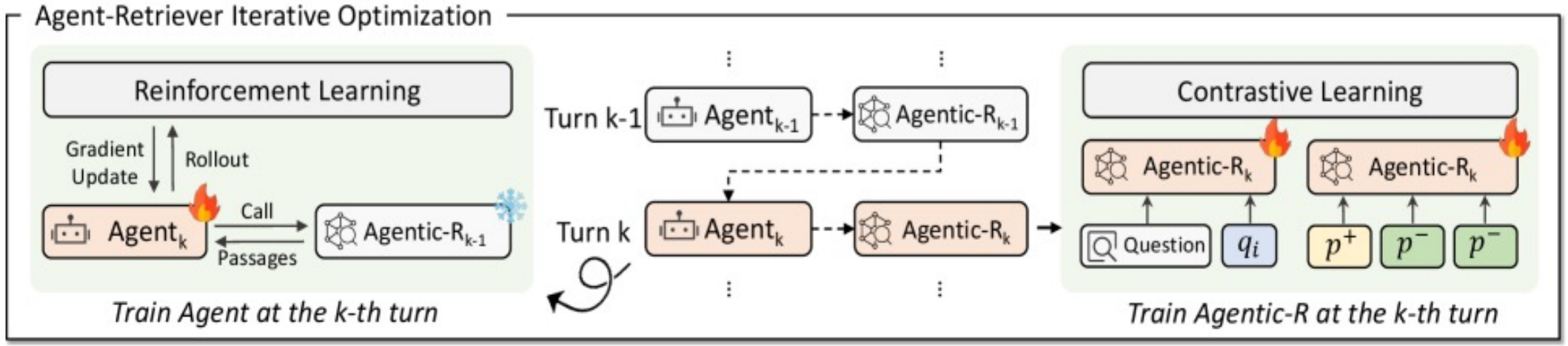
Agent and retriever train in an **alternating loop** — each freezes while the other learns.

1 Train Agent · retriever frozen

PPO with **EM on the final answer** as reward; the agent interacts to gather passages.

2 Train Retriever · agent frozen

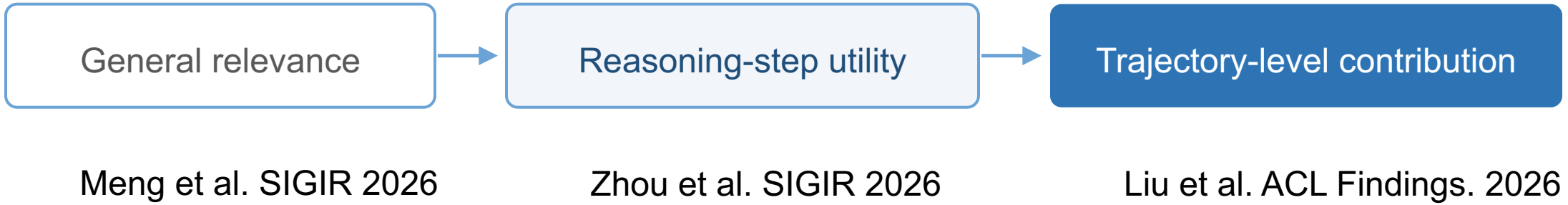
The improved agent yields better trajectories; **dual-signal labels** build its training data.



Liu et al. Agentic-R: Learning to Retrieve for Agentic Search. ACL Findings, 2026.

Document-side Utility — Recap

THE RETRIEVER OBJECTIVE SHIFTS



Document relevance ≠ document utility.

Return the evidence — at the right granularity — that best serves the current state.

The Three Utilities — Map

THE ACTION SPACE

```
retrieve? / browse / rewrite / verify / use tool / stop
```

The agent picks the action that best serves its need — including *whether/when to retrieve* — even bypassing the retriever.

Two questions: **which** action next, and **how** to reward it.

When to Retrieve

The first action-side call: retrieve now, or keep reasoning? Many state signals can trigger it.

SIGNAL OF NEED		OPERATIONALIZED BY
Low-confidence tokens	→	**FLARE** (Jiang et al., 2023)
Entropy · attention · stopwords	→	**DRAGIN** (Su et al., 2024)
Self-issued search actions	→	**Search-R1** (Jin et al., 2025)
Verifier / self-reflection failures	→	**Self-RAG** (Asai et al., 2024)
Contradictory evidence	→	**Conflicting Evidence** (Wang et al., 2024)

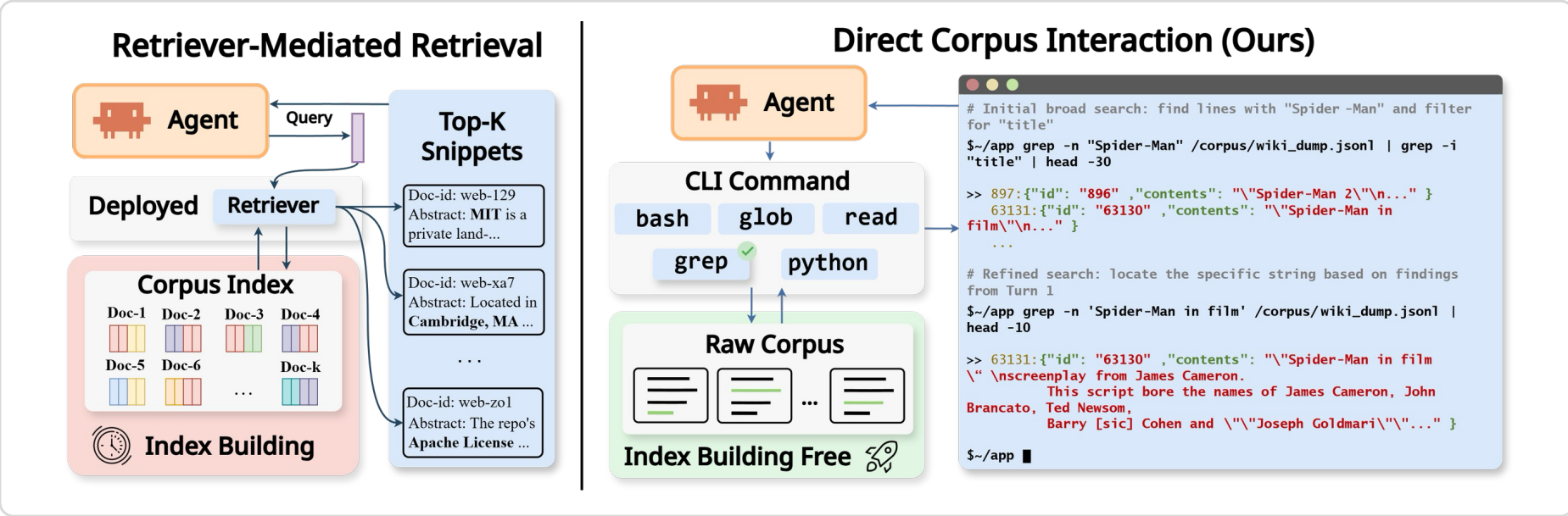
Direct Corpus Interaction

What about corpus access beyond retrieval?
an **action the agent chooses** — search raw files with **grep · rg · find**.

A fixed **top-k API** is just one hard-wired retrieval action. As agents get smarter, that interface — not their reasoning — becomes the bottleneck.

Li et al. Beyond Semantic Similarity: Retrieval for Agentic Search via Direct Corpus Interaction. arXiv, 2026.

Direct Corpus Interaction

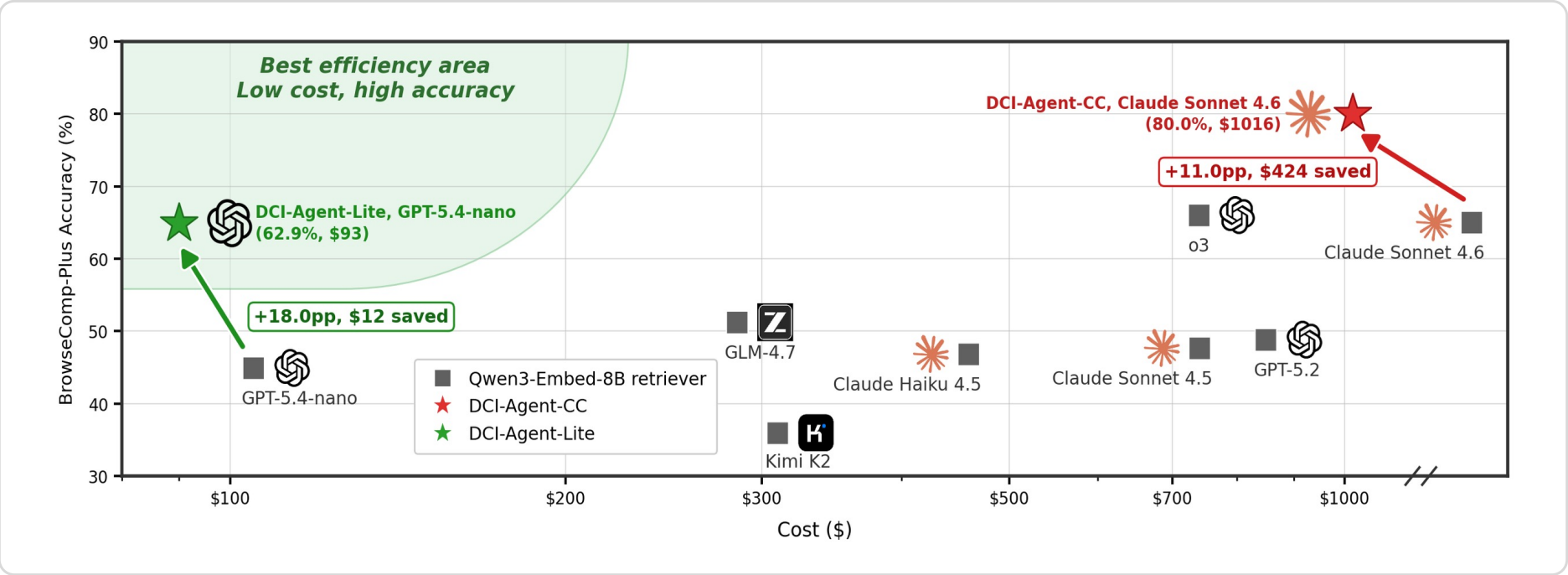


The far end of the action space — **bypass the retriever entirely**.

- **Zero-index** — no vector DB; adapts to new corpora instantly.
- **High-resolution control** — regex, joins, exact matches.

Li et al. Beyond Semantic Similarity: Retrieval for Agentic Search via Direct Corpus Interaction. arXiv, 2026.

Direct Corpus Interaction



Higher efficiency; higher accuracy

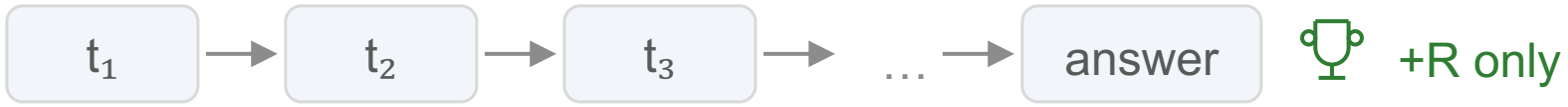
We've won 1st place on the multi-table QA subtask of NTCIR DAGRI task based on DCI!

Li et al. Beyond Semantic Similarity: Retrieval for Agentic Search via Direct Corpus Interaction. arXiv, 2026.

Intermediate Rewards

If the agent chooses among many actions, how does it learn which is right at each step?

Final-answer reward is sparse



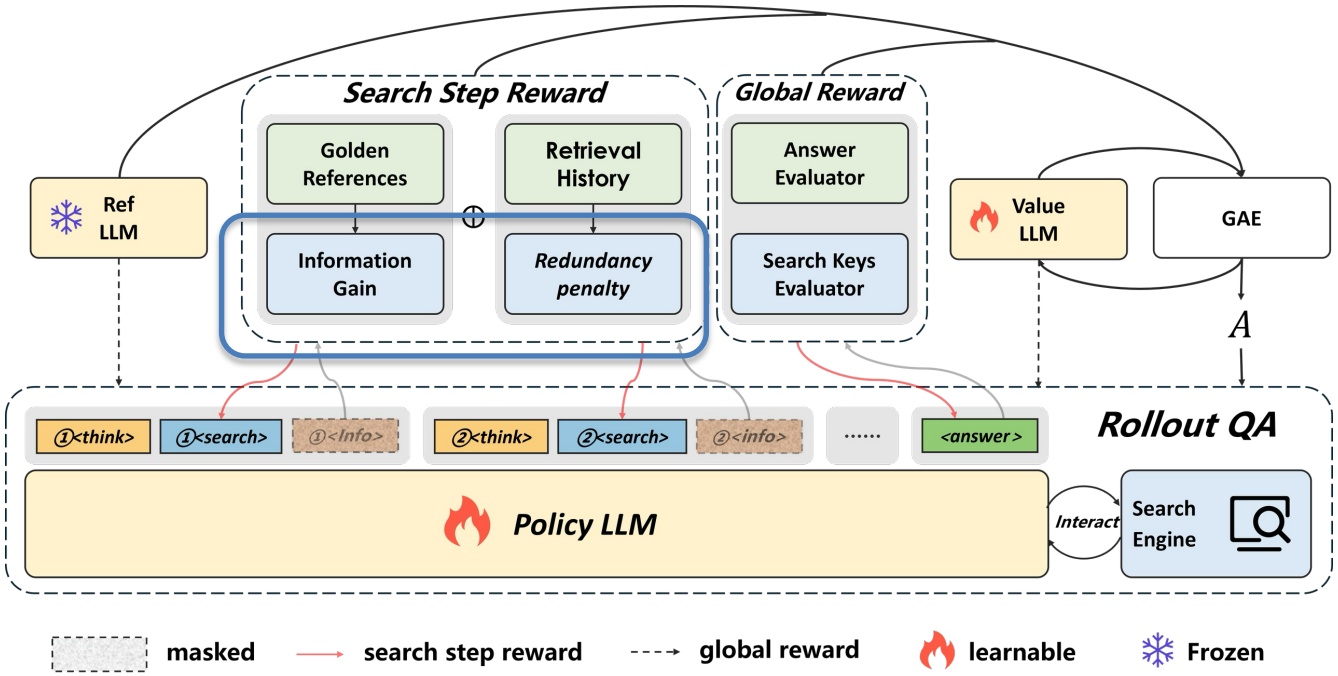
Intermediate rewards — dense per-step feedback

Translate utility metrics into intermediate rewards to guide step-by-step learning.

StepSearch

StepSearch rewards **each search step**, not just the final answer.

Global Reward = answer F1 + query alignment with ground-truth subqueries.
Step Reward = information gain − redundancy penalty.

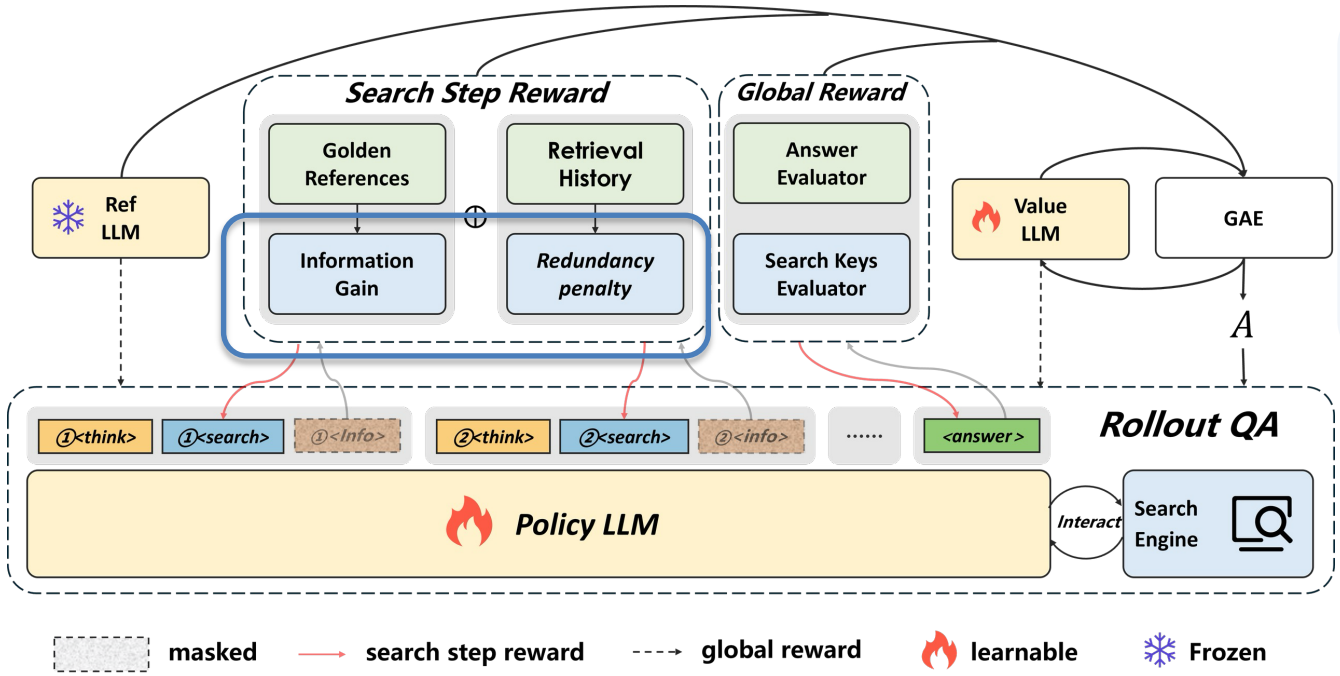


Wang et al. StepSearch: Igniting LLMs Search Ability via Step-Wise Proximal Policy Optimization. EMNLP, 2025.

StepSearch

StepSearch rewards **each search step**, not just the final answer.

Global Reward = answer F1 + query alignment with ground-truth subqueries.
Step Reward = information gain − redundancy penalty.



Information gain

- average marginal gain compared to the best historical match over n ground-truth docs

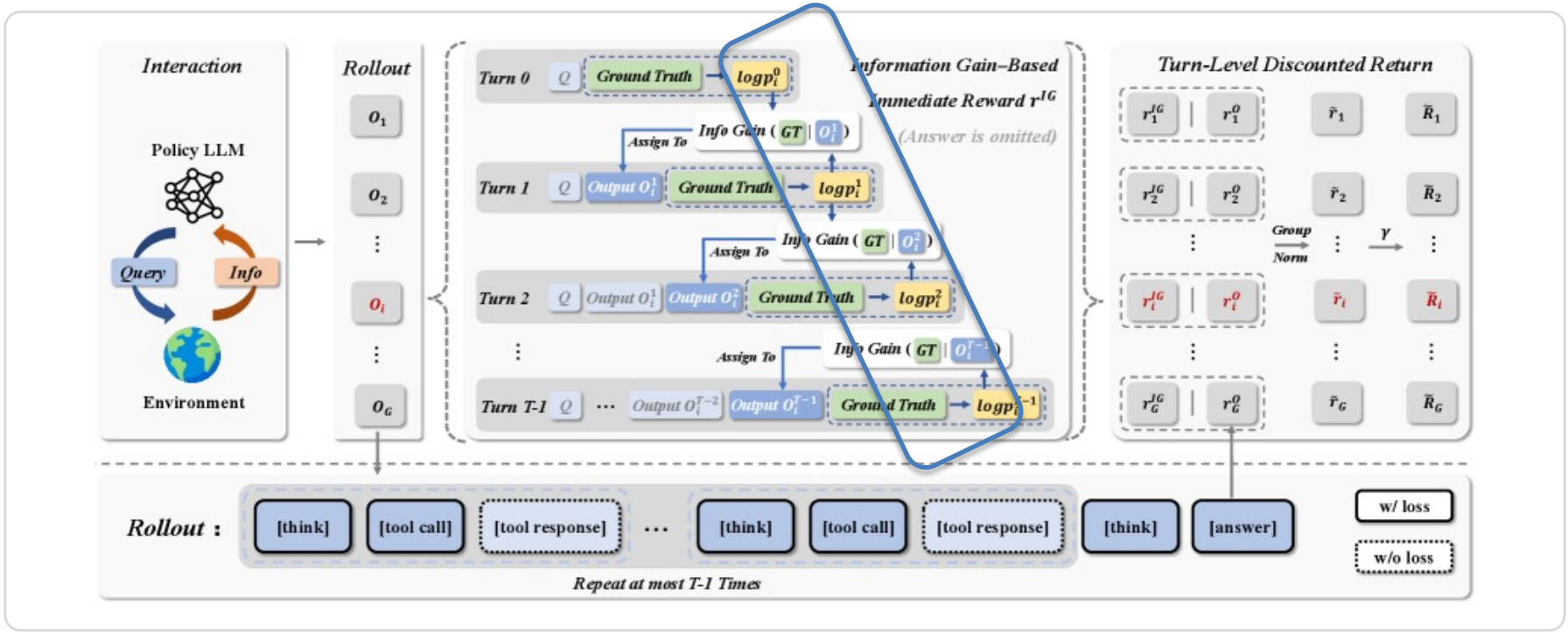
Redundancy Penalty

- fraction of retrieved docs already in the history

Wang et al. StepSearch: Igniting LLMs Search Ability via Step-Wise Proximal Policy Optimization. EMNLP, 2025.

IGPO — Belief Gain per Turn

IGPO rewards each turn by the **marginal rise** in the policy's own probability of the **ground-truth answer** after search.



Wang et al. Information Gain-based Policy Optimization: A Simple and Effective Approach for Multi-Turn LLM Agents. ICLR, 2026.

IGPO

Turn-level reward = the model's own **belief gain** about the ground-truth answer.

1. Log-probability of the gold answer under the current policy
2. **Info-gain reward** = marginal change in this probability after observing the tool (search) response
3. Group-normalized turn reward — info gain reward for t < T, outcome reward at t = T

Wang et al. Information Gain-based Policy Optimization: A Simple and Effective Approach for Multi-Turn LLM Agents. ICLR, 2026.

Action-side Utility — Recap

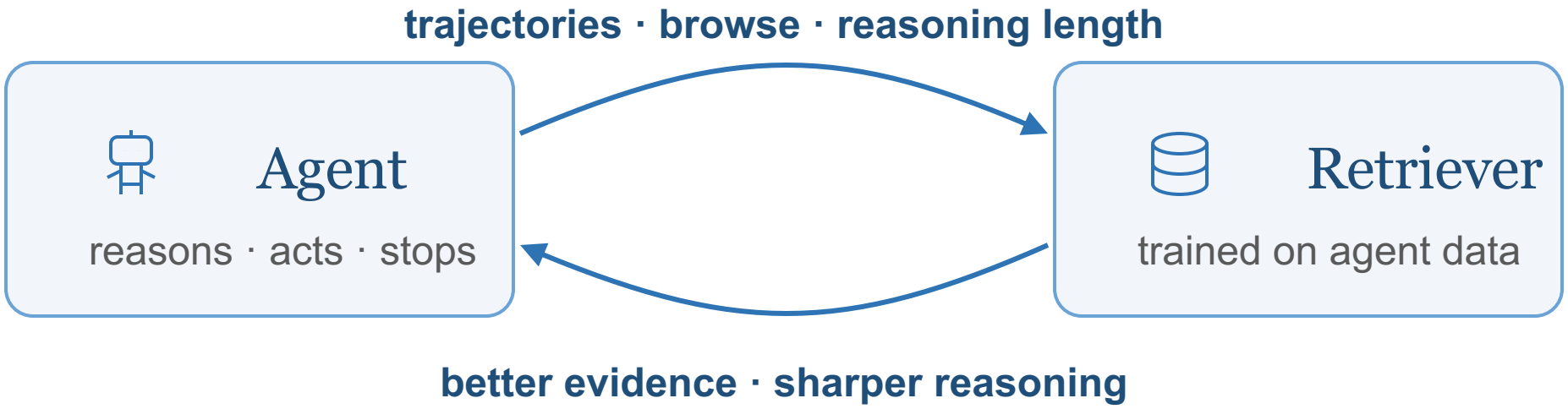
Action utility answers which action next — and how to reward it.

- choose the action that **maximizes progress** under the current state
- reward **marginal utility per step**, not just the final answer
- let **credit reach every intermediate decision**

> Sparse outcome reward → **dense, utility-shaped intermediate rewards.**

Retriever Adaptation

The retriever is not a frozen black box behind the agent.



The agent **shapes** the retriever; the retriever **sharpens** the agent.

The Three Utilities — Map

THREE PER-STATE UTILITIES

- **Query utility** — issue a state-relevant query
- **Document utility** — support the current step and the final answer
- **Action utility** — when to retrieve / browse / verify / stop / use tool

Retrieval should be optimized for trajectory utility, not only topical match.

Section 4 — Takeaways

1 **Agentic RAG makes utility state-dependent.**
Three per-state utilities — query, document, action — bound by a trajectory-level credit axis.

2 **Information need is dynamic.**
Signals decide when to retrieve (action) and what query to issue (query).

3 **Document utility is local and global.**
A passage relevant now can still hurt the final answer.

4 **Intermediate rewards turn utility into step-level signals.**
And retrievers should be trained from agent data, not human relevance.

NEXT — SECTION 5

We followed utility across a trajectory.

What's still open?

synthesis → one unifying map → open problems

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

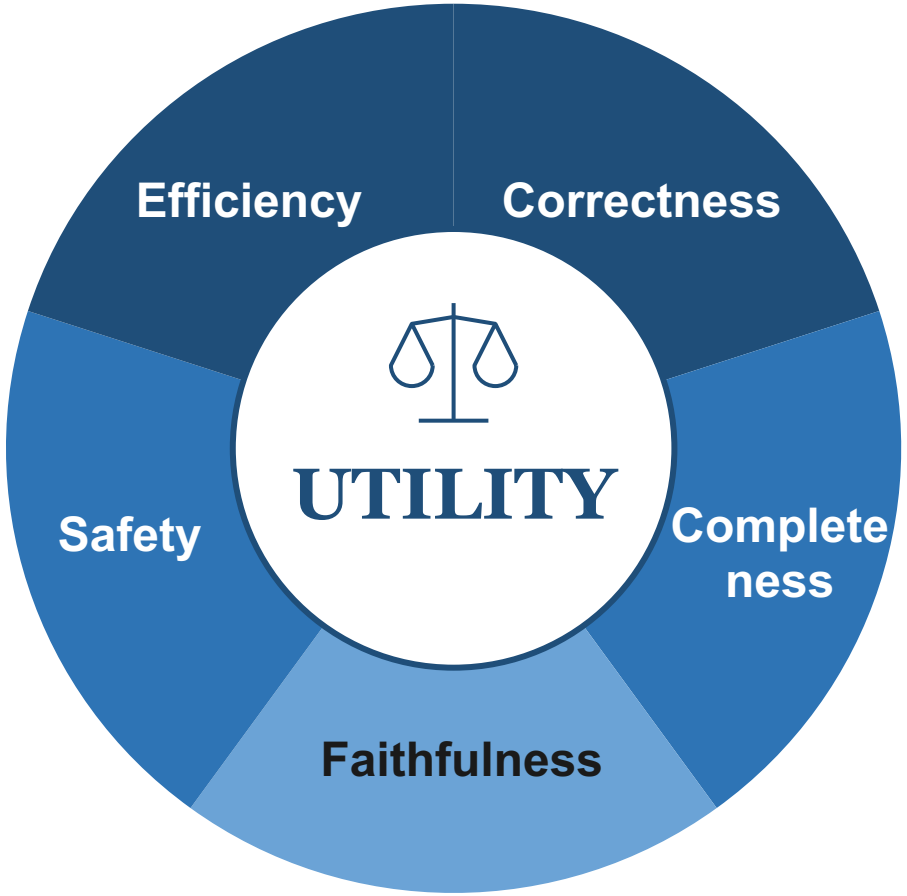
Tutorial website

SECTION 5

Open Problems

Five open problems on the path to reliable utility-centric retrieval.

Utility Definition



OPEN QUESTIONS

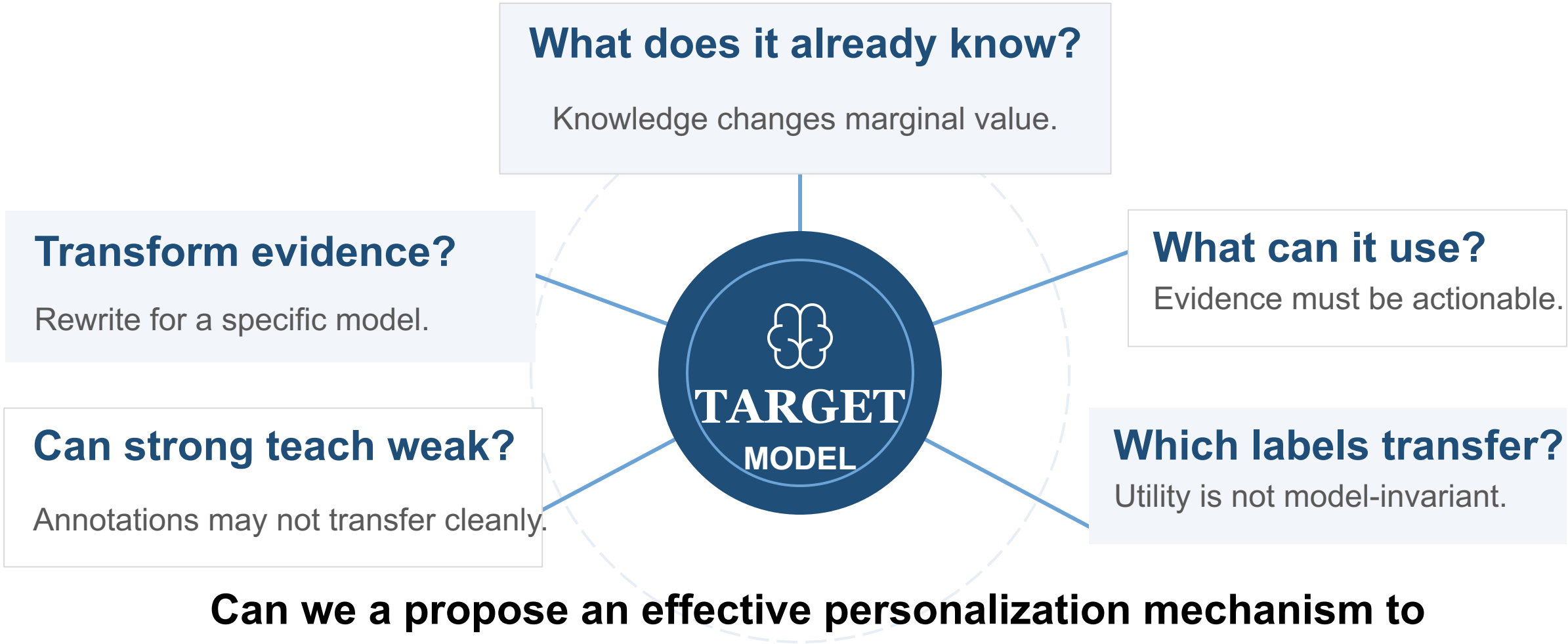
How do we define utility beyond factoid QA?

How do we compare utility across domains?

How do we represent negative utility?

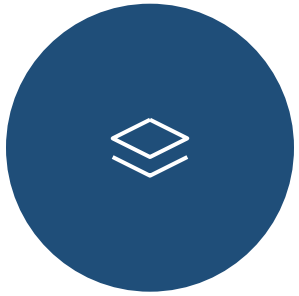
How to combine correctness, completeness, faithfulness, safety, and efficiency?

LLM-Specific Utility



Can we a propose an effective personalization mechanism to identify LLM-specific high-utility info (query & doc)?

Set & Context Utility



A budgeted, complementary, attributable set of evidence.

OPEN QUESTIONS

- How to balance coverage and complementarity?
- How to avoid redundancy and conflict?
- How to evaluate marginal contribution?
- How to assemble context under token budgets?
- How to preserve provenance and attribution?

Context utility is a composition problem; combinatorial explosion

Scalable Training & Evaluation

Scale labels, retrievers, judges, and evaluation.

OPEN QUESTIONS

How to scale high-fidelity utility labels?

How to train corpus-scale retrievers?

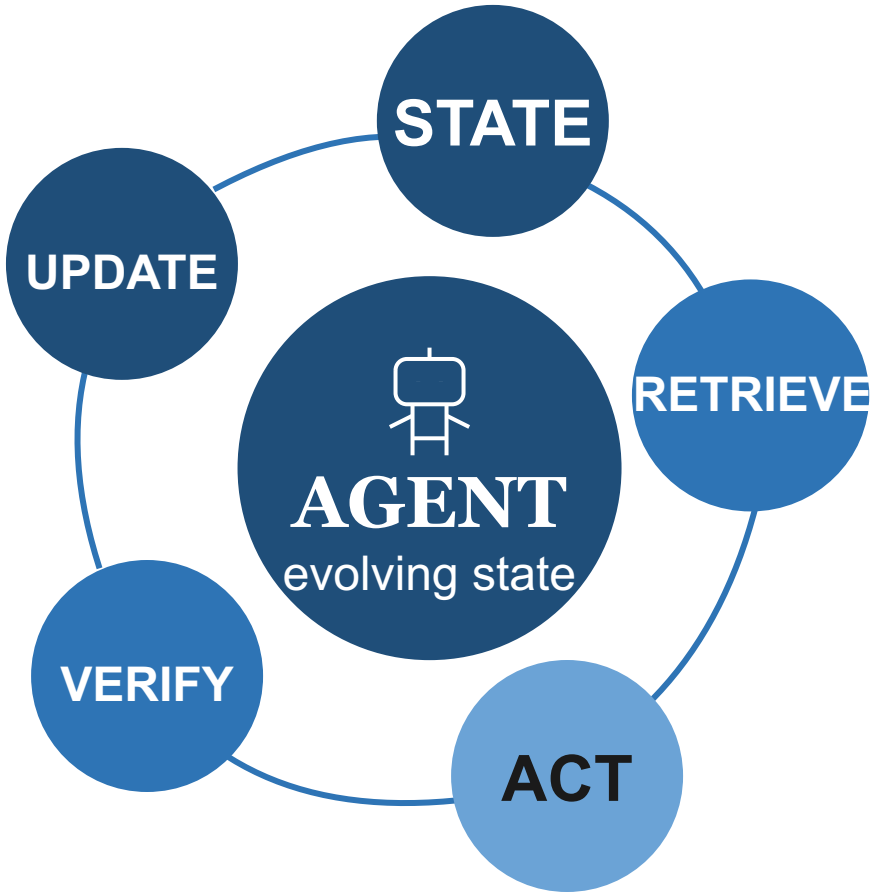
How reliable are LLM judges?

How to isolate retrieval from the generator?

How to evaluate open-ended & agentic tasks?

RESEARCH NEED **Scale labels — and isolate retrieval from generator ability.**

Agentic Retrieval



A MAJOR FRONTIER

Search agents make retrieval dynamic again— and the retriever must catch up.

Next-step or final-answer utility?
Trajectories as click logs?
Reward assignment across long trajectories
Evidence verification

Tutorial Recap

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

Thank You

Questions & Discussion

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

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

Tutorial website