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?

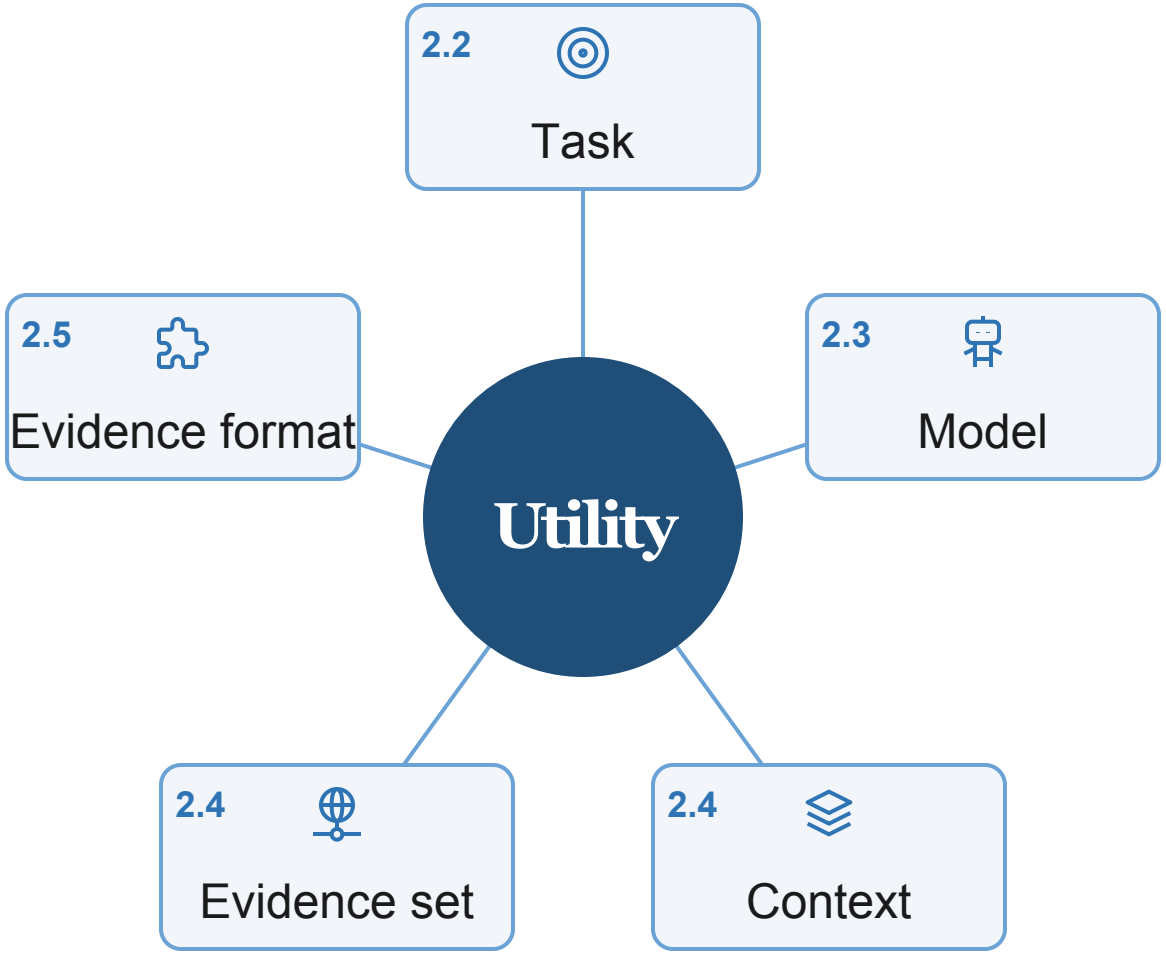
"evidence format" ↓

Format

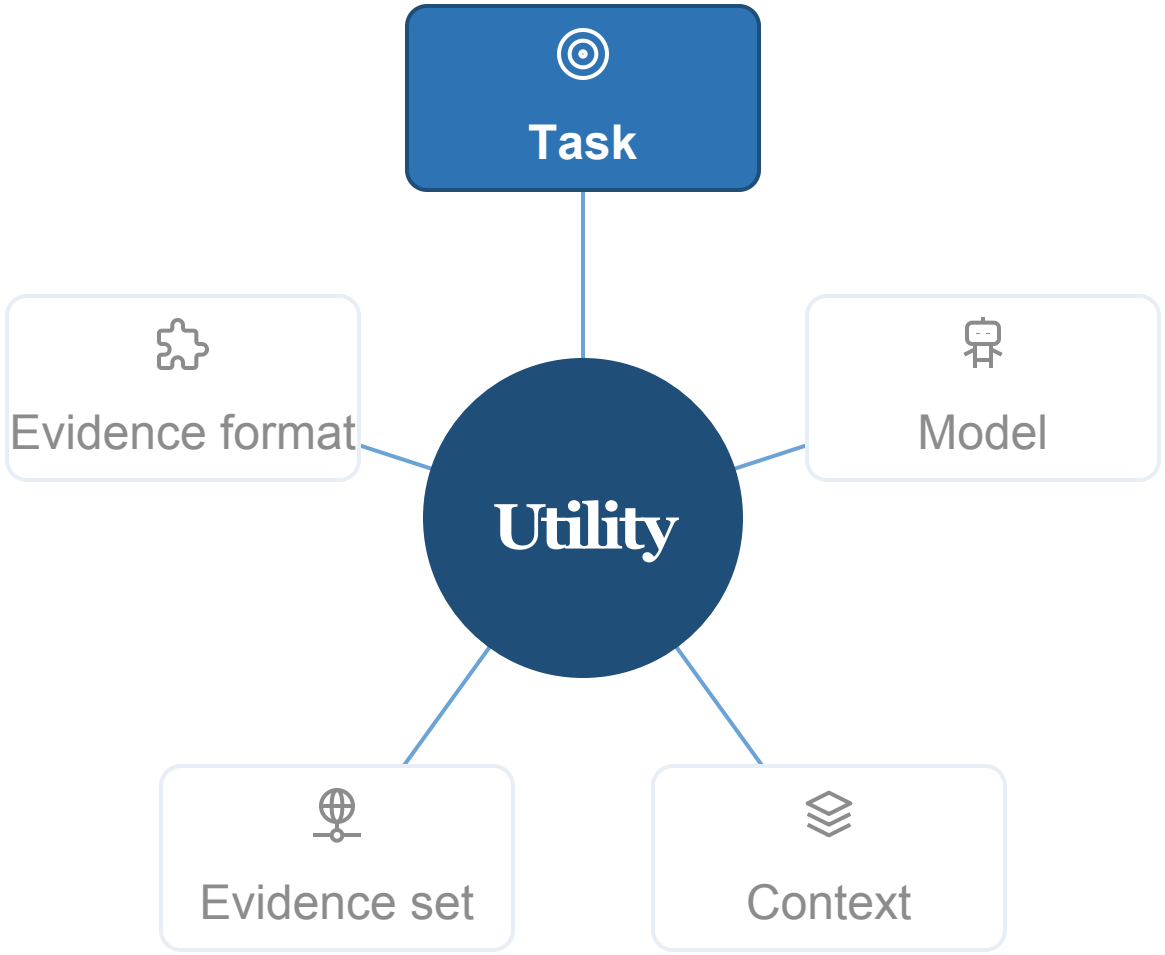
Is the form usable?

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
> - same entity, missing fact, **no utility**.

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

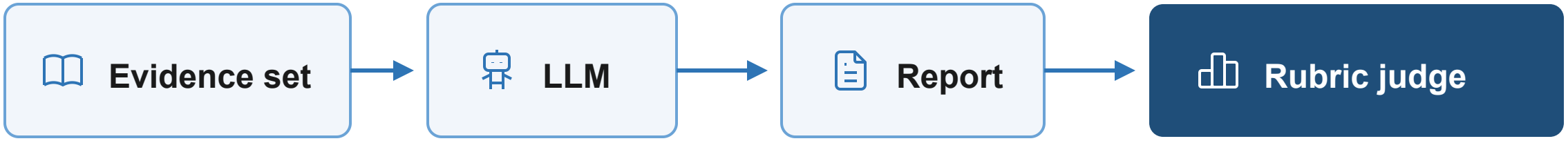
Competing views are weighed fairly.

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

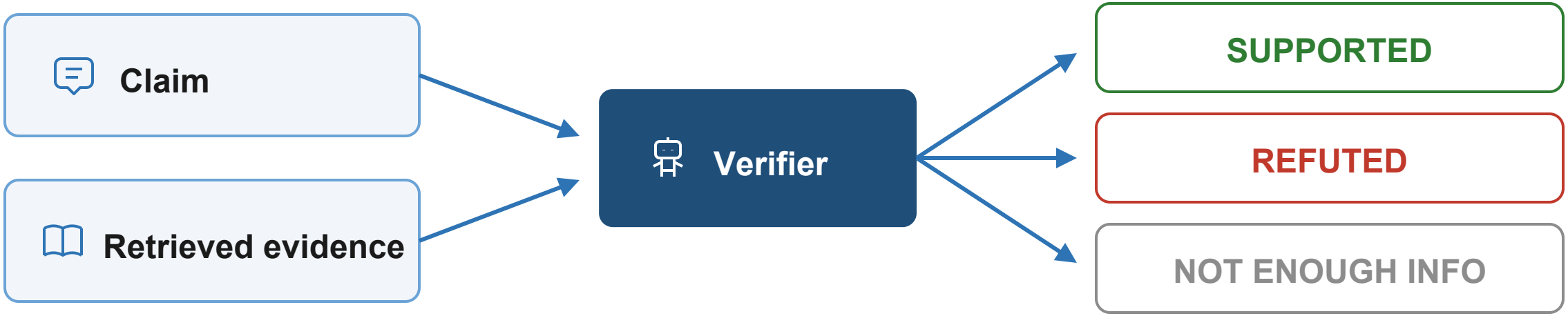
- comprehensiveness
- insight / depth
- instruction-following
- readability

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

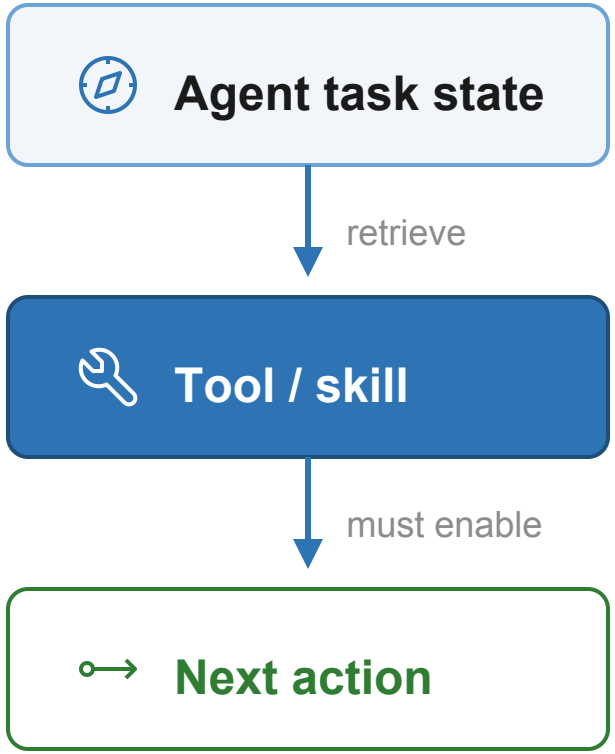
Whatever helps produce executable, context-compatible code next step.

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

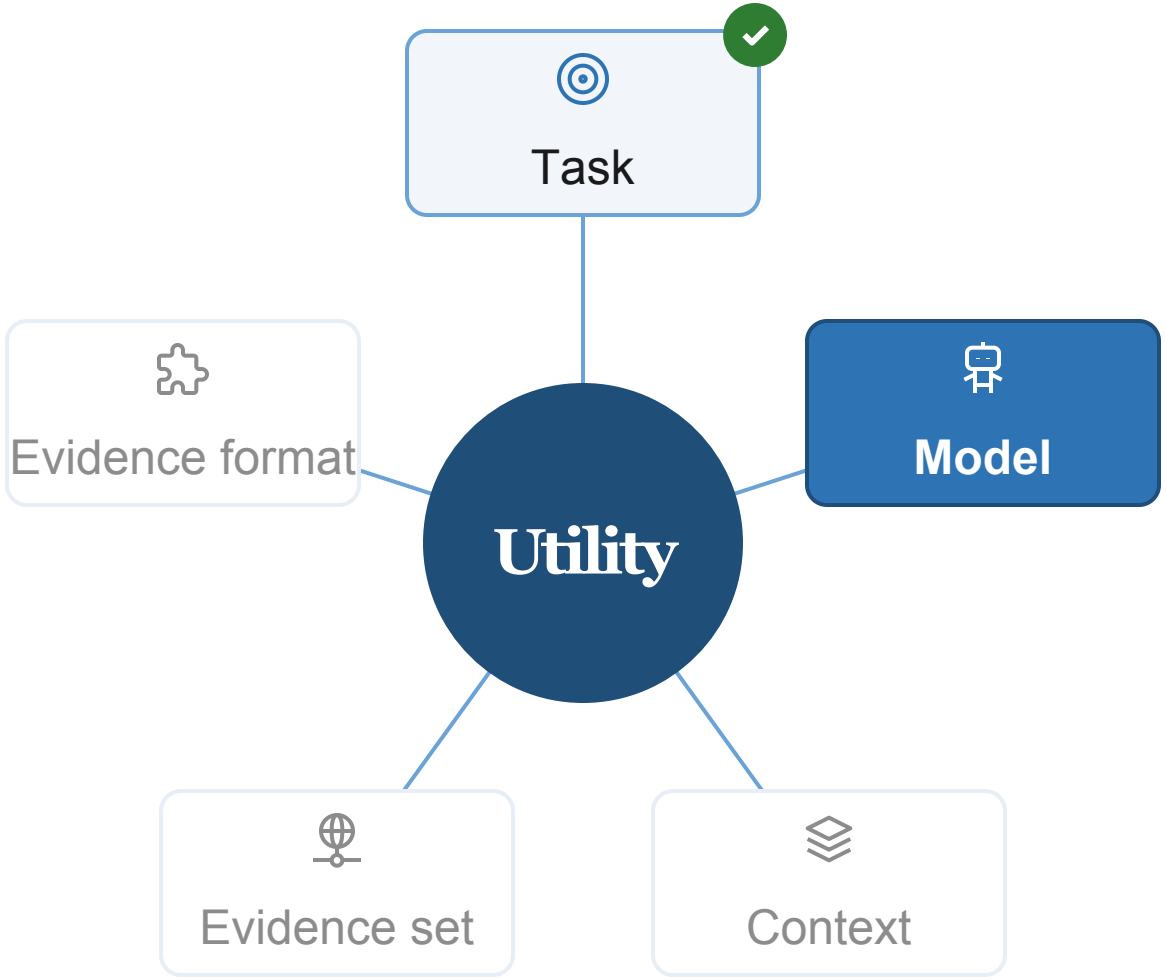
Underexplored — an open research direction. Agent-trajectory methods return in Section 4.

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

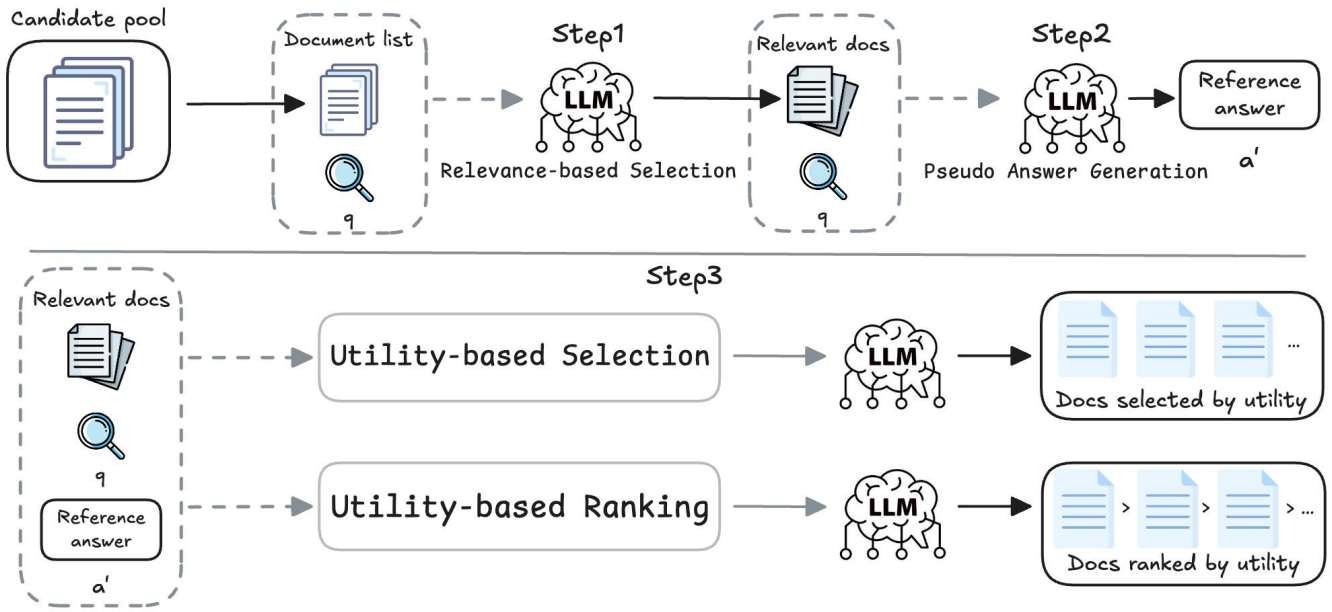
The LM's answer likelihood indicates utility

(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**



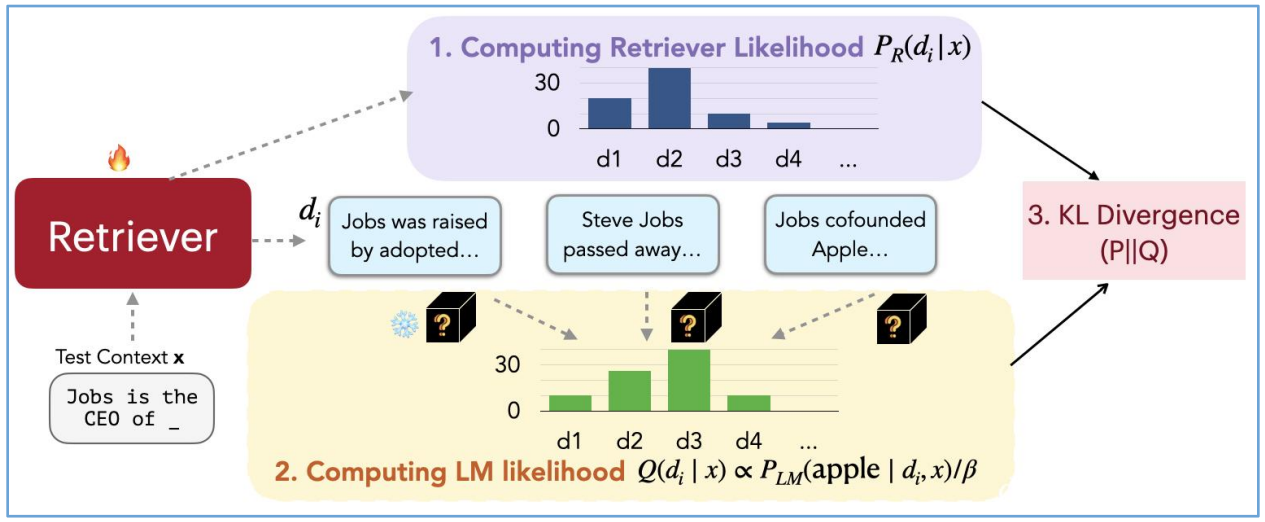
LLM judgments without manually annotated answers

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

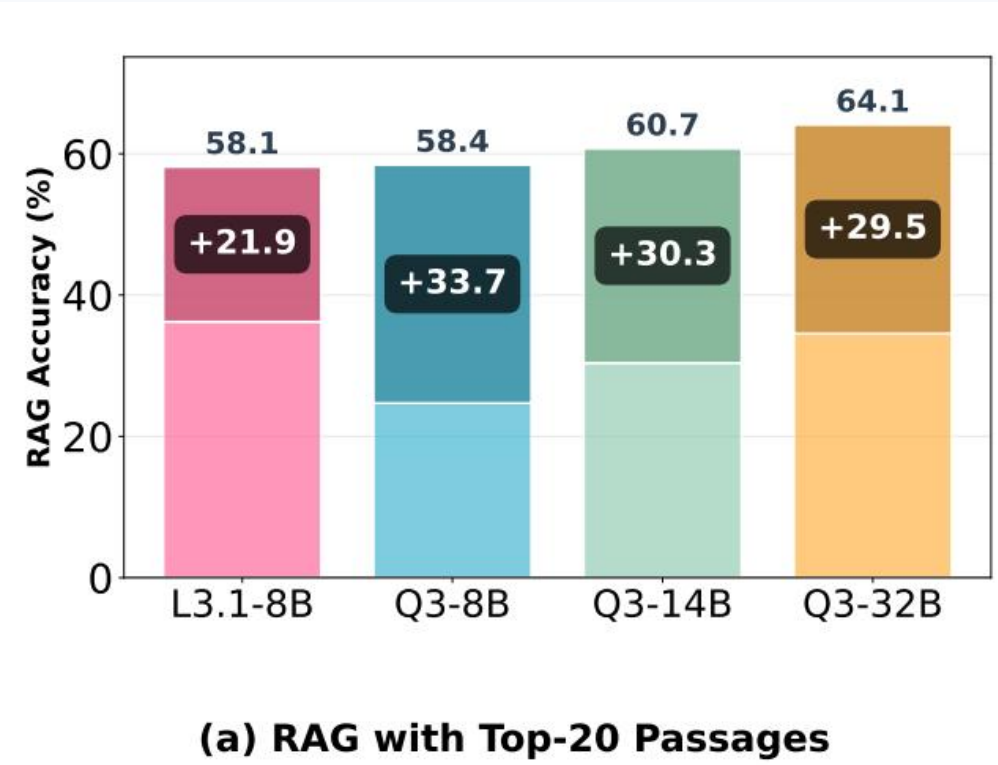
MODELS DIFFER IN

- 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

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

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

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

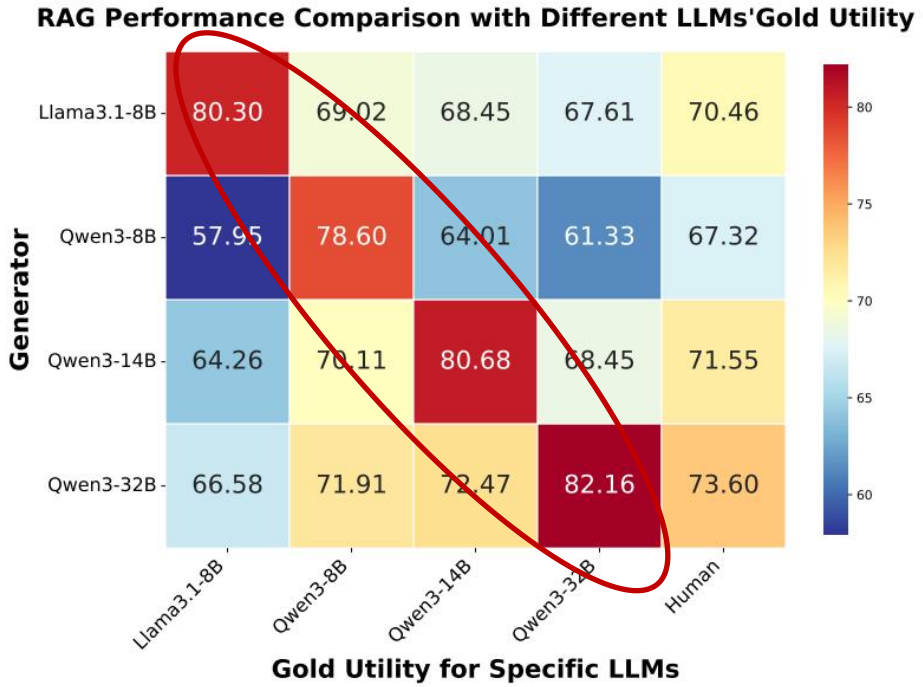
3 candidate pools / query

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

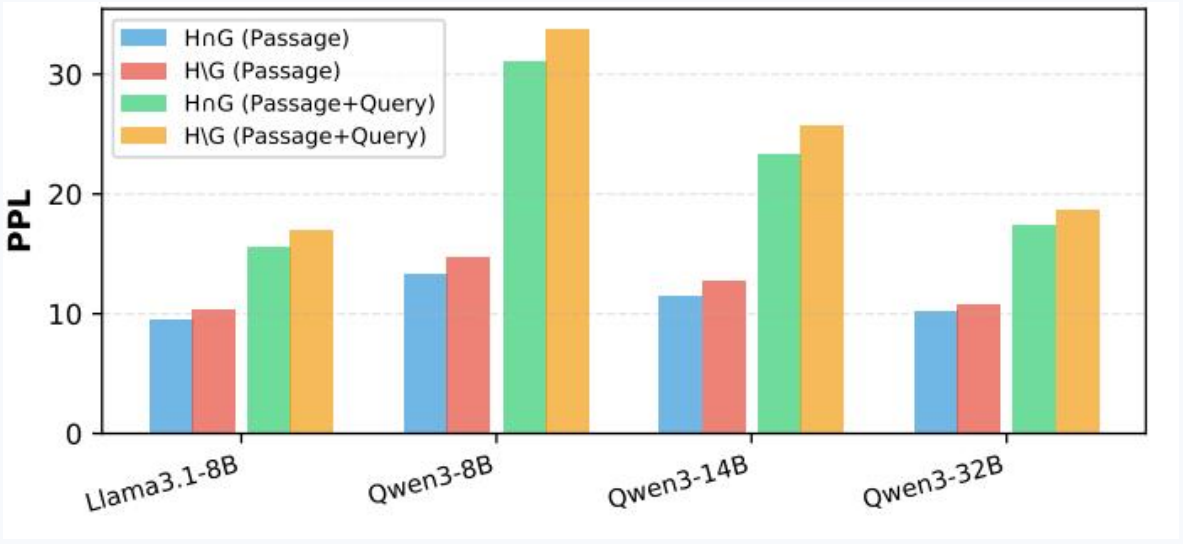
3rd **Passages picked for other LLMs**

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

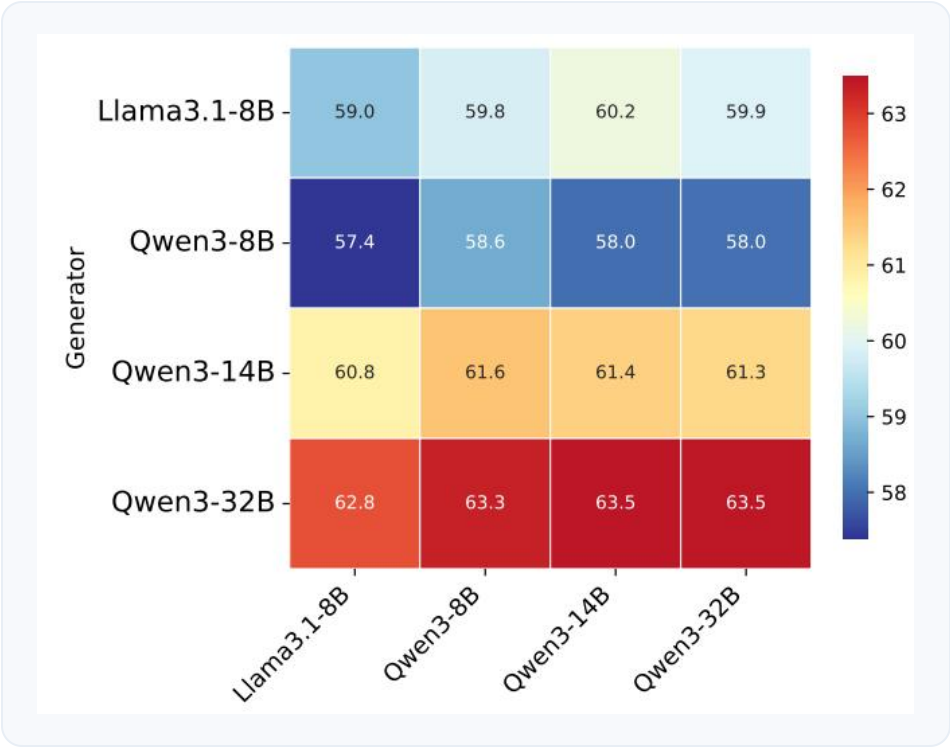
The LLM's own gold passages have **lower perplexity** than other human positives — the model reads them more easily.

> 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**.



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.

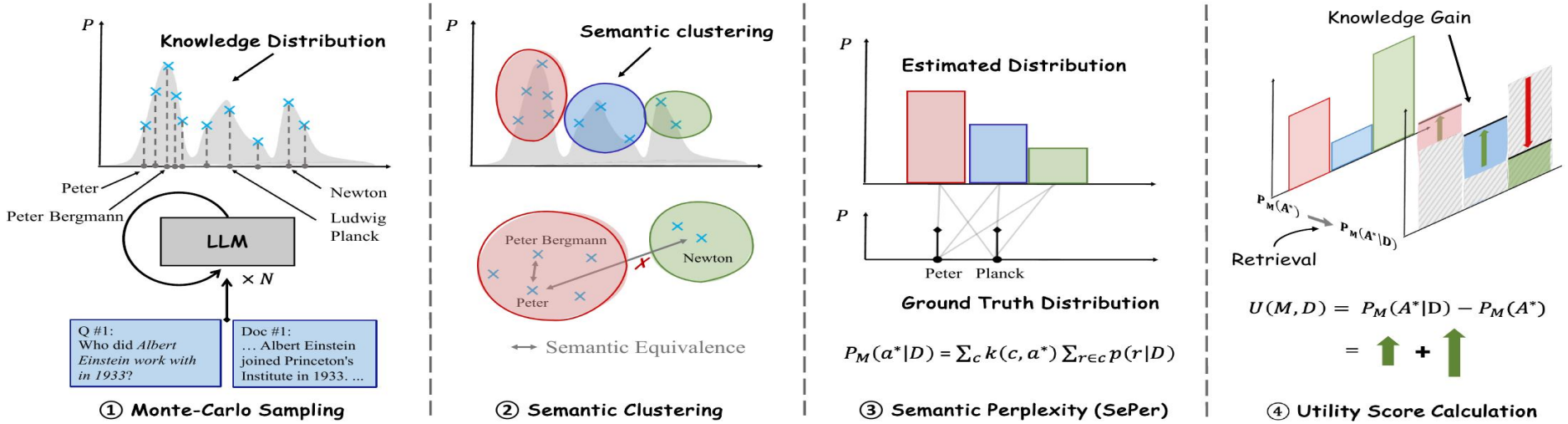
Utility judged by different LLMs looks alike — columns barely differ.

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.

belief after retrieval minus the model-specific baseline

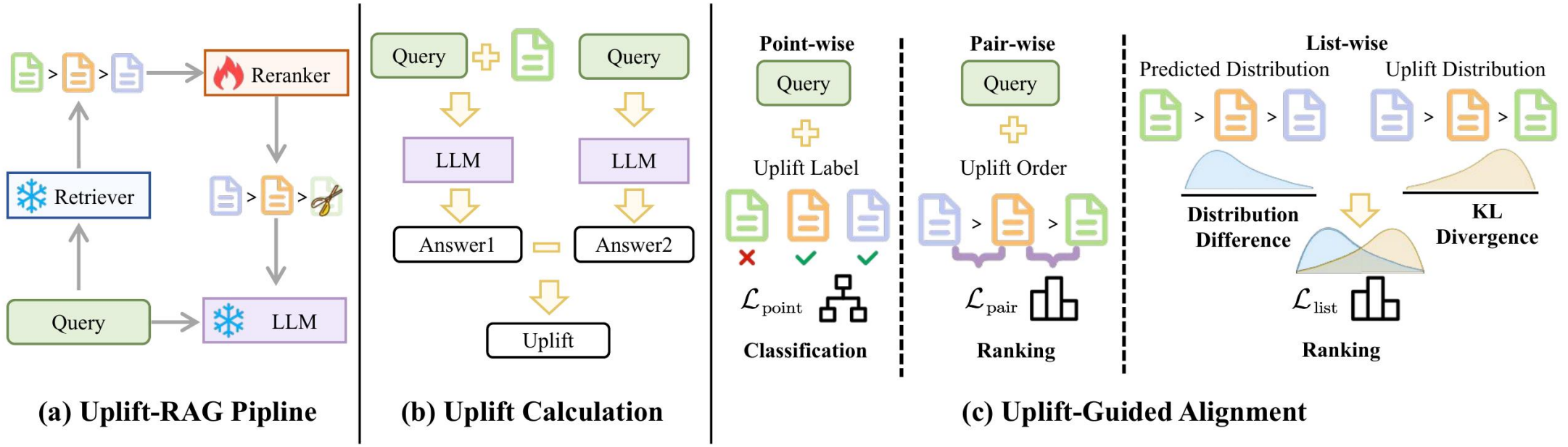


Monte-Carlo samples → semantic clusters → ground-truth belief → retrieval-induced knowledge gain.

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

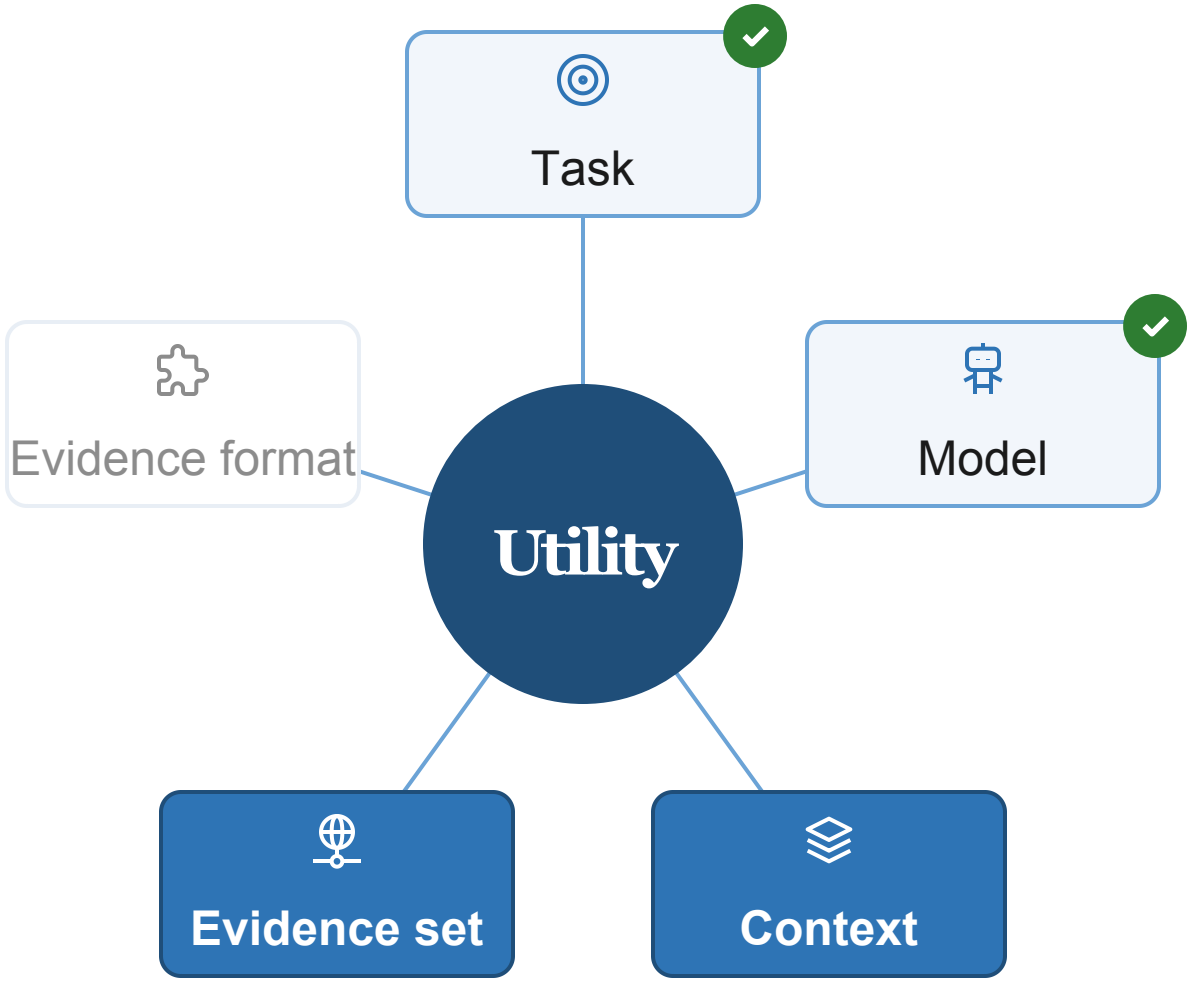
Top two rows: **scalable, reusable** supervision. Bottom row: **aligned to one model**.

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

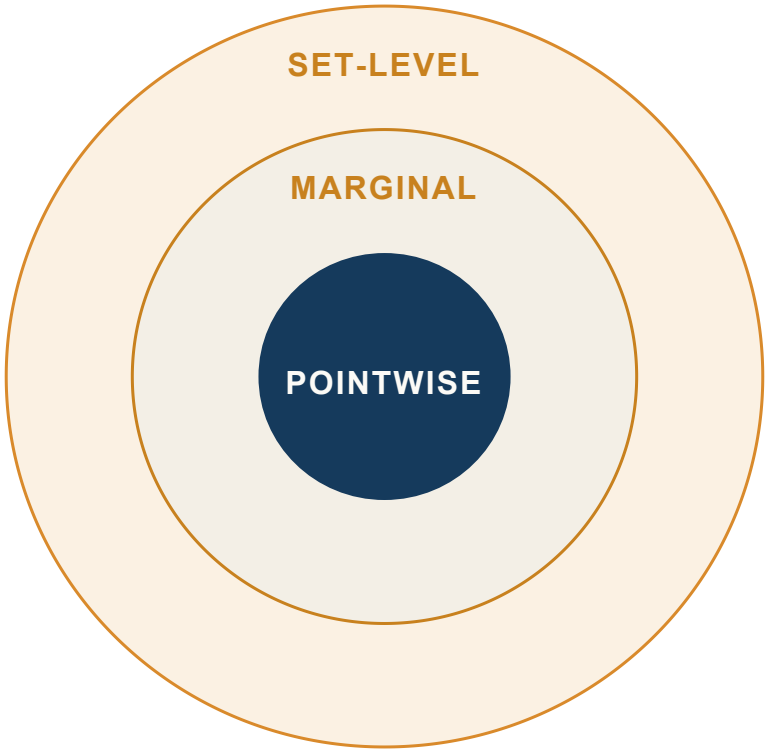


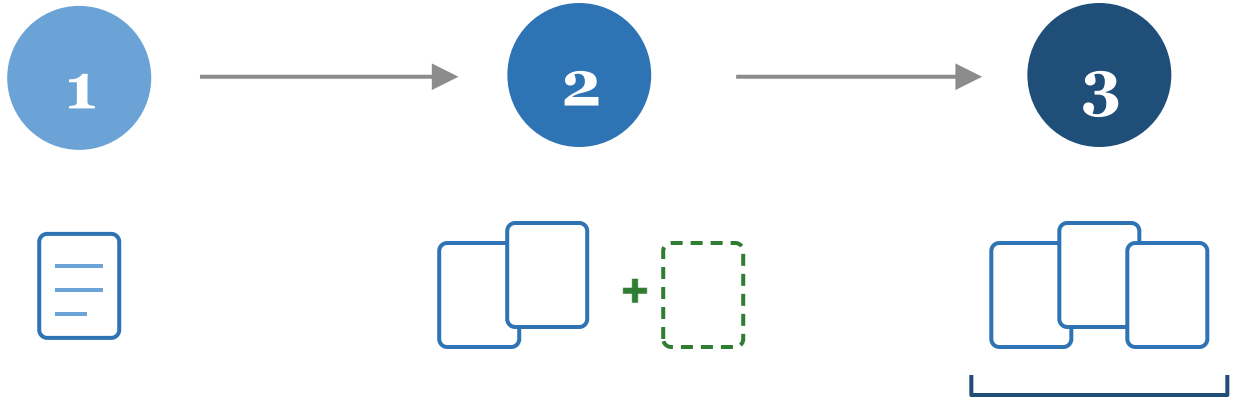
THIRD DIMENSION

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

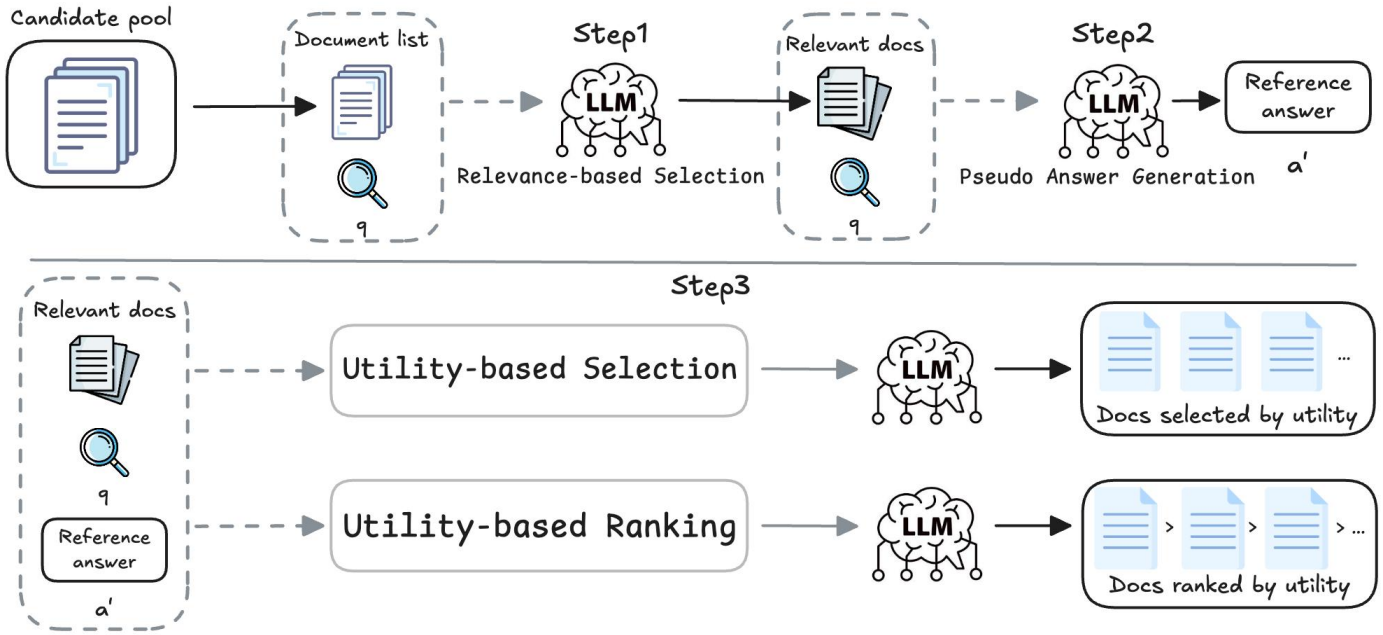
- simple to label
- easy to train and deploy
- fits one-at-a-time scoring
  selectors and rerankers

Blind to

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

Pointwise Labeling

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



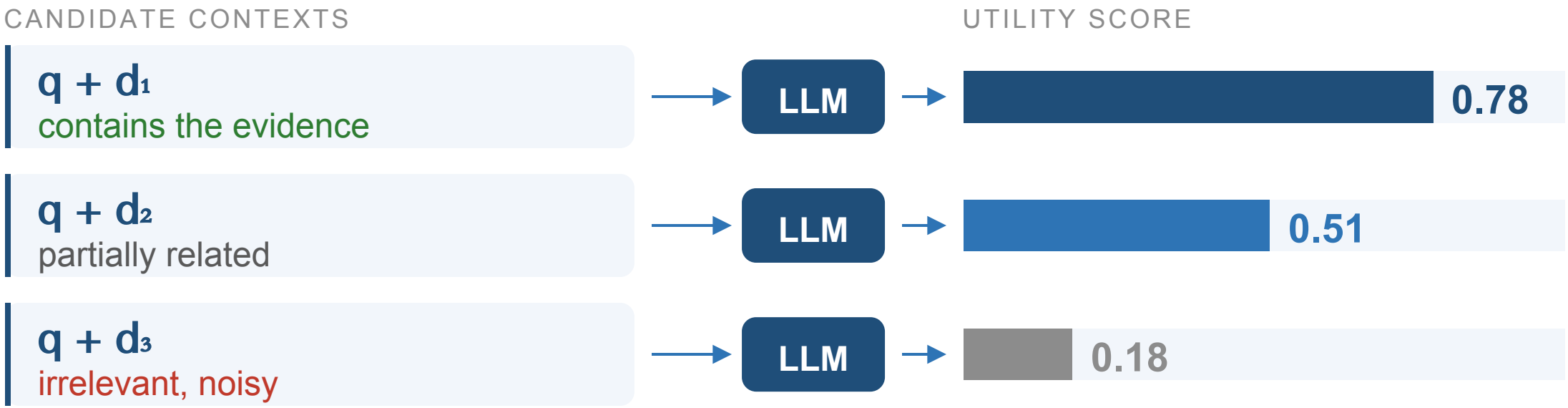
Per-document Utility Labels

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?

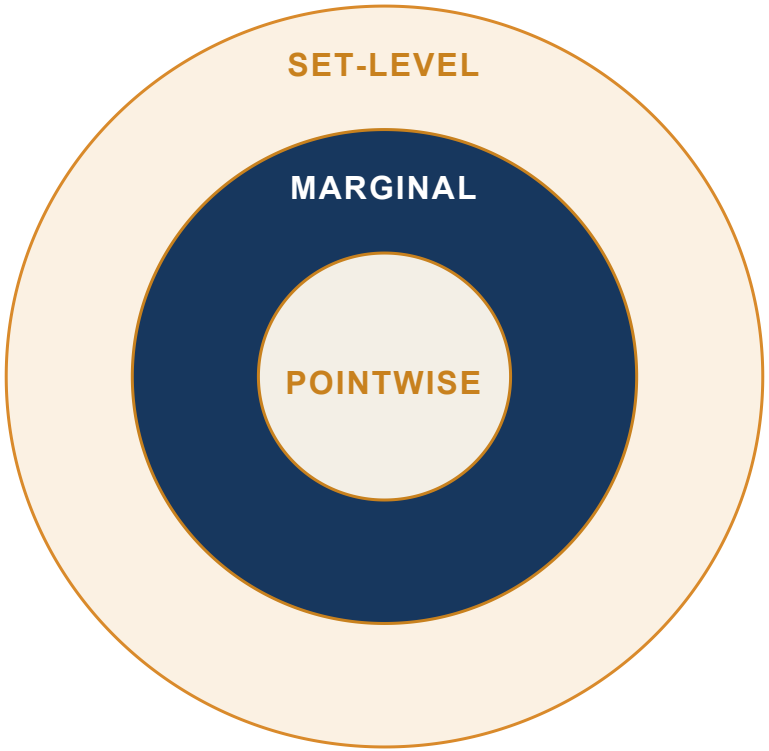


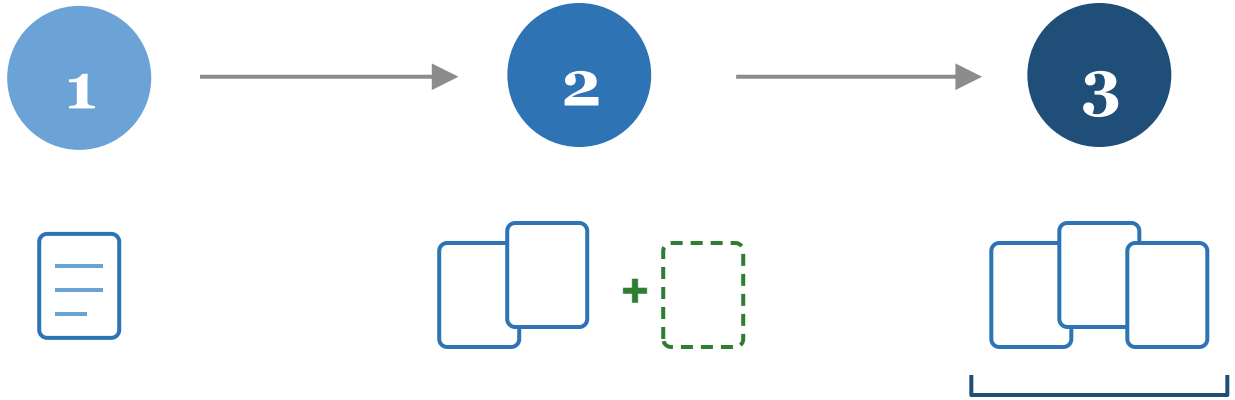
Practical signal: the model's likelihood of the gold answer

$U(d) \propto \log P(a \mid q, d)$ — 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

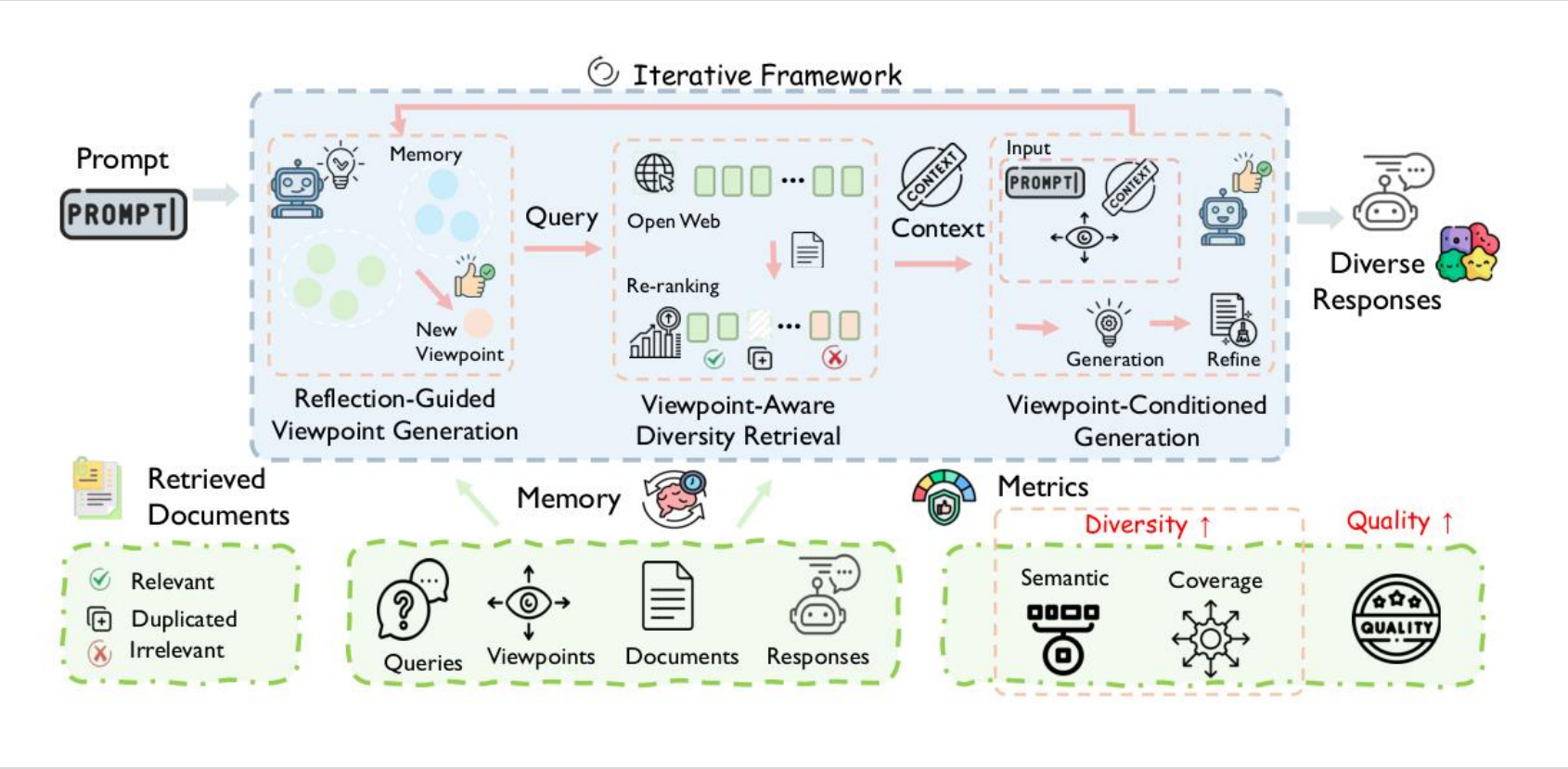
$$u(d \mid S, q, M) = \text{Perf}(M, S \cup \{d\}) - \text{Perf}(M, S)$$

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

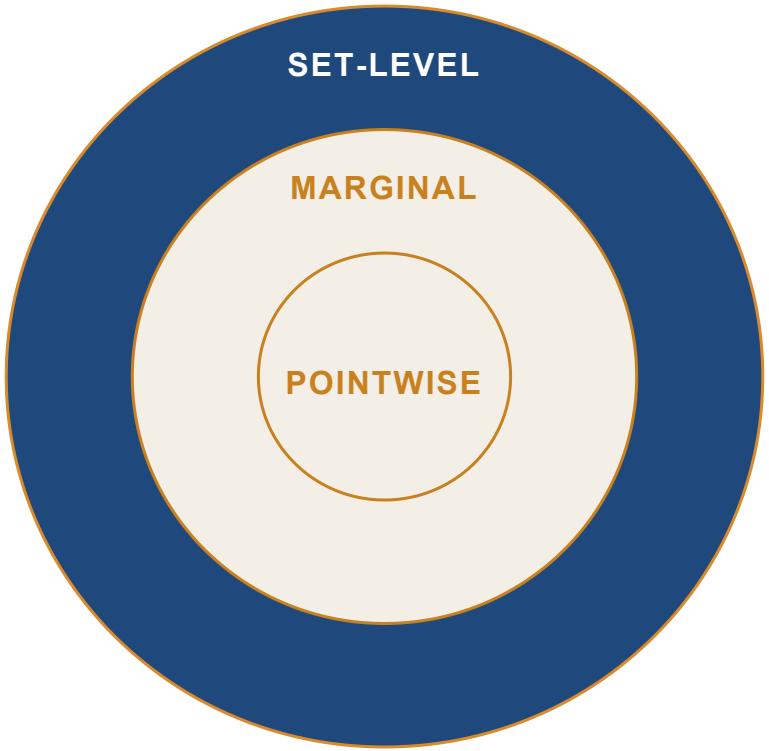
GPT-4o assigns utility 1–5 from role and criticality

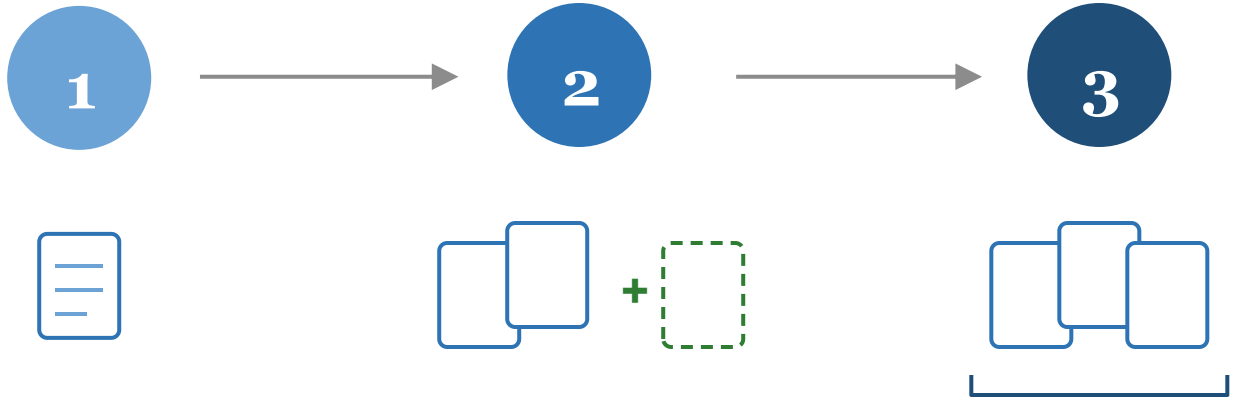
3 · Distill

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?

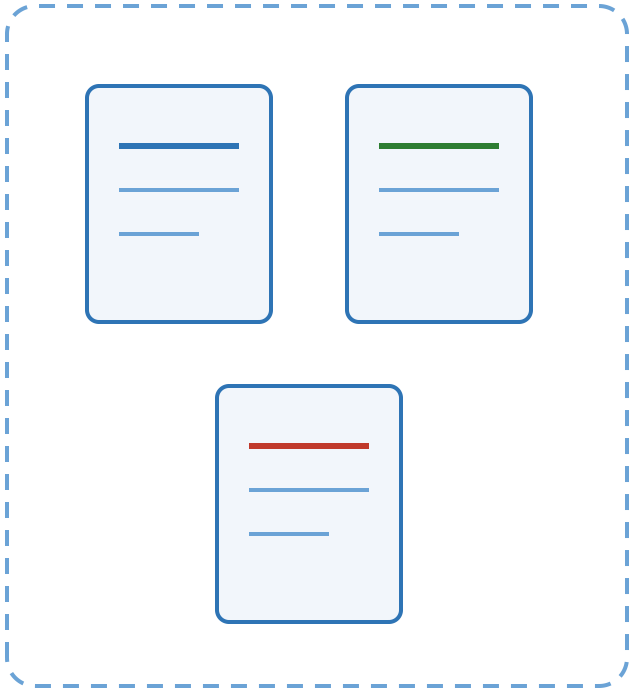
Marginal

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

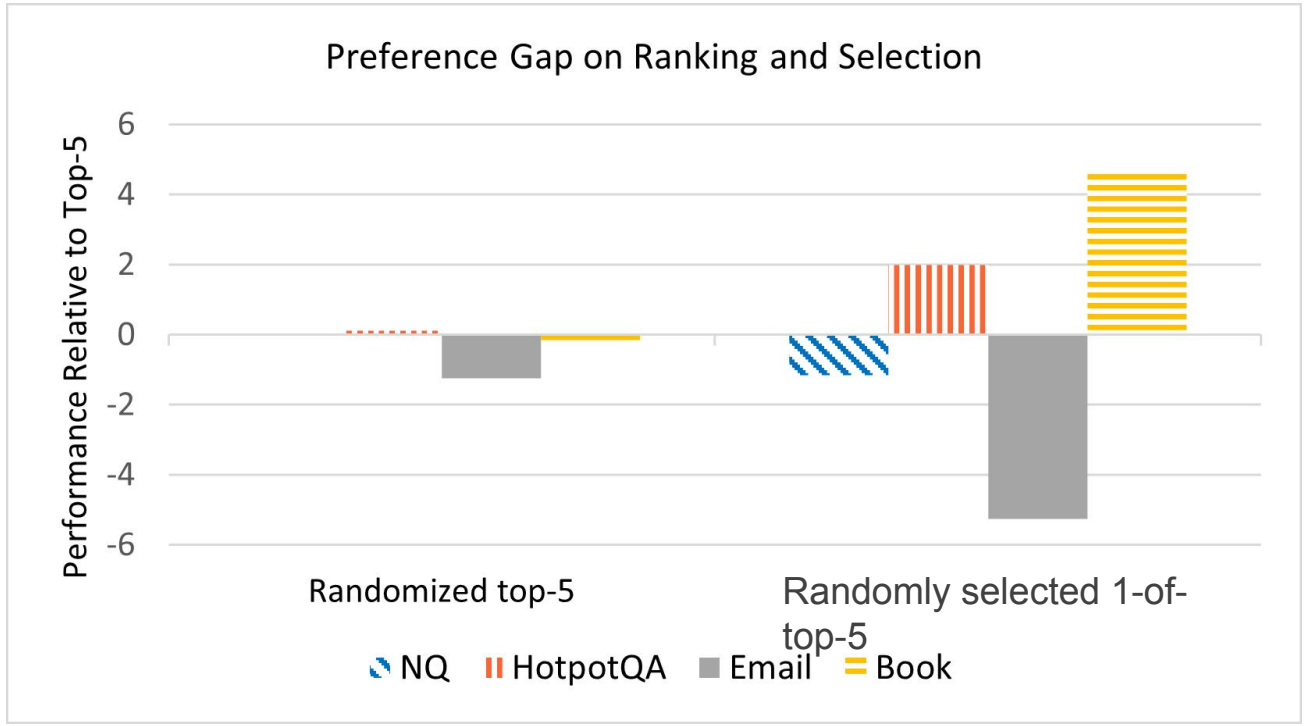
Conflict management

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.

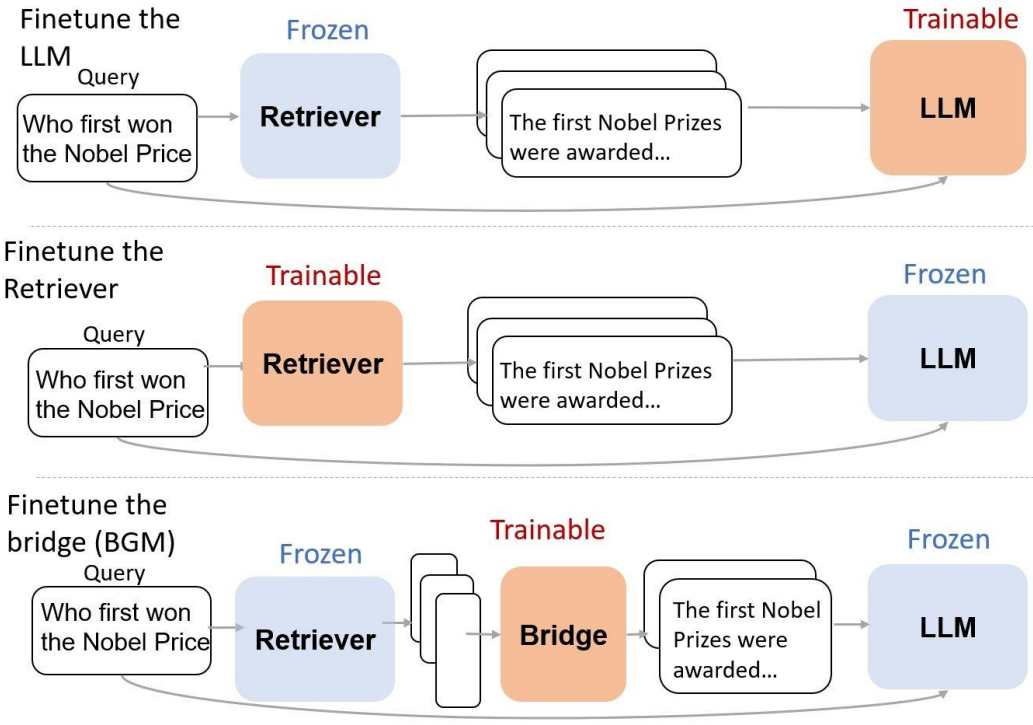
Selection does

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.

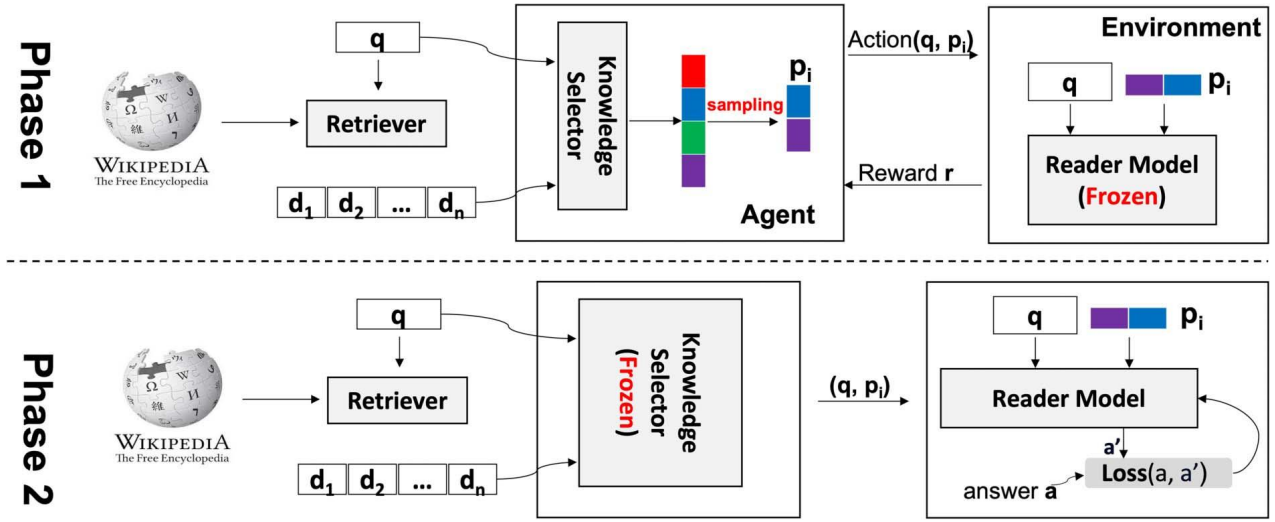
3 Reinforcement learning

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.

Mutual Learning: The Knowledge Selector 2 — LLM-Centric Utility



a knowledge selector sits between retriever and reader; the two phases alternate each epoch.

Phase 1 · train the selector

Reader frozen; selector updated by RL using the reader's reward — no labeled pairs needed.

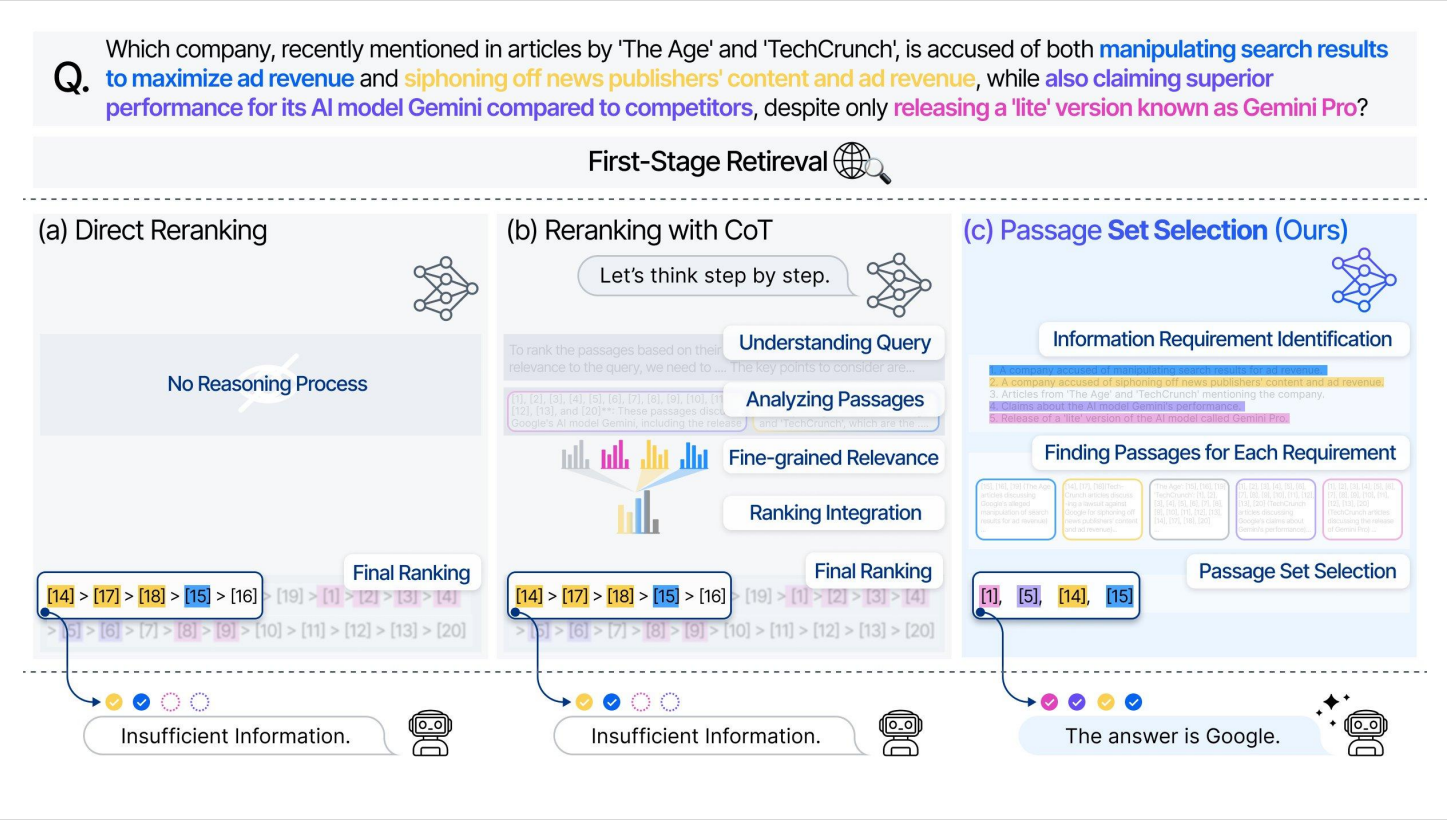
Phase 2 · train the reader

Selector frozen; reader updated on the compact, high-utility subset the selector chose.

retrieve what you *need* — selector and reader teach each other to keep only the useful passages.

Wang et al. Retrieve What You Need: A Mutual Learning Framework for Open-domain Question Answering. TACL, 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

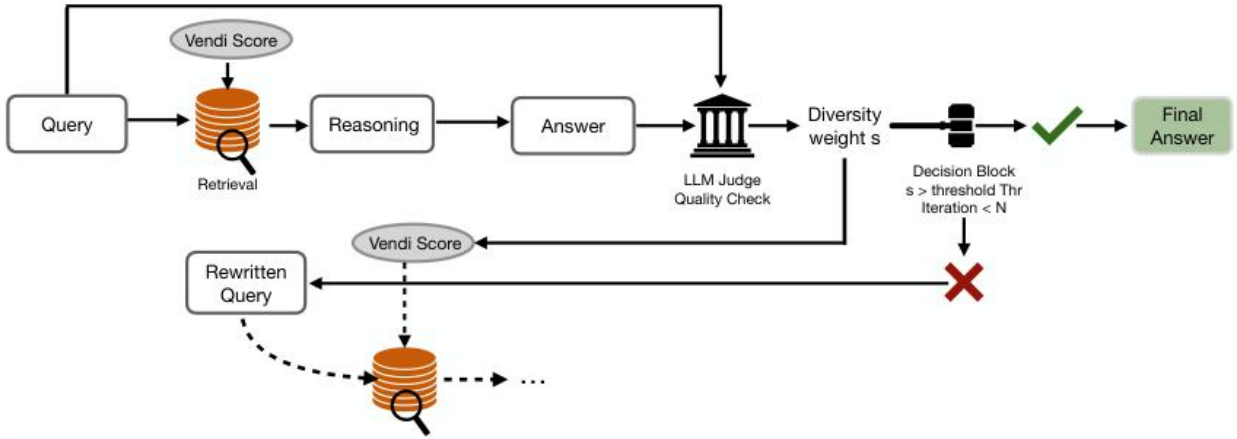
SET-LEVEL DIVERSITY

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

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



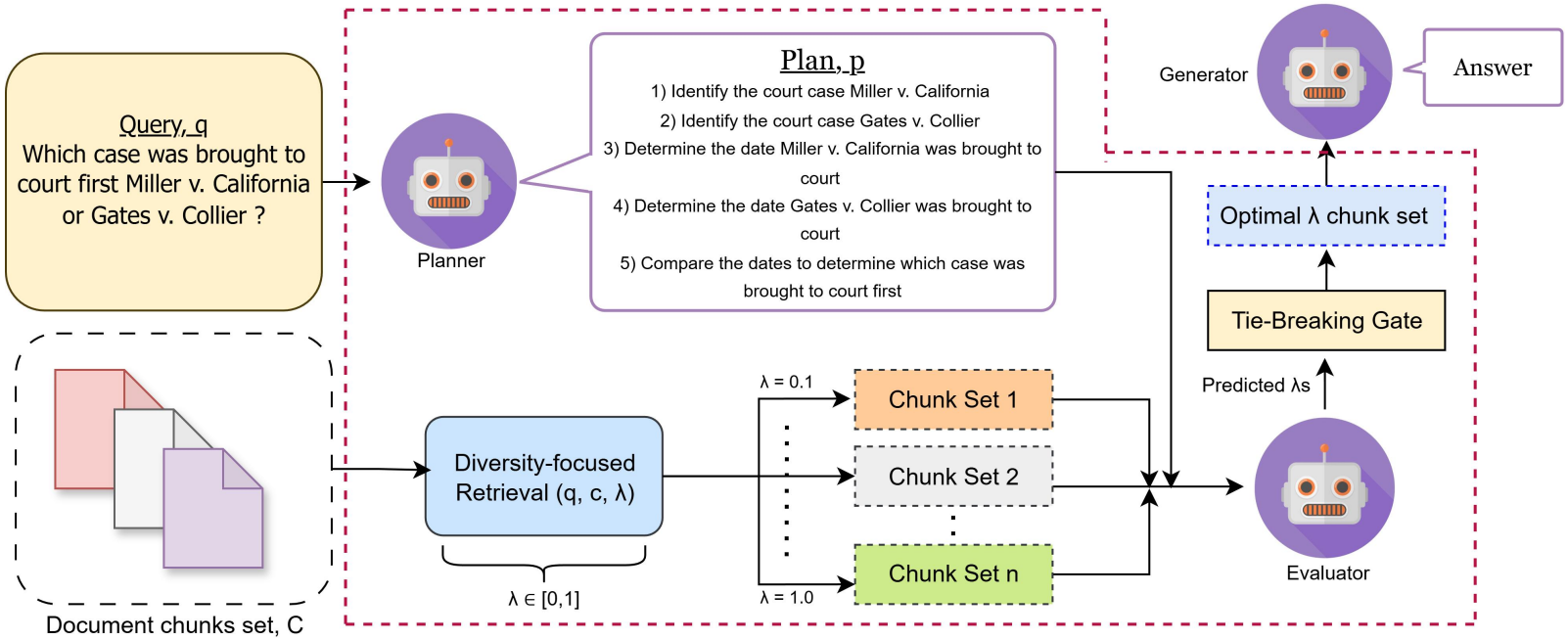
Utility: s · Vendi Score + (1−s) · relevance

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



c_S is the centroid of the already selected chunks.

QUERY-AWARE CONTROL

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

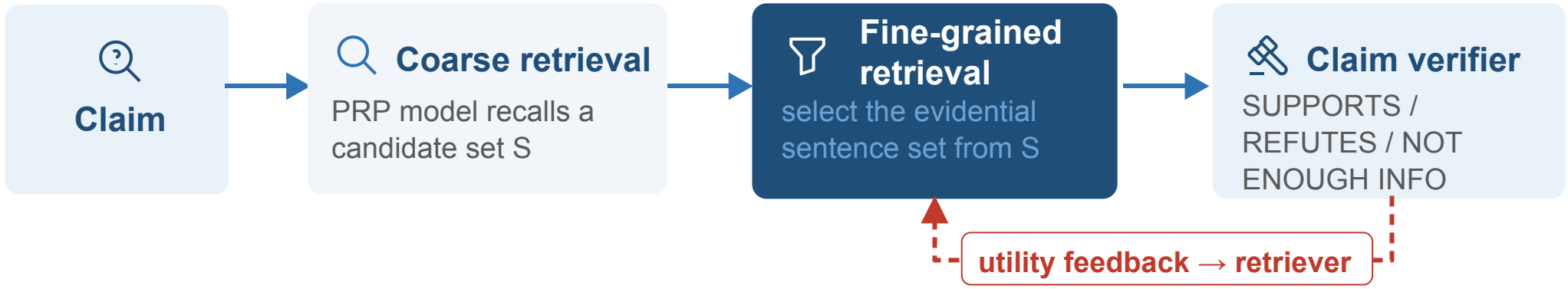
Evaluator selects among multiple λ-conditioned sets.

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.



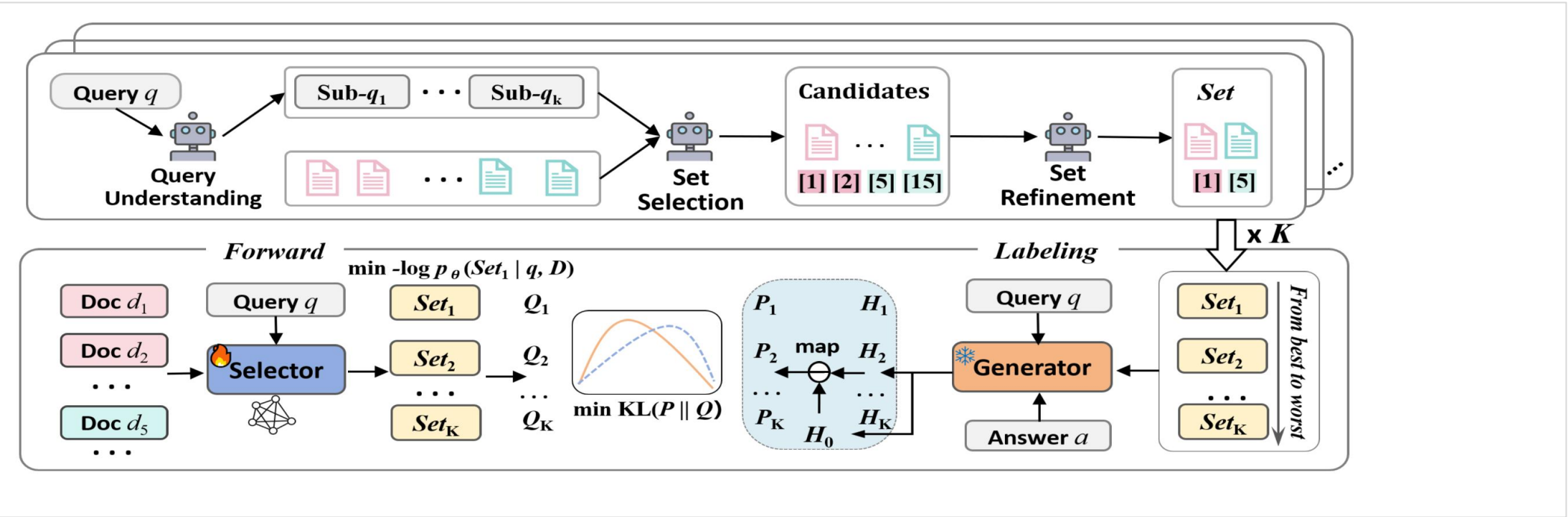
A CLS token before each sentence; use gumbel-softmax to learn which sentence should be kept

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.

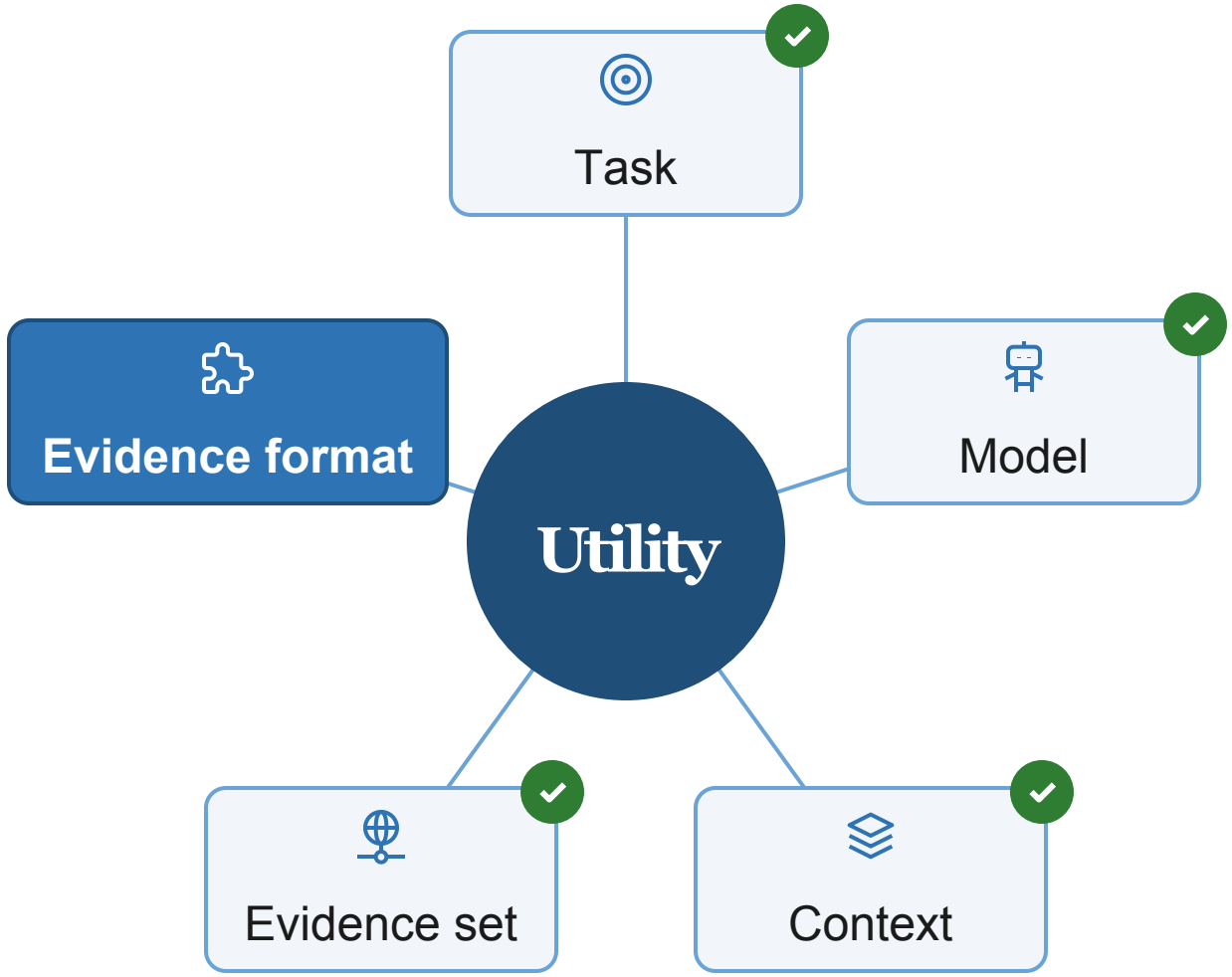
Jiang et al. OptiSet: Unified Optimizing Set Selection and Ranking for Retrieval-Augmented Generation. arXiv, 2026.

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



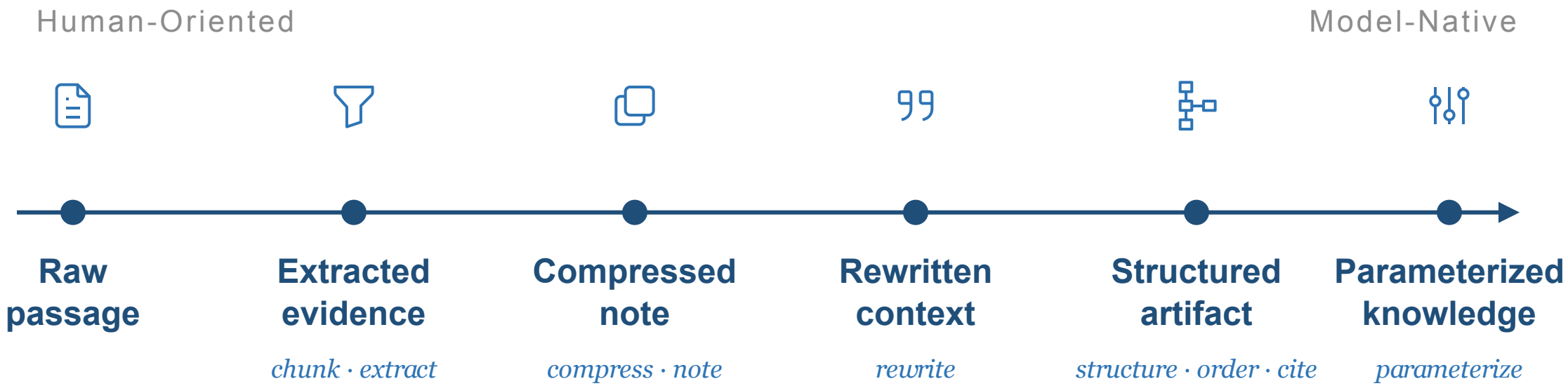
LAST DIMENSION

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

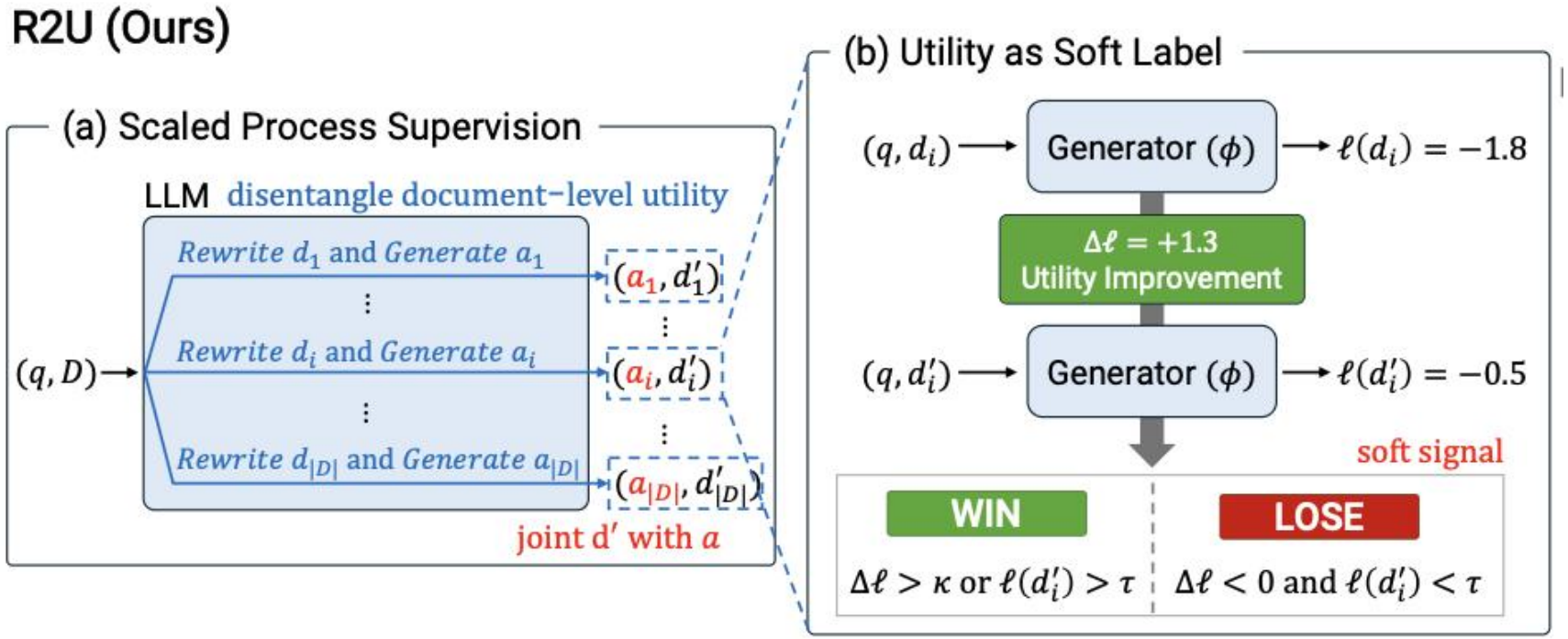
Relevance-trained rewriters keep fluent but useless spans.

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

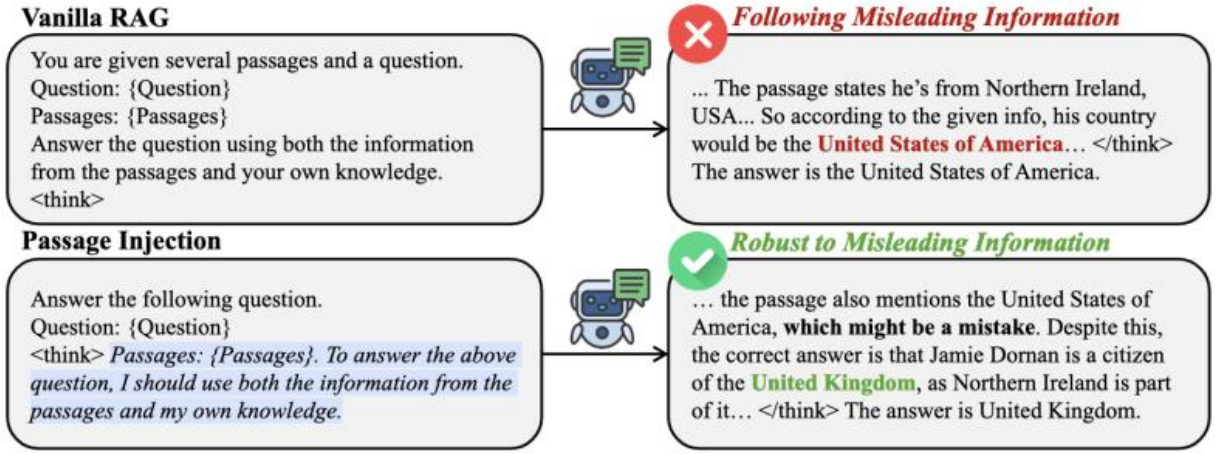
Put retrieved passages **inside the reasoning**, not the input prompt.

➢ **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

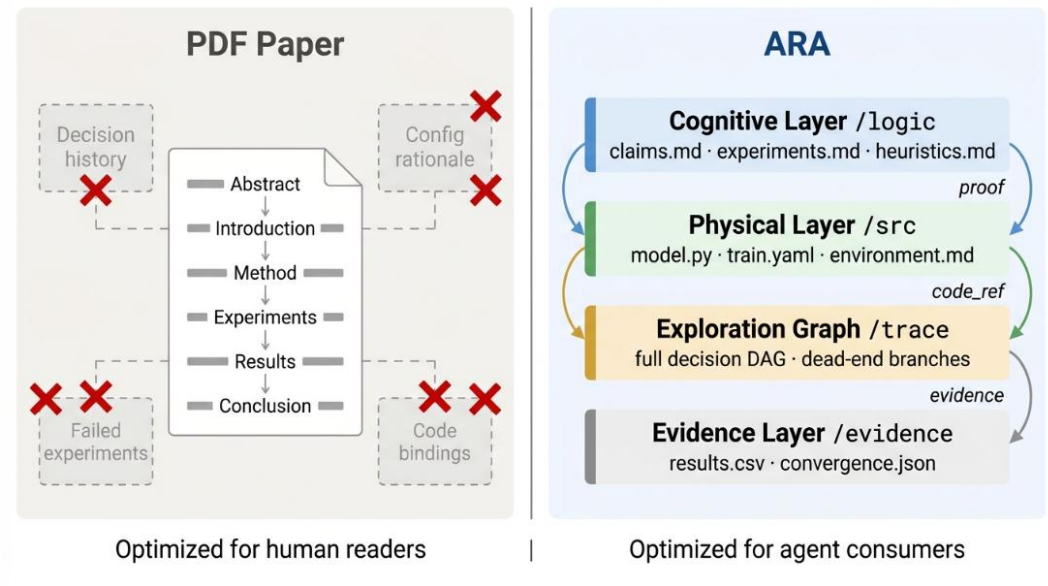
- scientific logic
- executable code
- exploration graphs
- evidence grounding

Future retrieval targets: structured, executable, directly inspectable.

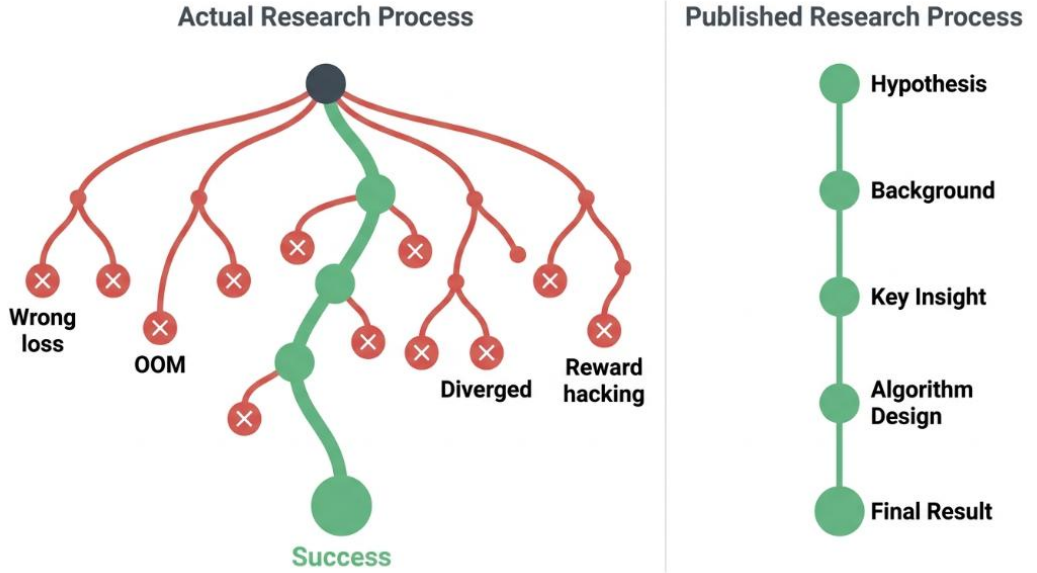
Liu et al. The Last Human-Written Paper: Agent-Native Research Artifacts. arXiv, 2026.

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.

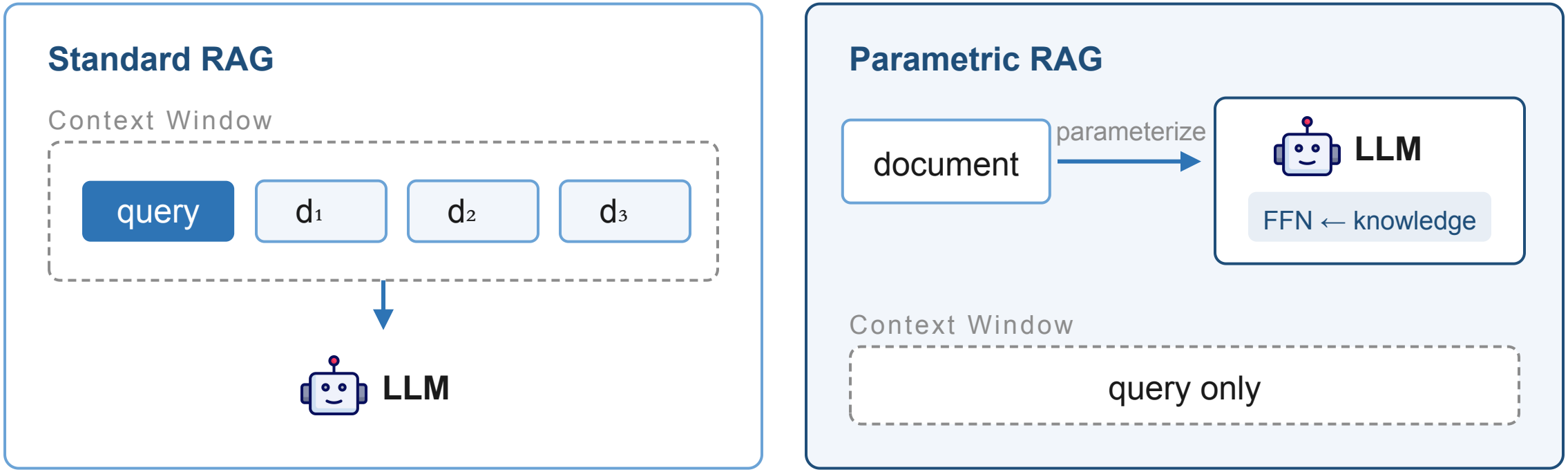
Agents load only the layer they need —

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**.

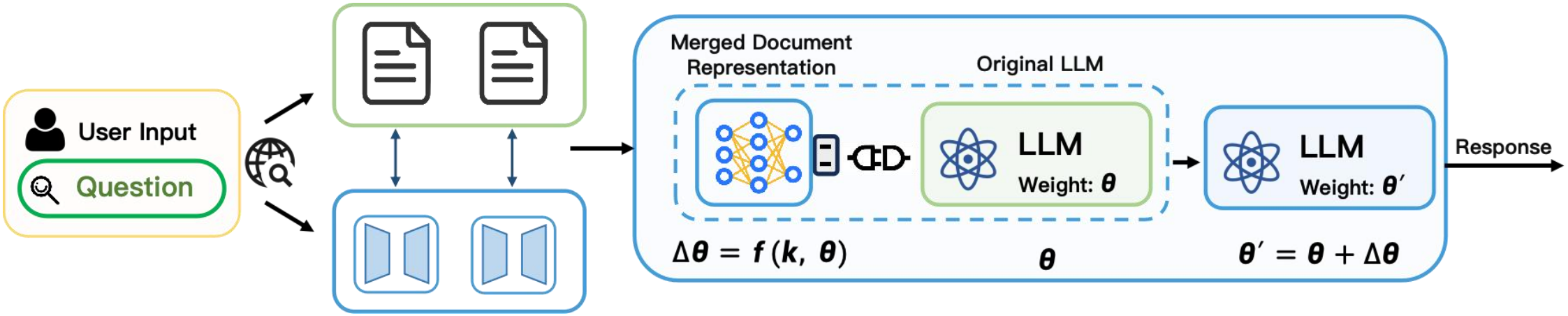


in-context text → parameterized knowledge

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

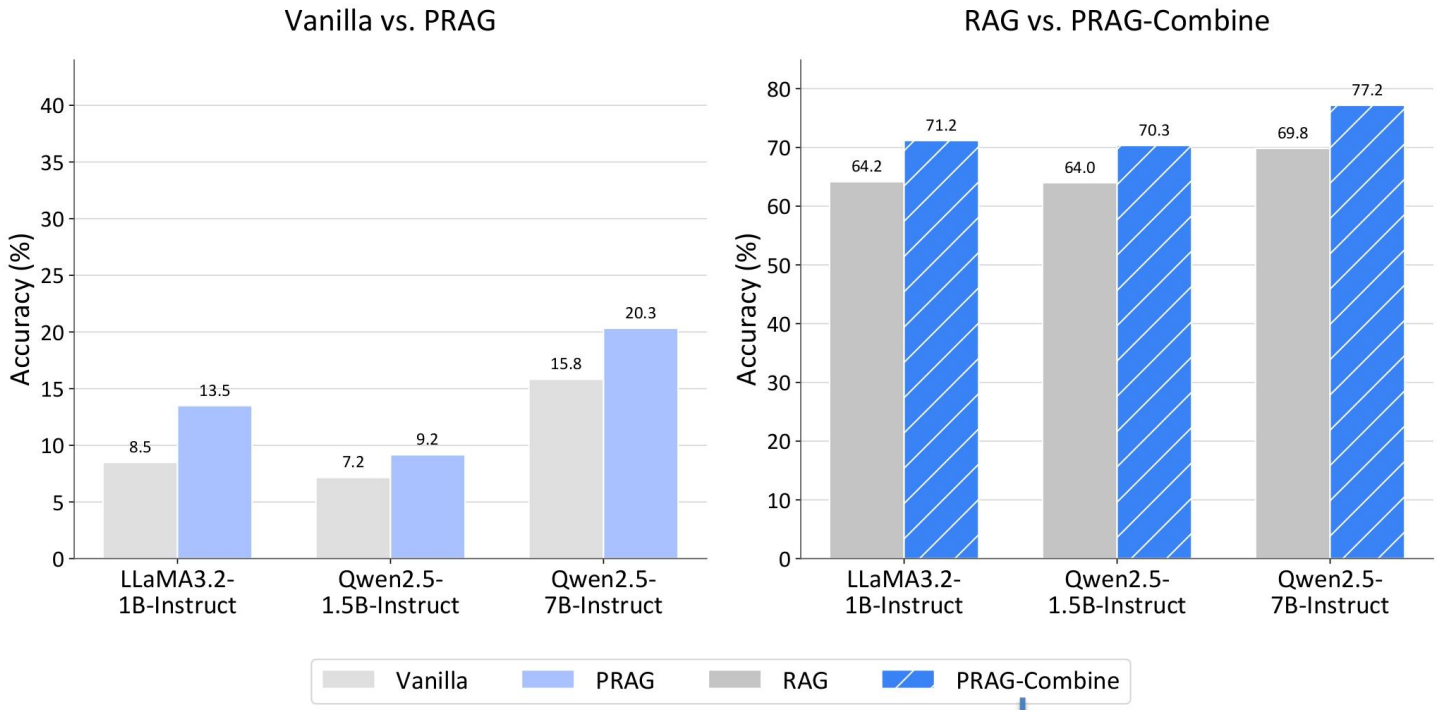
Online

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.

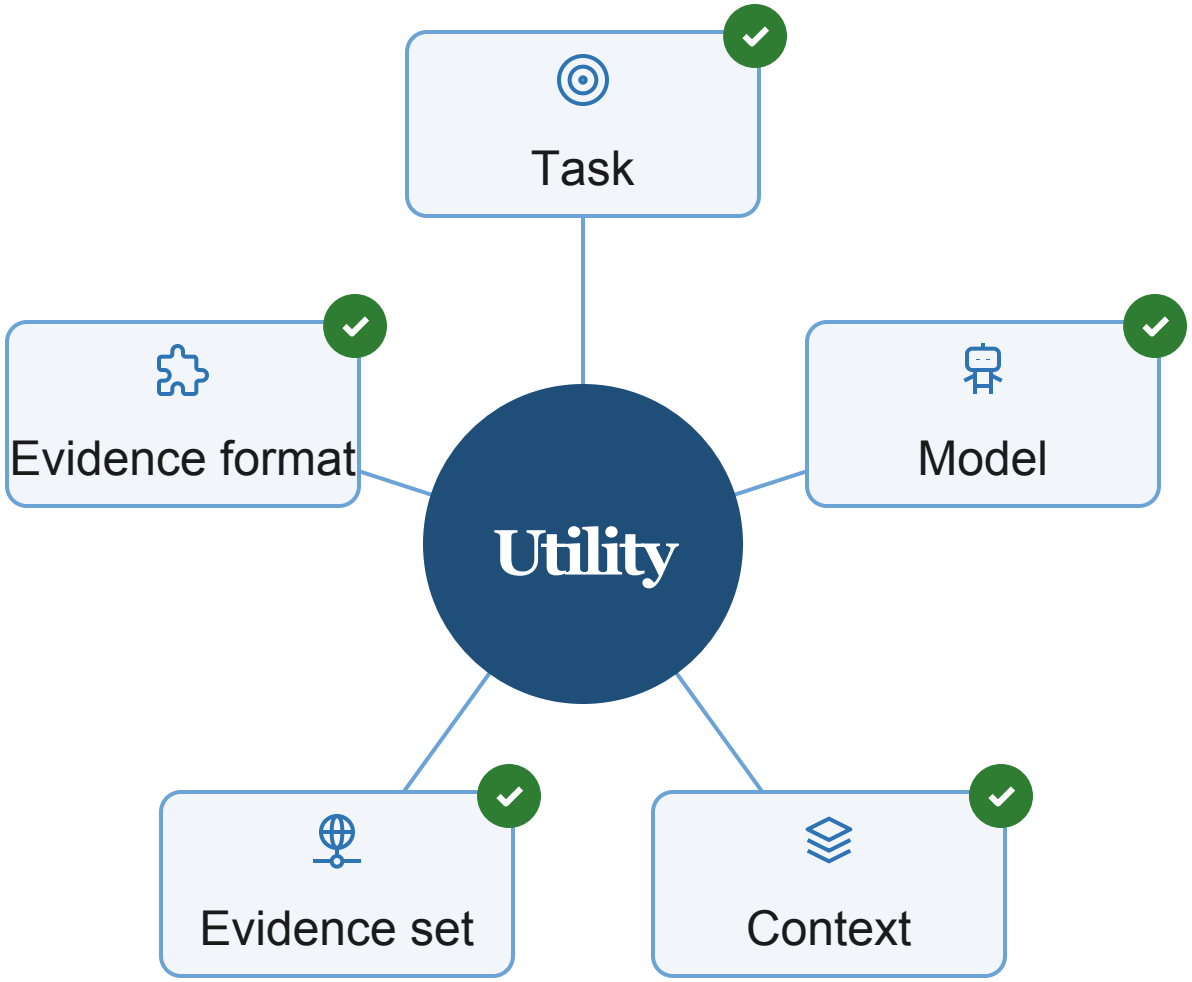
Evidence Format as Utility Transformation

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

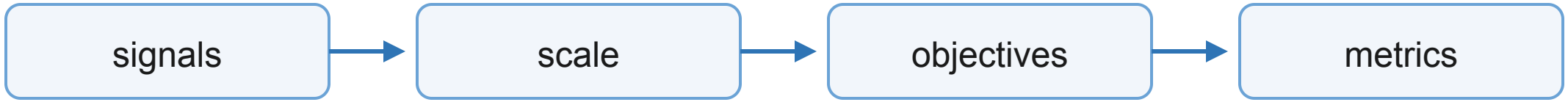
presentation changes utility

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