SECTION 3

Utility Modeling & Optimization

Measurement, training, and evaluation — utility made operational.

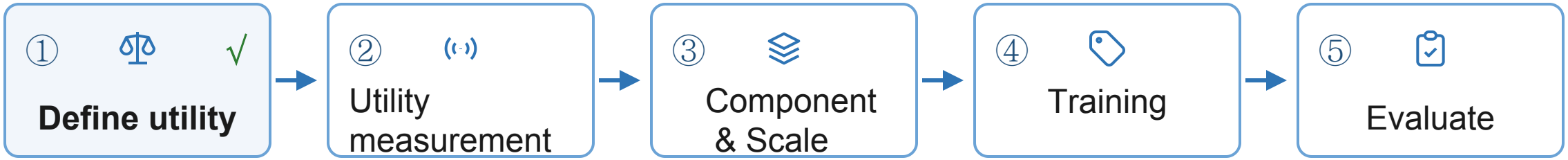
The Core Question

PICKING UP FROM SECTION 2

We defined utility. Can we optimize it?

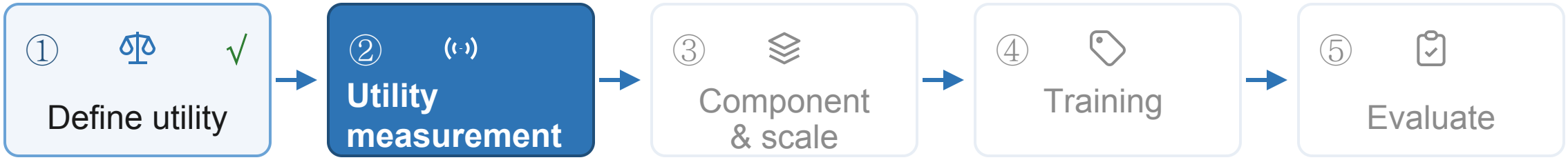
> How can we **modeling** utility — and **optimize** utility-focused retrievers/selectors/rerankers, and **evalute** utility-focused RAG?

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

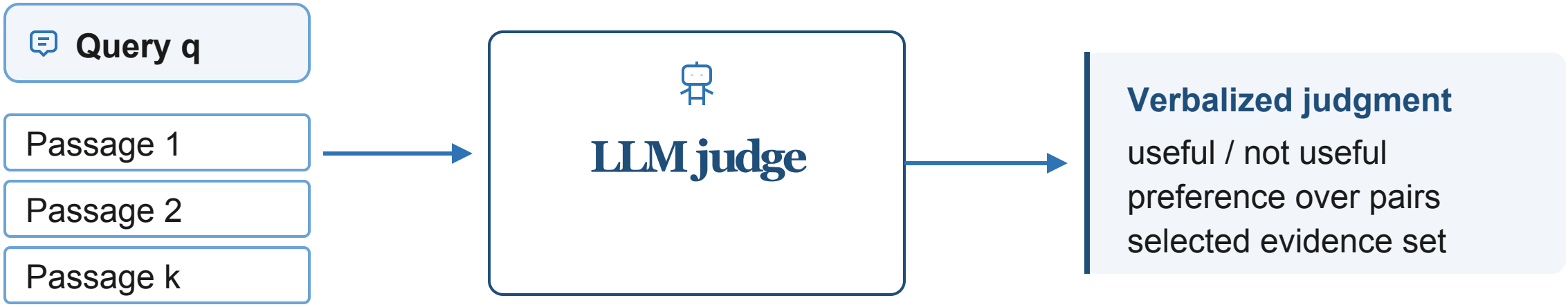
Use attention weights as soft utility to measure how the reader model **actually used** the passages.

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

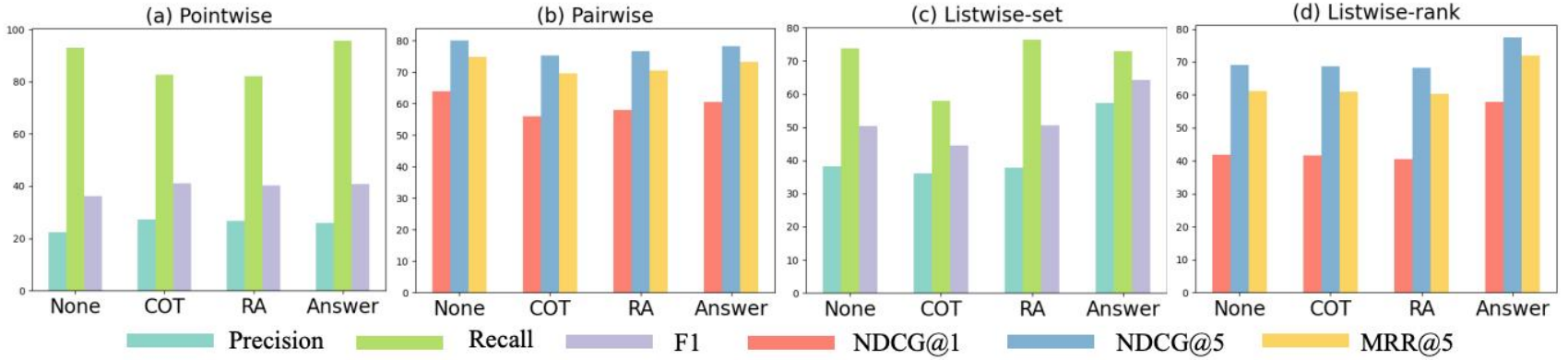


Figure 6: The performance (%) of utility judgments using ChatGPT under different forms, i.e., pointwise, pairwise, listwise-rank and listwise-set, on the NQ dataset using the prompts with different requirements.

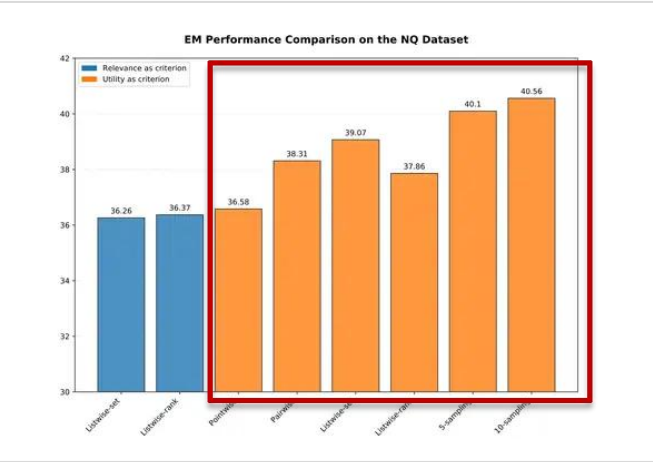
Reasoning and **pseudo-answer generation** markedly improve the LLM's utility judgments performance.

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



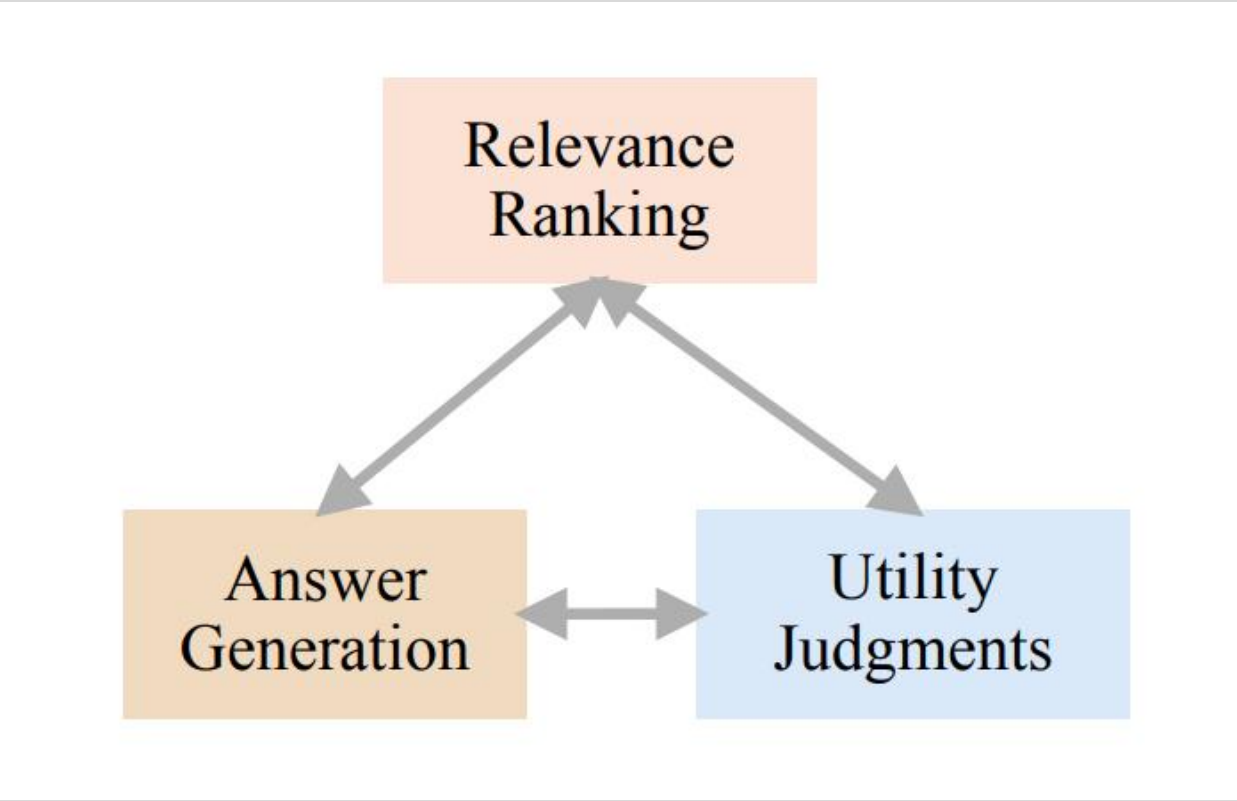
- ✓ **Utility-based selection** is better aligned with generation than relevance-based selection.
- ✓ **Listwise** selection achieves the best RAG performance compared to other judgment methods.

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**.

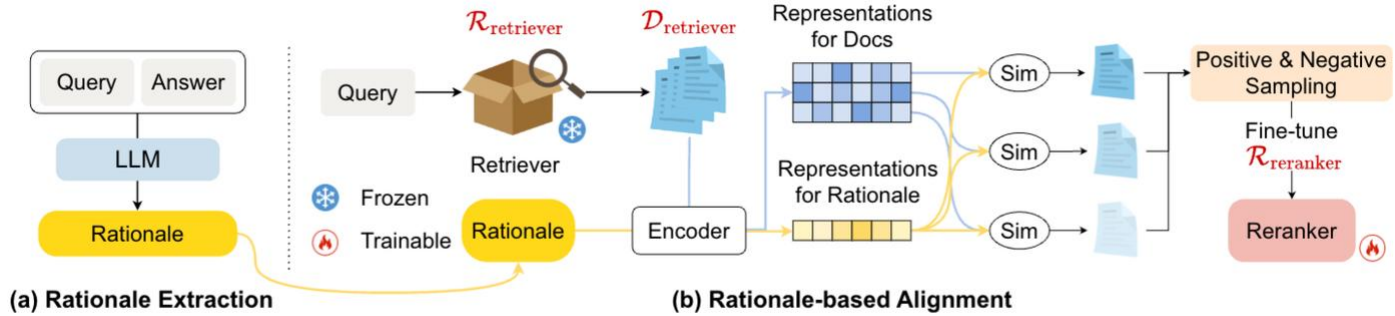


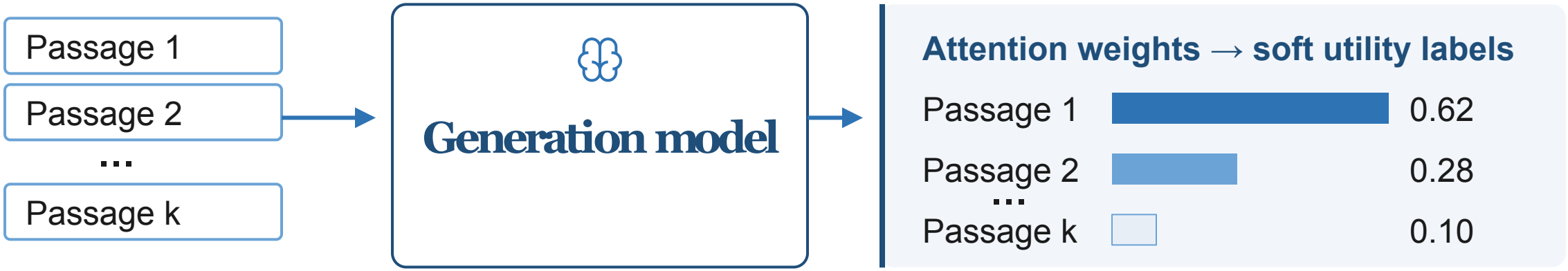
Figure 1: Overview of RADIO.

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

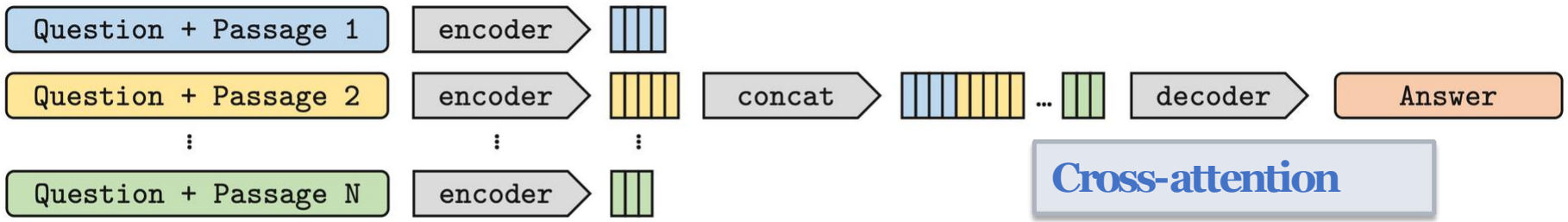


Figure 2: Architecture of the Fusion-in-Decoder method.

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

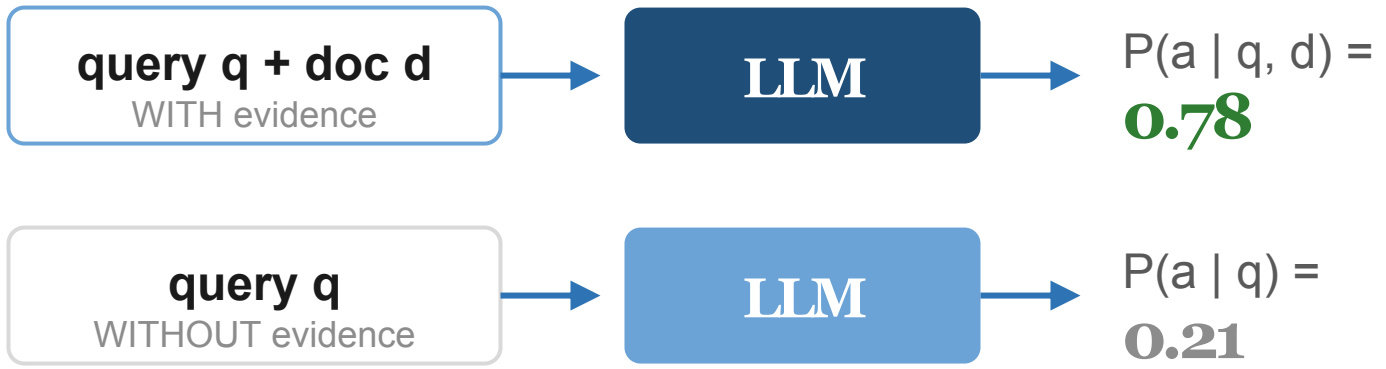
SePer

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

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

$$= 0.57$$

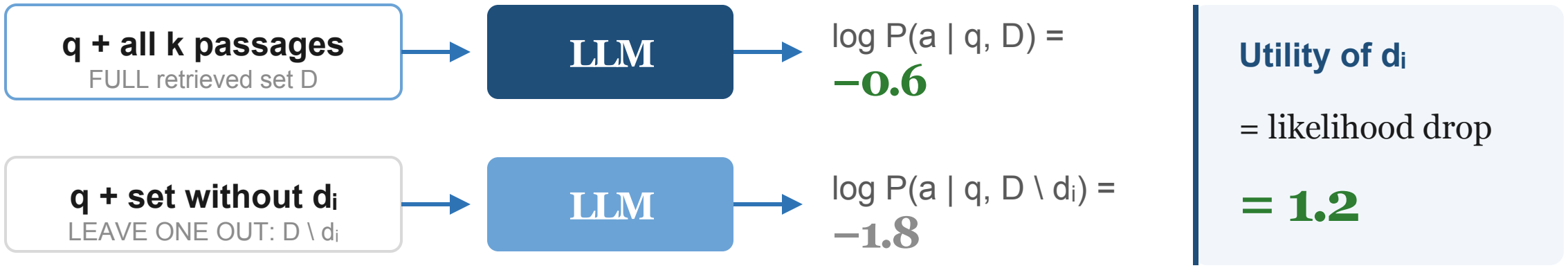
Practical scoring: the normalized log-probability of the ground-truth answer under the prompt containing q and d — a dense, differentiable, model-specific utility signal.

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.



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

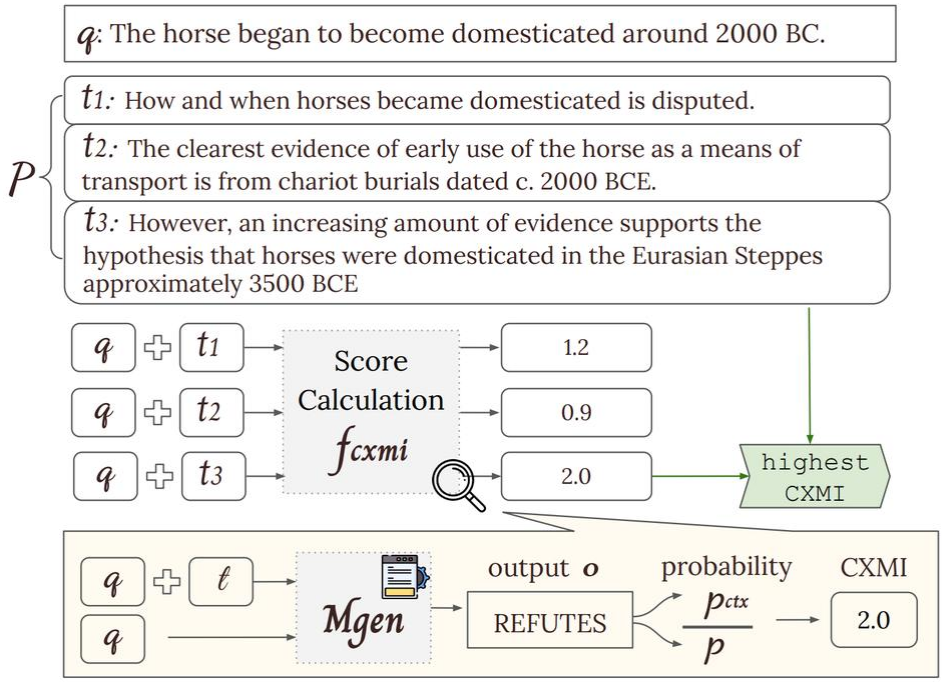


Figure 3: An example illustration of context filtering with the CXMI strategy.

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

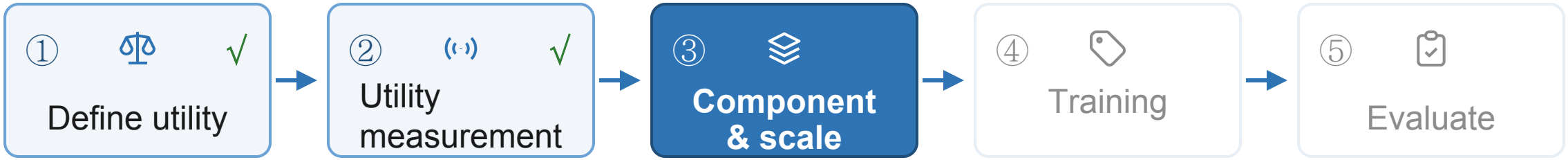
Filtering noisy context improves the generator's accuracy and robustness.

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

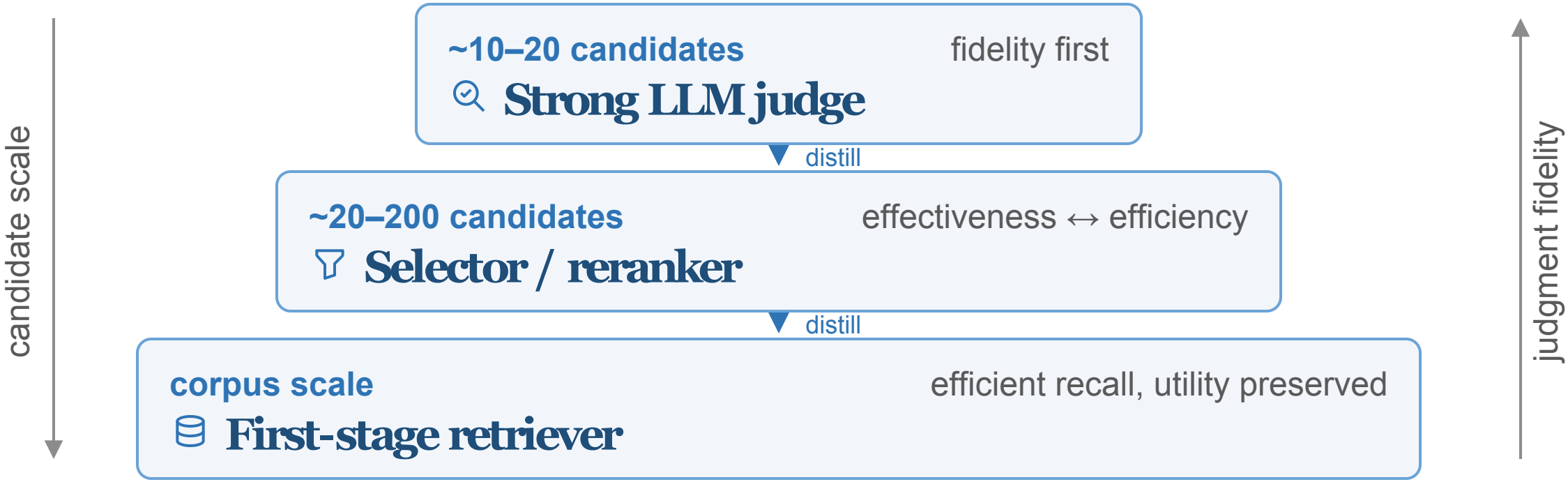
3.4 · The Efficiency Ladder



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

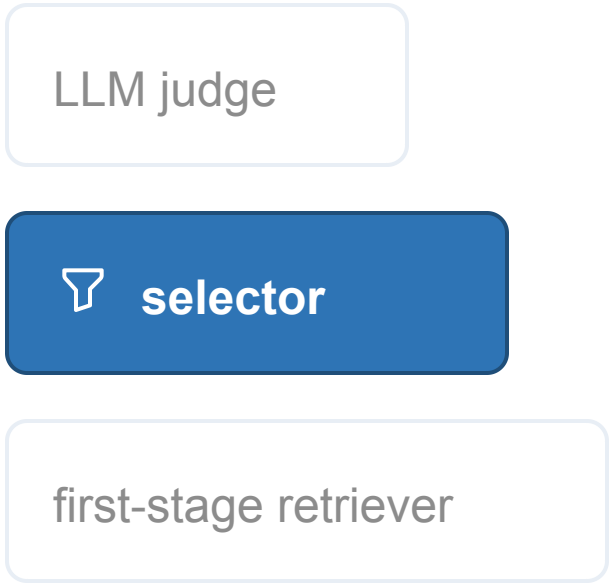
Using stronger LLMs in utility judgments for **utility-focused annotation**

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

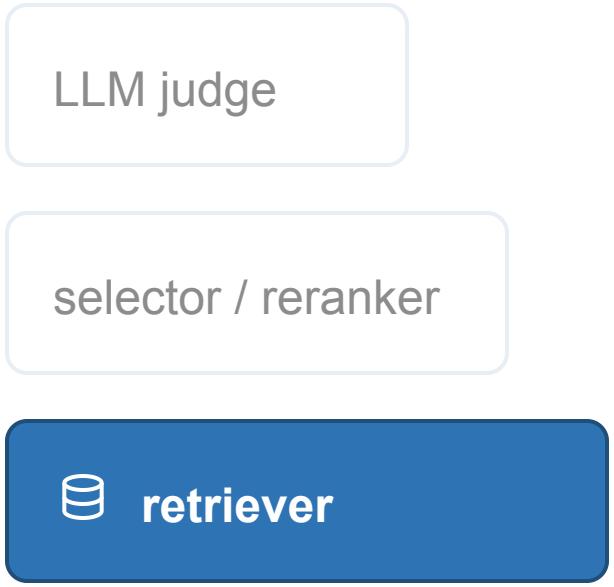
After **distillation from LLMs**, small model selects the **set** that meets the query's information needs.

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

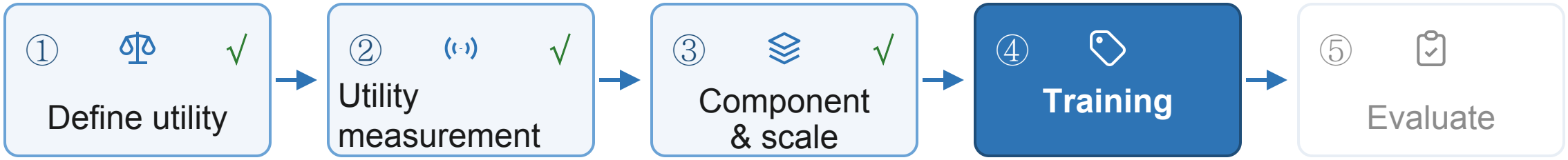
Documents as **latent variables** — EM-style joint training from answers alone.

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

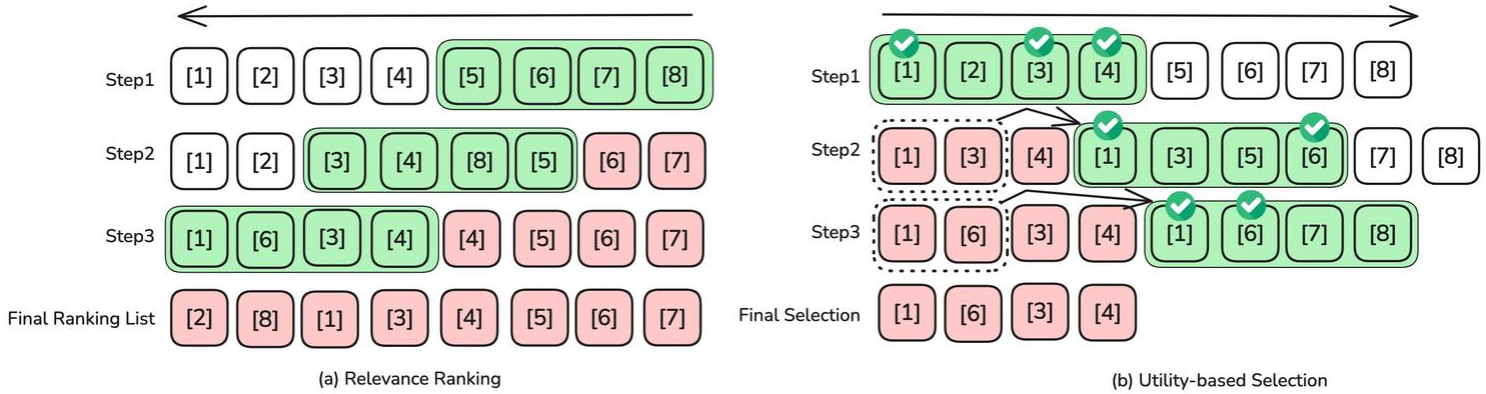
reward the selector when it helps the generator

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.

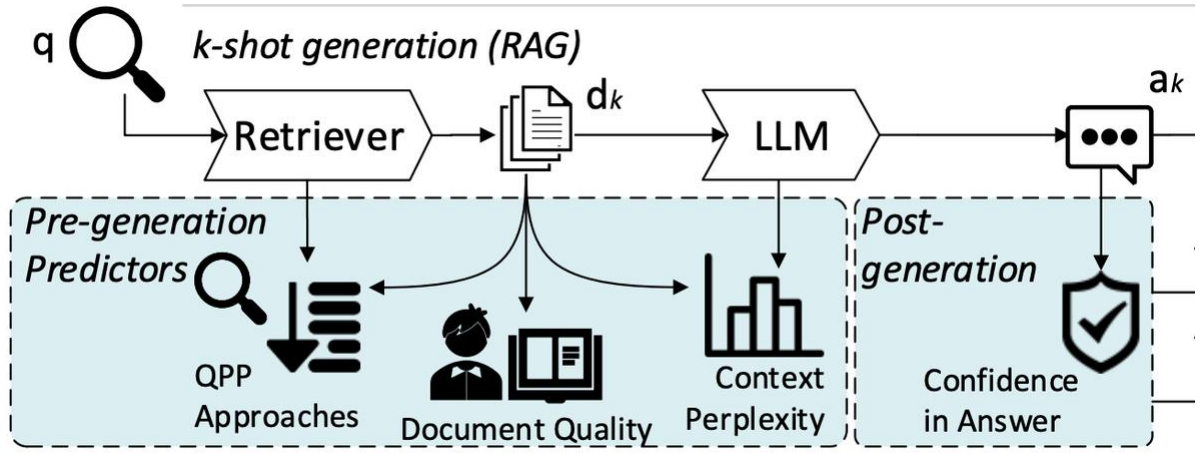
Utility measurement method

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

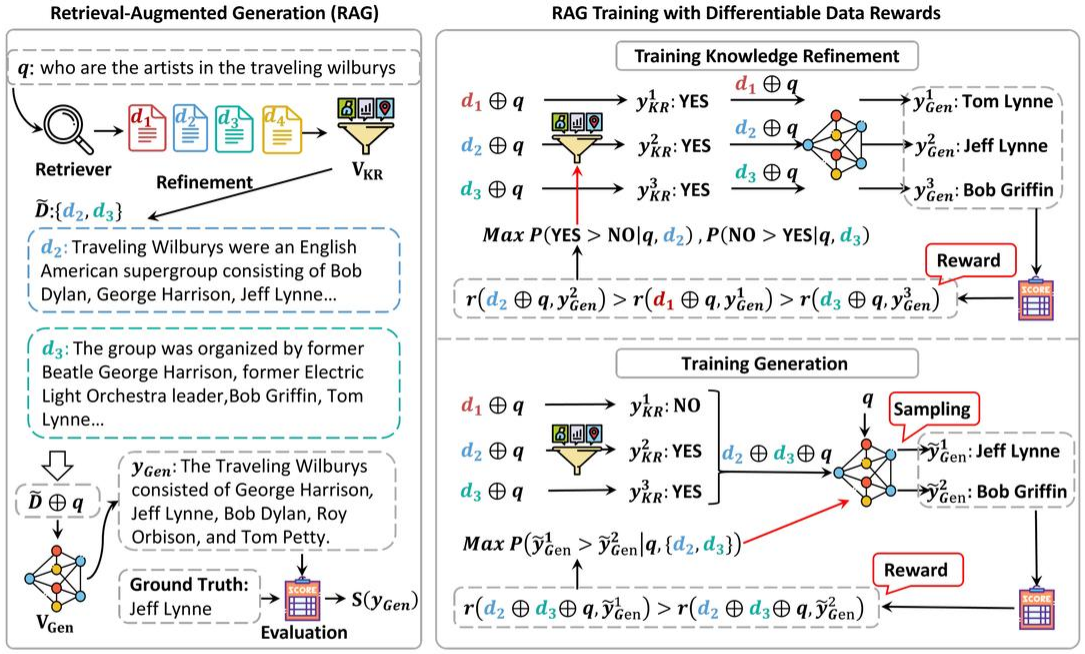


Figure 1: The Illustration of End-to-End Retrieval-Augmented Generation (RAG) Training with Our Differentiable Data Reward (DDR) Method. During training, we iteratively optimize the Generation module (V_{Gen}) and Knowledge Refinement module (V_{KR}).

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.

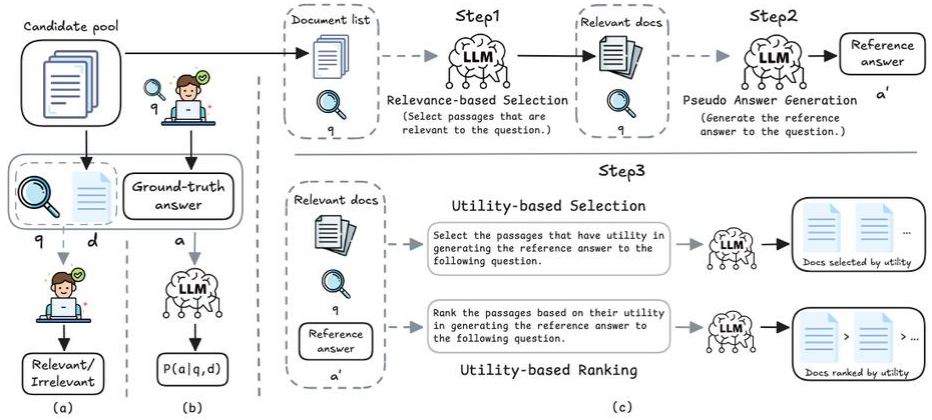
RAG
Stochastic RAG

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

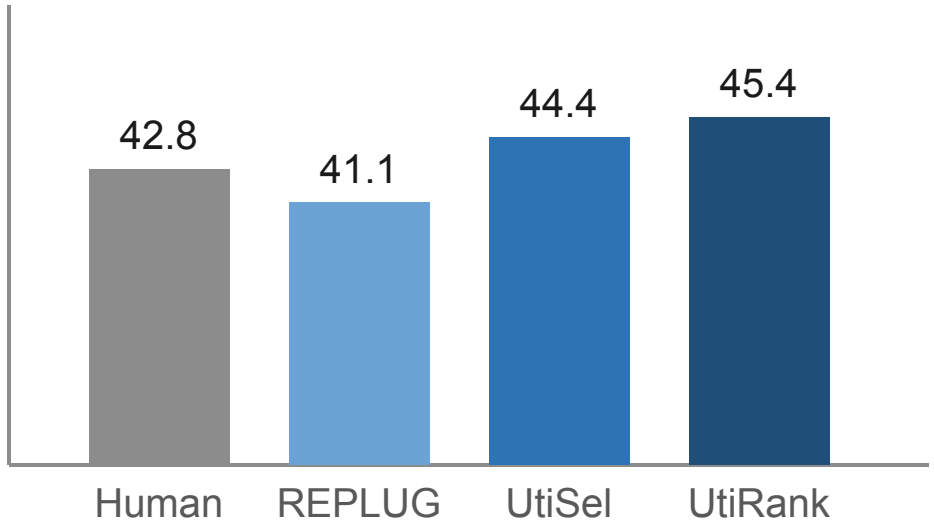
① Human annotation ② Ground-truth answer likelihood ③ **LLM utility judgment**

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.

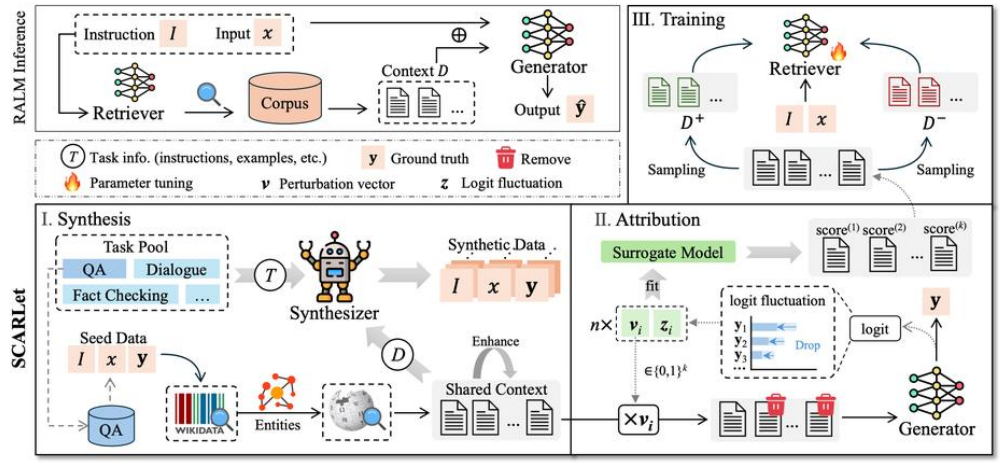
Objectives: Single / Joint / Sum-of-marginal likelihood — SumMargLH is most robust to noise.

> 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.

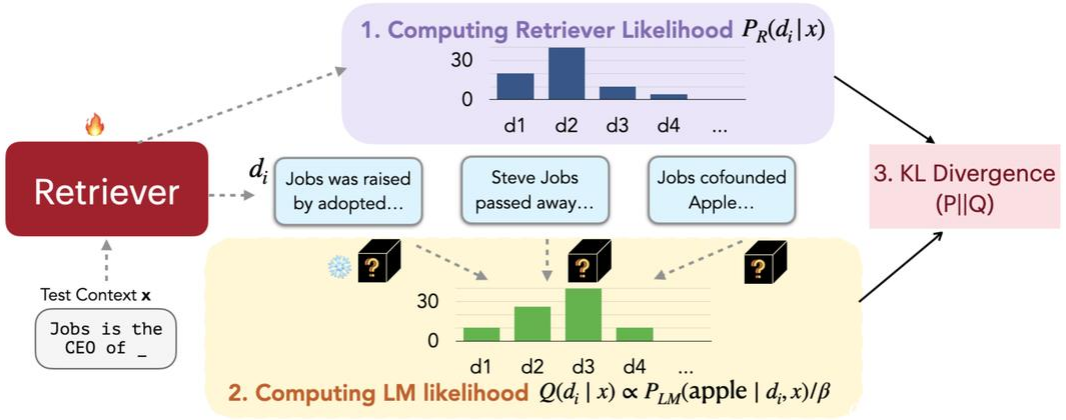
2. repeated retrieval and distillation, the retriever gradually becomes utility-aware..

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

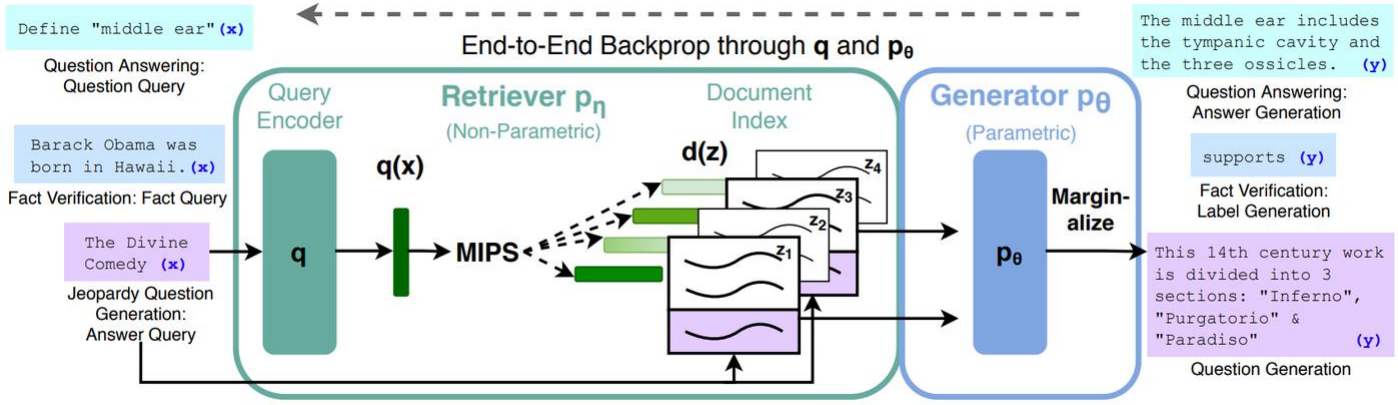
Likelihood as the teacher

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.

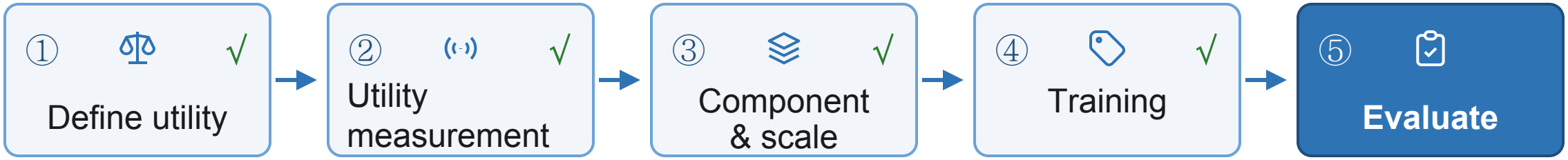
Training the Retriever & System

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