Goldilocks Zone

PREPRINT · 2026

Rethinking the Reranker

Boundary-Aware Evidence Selection for Robust Retrieval-Augmented Generation

BAR-RAG

Jiashuo Sun · Pengcheng Jiang · Saizhuo Wang · Jiajun Fan · Heng Wang
Siru Ouyang · Ming Zhong · Yizhu Jiao · Chengsong Huang · Xueqiang Xu
Pengrui Han · Peiran Li · Jiaxin Huang · Ge Liu · Heng Ji · Jiawei Han

UIUC · HKUST · WashU · Texas A&M

MOTIVATION

Higher recall ≠ better answers

RAG is brittle even when the right evidence **is in the top-K.**

Trivial shortcuts

Answer-revealing passages encourage pattern matching, not reasoning.

Insufficient evidence

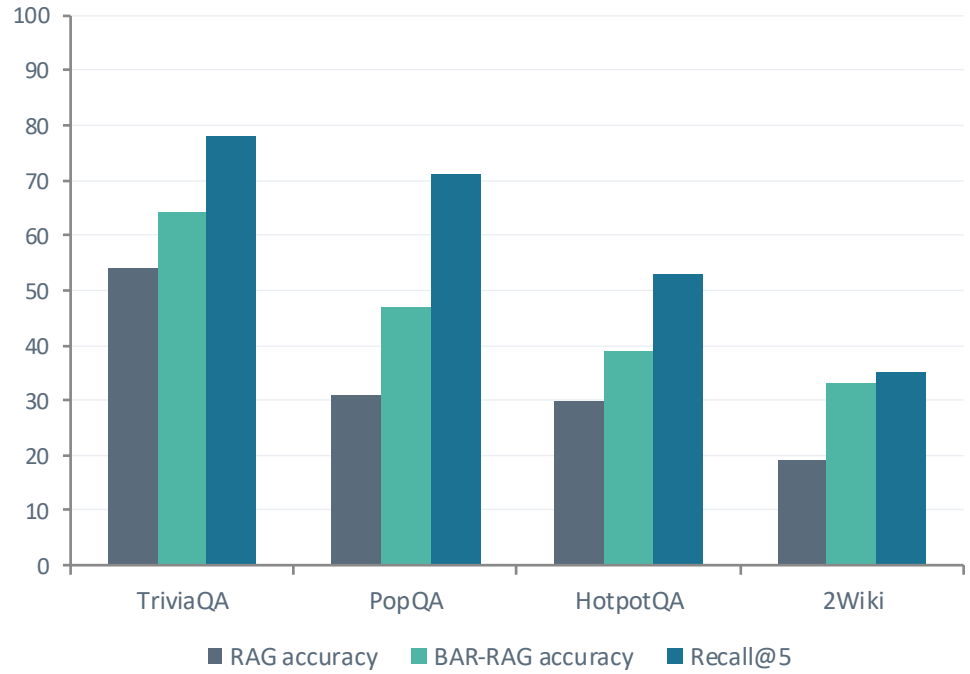
Top-K may include relevant but incomplete documents that can't support inference.

Train–test mismatch

Generators trained on curated evidence meet noisy retrieval at deployment.

Recall@5 vs. QA Accuracy

Retrieval recall stays high, but a vanilla RAG generator captures only a fraction of it.



KEY INSIGHT

From relevance scorer to evidence selector

STANDARD RERANKER

Maximize relevance score

Optimizes **query–document similarity** only — ignores what the generator actually needs.

Failure modes

- Surfaces answer-revealing passages → shortcut learning.
- Picks documents lacking the critical inference step.
- No estimate of generator competence.

BAR-RAG · EVIDENCE SELECTOR

Target the Goldilocks Zone

Pick evidence that is **challenging yet sufficient** for inference — near the generator's competence boundary.

What changes

- Reward = relevance + boundary signal from generator rollouts.
- Forces the generator to integrate, not pattern-match.
- Closes the train–test evidence-distribution gap.

CORE IDEA

The Goldilocks Zone of evidence

Goldilocks Zone: evidence sets S where the empirical correctness $\hat{p}(S) \approx c$ *(e.g., $c = 0.5$)*

Solvable enough that the generator can succeed sometimes — hard enough that it must reason.

Too easy

$\hat{p} \approx 1$

Answer is obvious from a single passage. Generator memorizes shortcuts.

shortcut learning

Just right

$\hat{p} \approx c$

Multi-document integration is required, but a correct answer is reachable.

strongest learning signal

Too hard

$\hat{p} \approx 0$

Critical evidence is missing. Reward is uniformly low — no signal.

uninformative

METHOD OVERVIEW

Two-stage reinforcement learning pipeline

Query +
Top-25 docs

STAGE 1 · Selector training

Train selector πr while generator πg is frozen.

Boundary reward R_bdy

Triangular peak at p̂(S) = c — reward concentrated near the competence boundary.

Relevance reward R_rel

Average retrieval score of selected docs (rescaled).

Format + count penalty

Gates malformed outputs; nudges |S| toward k* = 5.

⟶ *Optimize via GRPO with group-normalized advantages.*

STAGE 2 · Generator fine-tuning

Freeze πr, fine-tune πg on its induced evidence distribution.

Accuracy reward R_acc

Weighted F1 + Exact Match against gold answers.

Citation reward R_cite

Peaked at n* cited documents — grounds answers in evidence.

Format gate

Zero reward unless <answer>…</answer> is well-formed.

⟶ ***Closes the train–test evidence-distribution gap.***

Iterate Stage 1 ⇄ Stage 2 for T rounds

STAGE 1 DETAIL

Estimating p̂(S) and shaping the boundary reward

Procedure for each query q

1 Sample evidence sets

`S(m) ~ πr(· | q, C)  for  m = 1..M  (M = 8)`

2 Roll out the generator

`K answers per S(m) (K = 10), gated by reward threshold δ = 0.8`

3 Estimate solvability

`p̂(S) = (1/K) Σ 𝟙[Rg(a) ≥ δ]`

4 Shape the reward

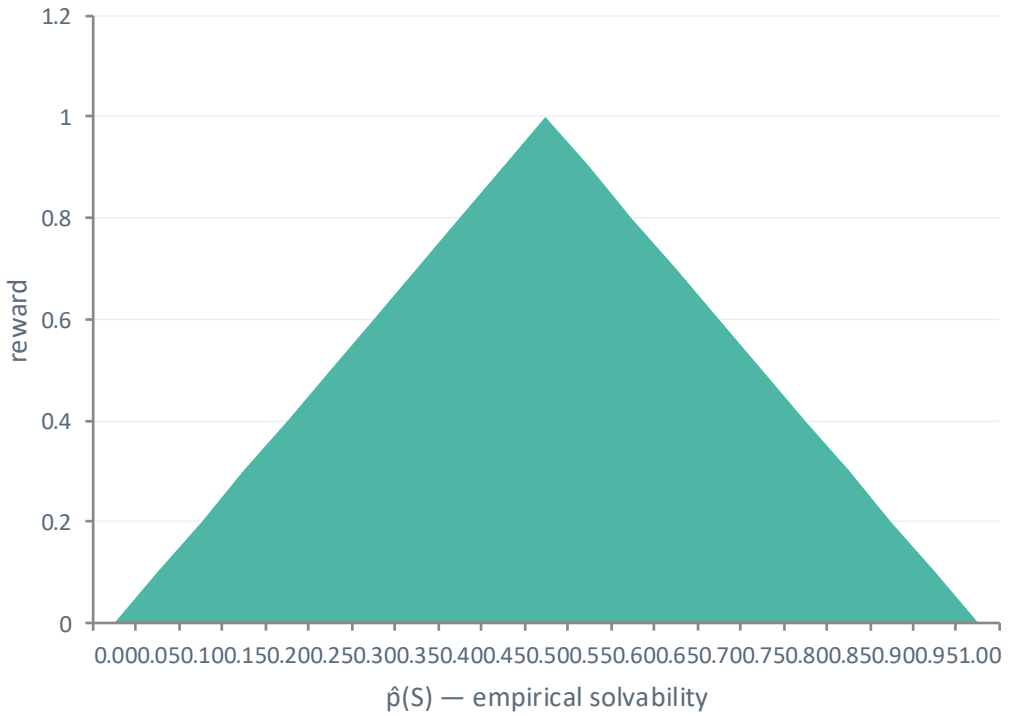
`Triangular peak at c = 0.5, then add relevance + count terms`

Training-time filtering

Drop queries with near-deterministic outcomes (always solved or never solved): they yield no usable RL signal.

Boundary reward R_bdy(S)

`R_bdy(S) = min( p̂/c ,  (1−p̂)/(1−c) )`



MAIN RESULTS

Consistent gains across 7 QA benchmarks

Method	NQ	TriviaQA	PopQA	HotpotQA	2Wiki	MuSiQue	Bamboogle	Avg.
Direct Inference	13.4	40.8	14.0	18.3	12.6	3.1	12.0	16.3
Chain-of-Thought	4.8	18.5	5.4	9.2	10.8	2.2	23.2	10.6
RAG	39.3	53.7	26.7	28.9	18.9	4.7	16.0	26.9
IRCoT	22.4	47.8	30.1	13.3	14.9	7.2	22.4	22.6
RAG + Reranker	40.5	55.3	27.3	28.1	20.4	5.5	18.7	28.0
RAG SFT	42.7	58.6	32.3	32.4	22.6	6.8	27.1	31.8
BAR-RAG (3 iter)	**46.9**	**64.5**	**46.9**	**38.8**	**29.8**	**9.1**	**39.6**	**39.1**

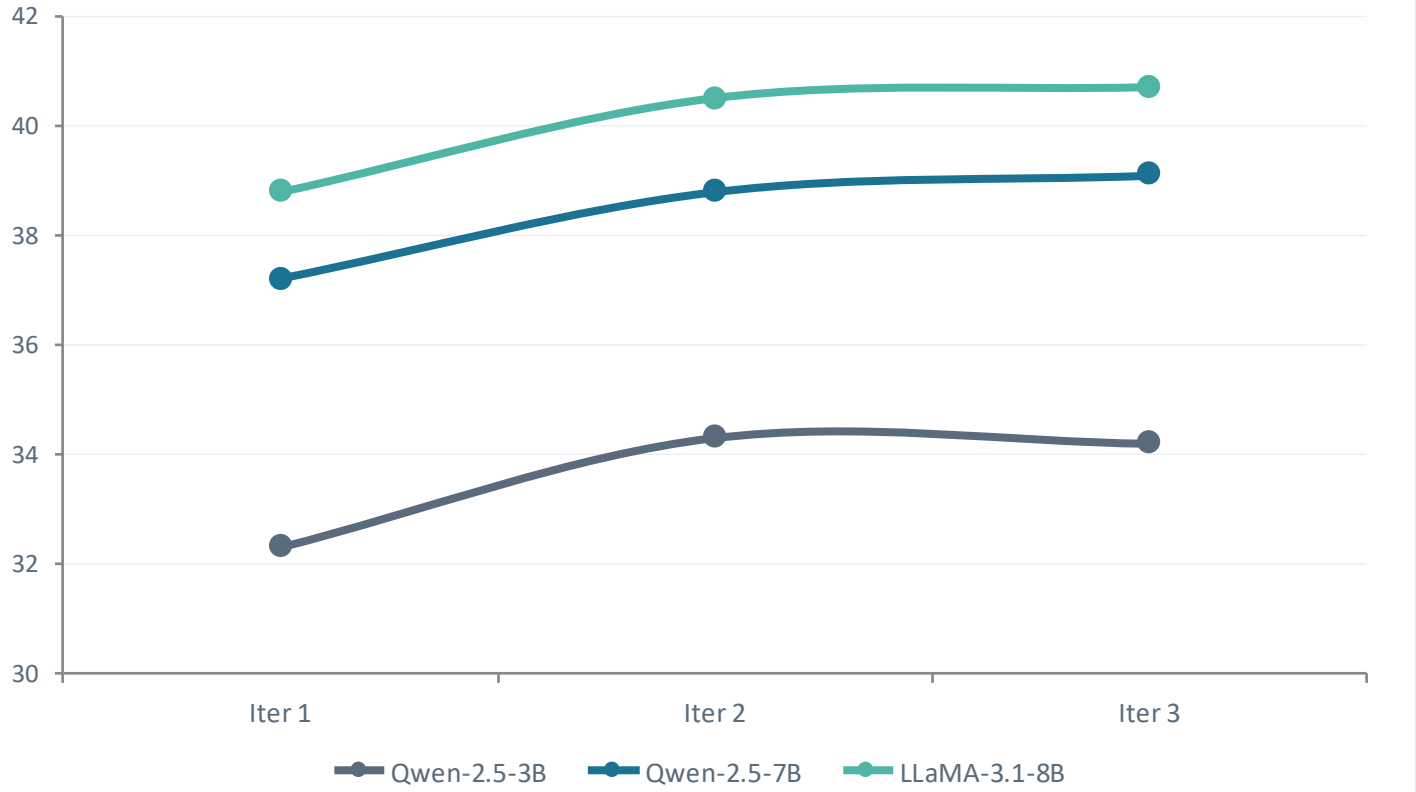
Backbone: Qwen-2.5-7B-Instruct

+10.3 | EM avg. gain

over the strongest RAG and reranking baselines, across 3 backbones × 7 datasets.

ITERATIVE TRAINING

Co-adaptation converges in ~2 rounds



Iter 1 → 2

Largest gains. Selector and generator align on a competence-matched evidence distribution.

Iter 2 → 3

Diminishing returns — the procedure converges to a stable solution.

Implication

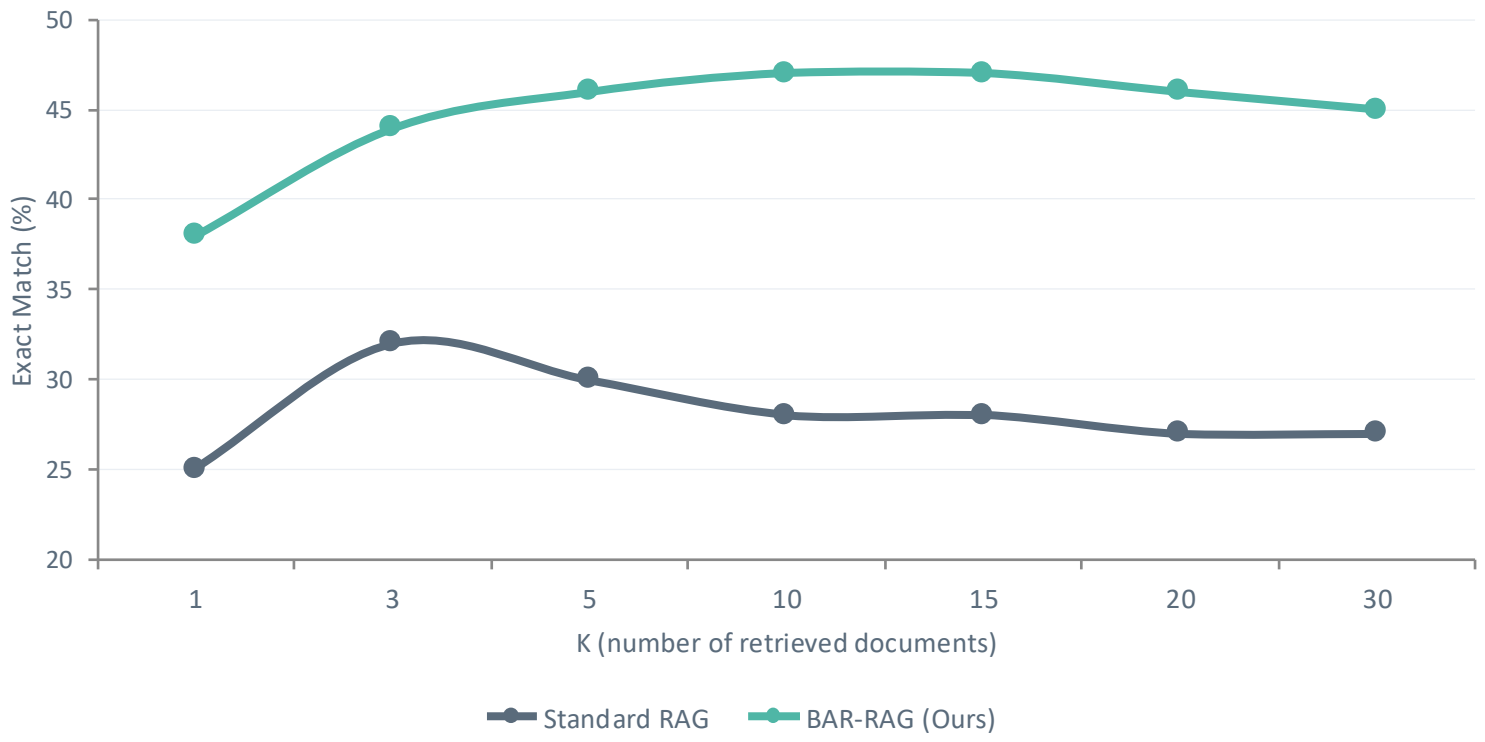
A few co-training rounds are enough; no expensive long-horizon RL needed.

ROBUSTNESS

BAR-RAG generators stay strong as K varies

Top-K accuracy on PopQA · LLaMA-3.1-8B-Instruct

Same naive retriever; only the generator differs.



WHY

Trained near the boundary, deployed anywhere

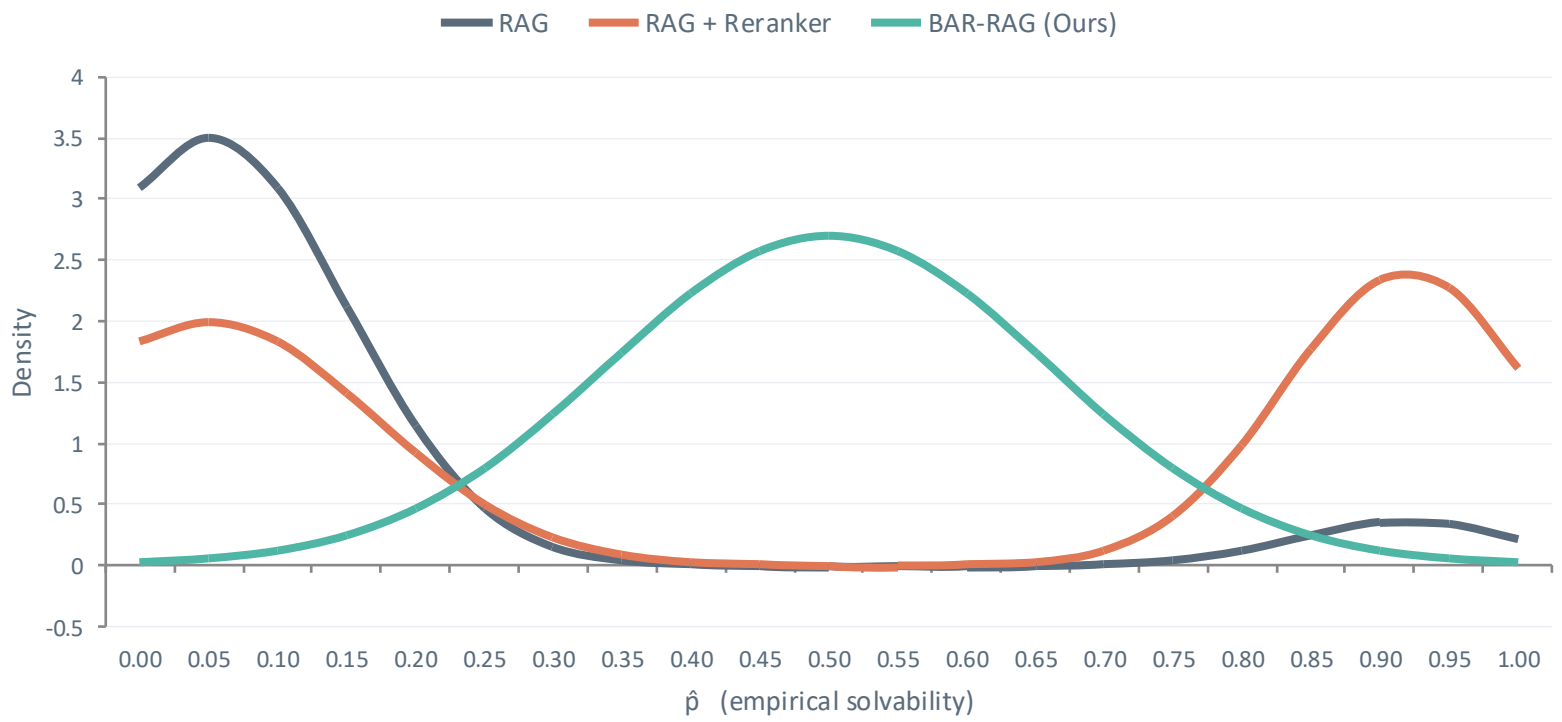
- **Largest gains in low-K**
  where evidence is sparse and reasoning robustness matters most.
- **Stable as K grows**
  BAR-RAG resists the noise that degrades vanilla RAG.
- **No selector at inference**
  Selector used only in training; serving cost is identical to standard RAG.

WHAT THE SELECTOR LEARNS

Reshaping the evidence-difficulty distribution

Empirical solvability p̂(S) of selected evidence · Qwen-2.5-7B

Naive RAG and reranker-based RAG are bimodal (extremes); BAR-RAG mass concentrates around c = 0.5.



Naive RAG

Mass at p̂ ≈ 0 — most evidence sets are unsolvable for the generator.

+ Reranker

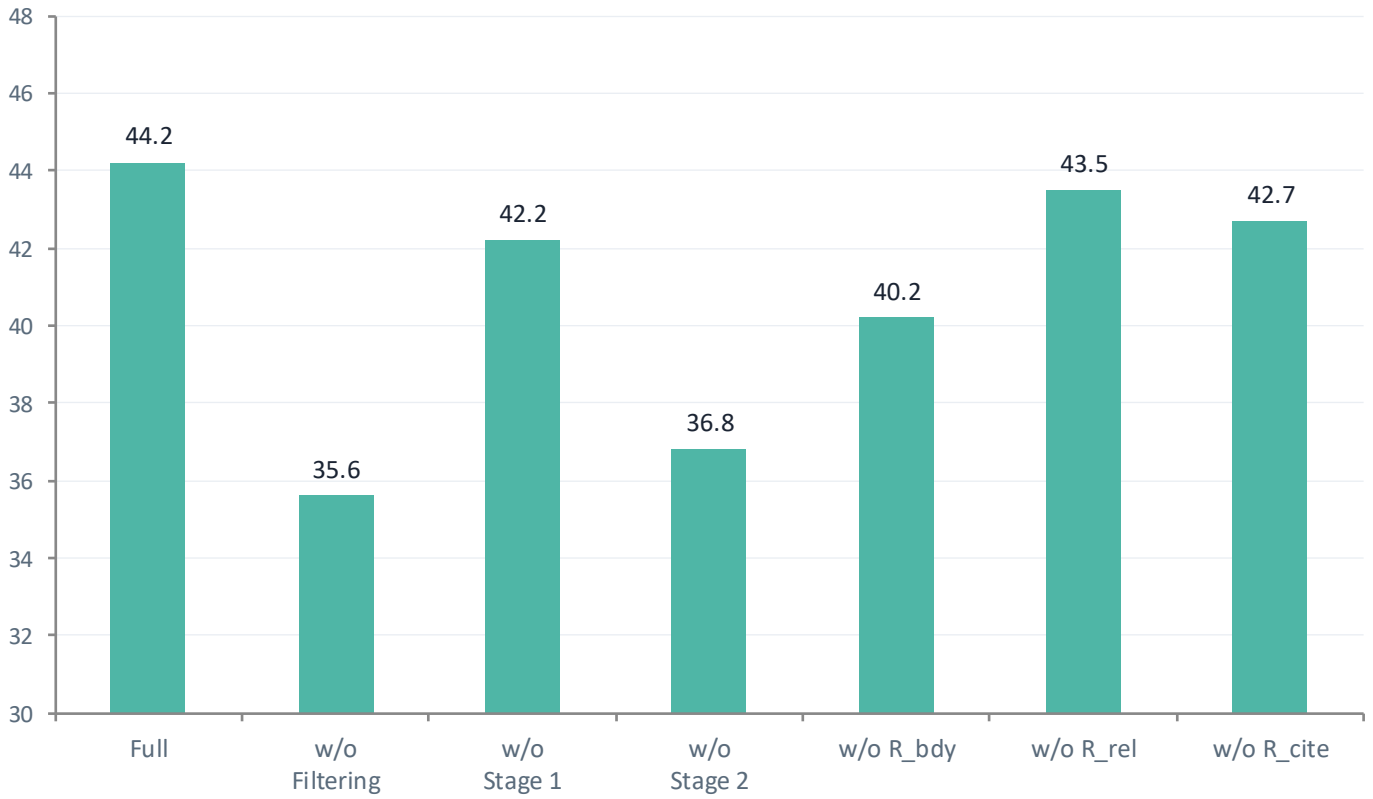
Shifts mass to p̂ ≈ 1 — answer-revealing passages, shortcut friendly.

BAR-RAG

Concentrates around p̂ ≈ c — strongest learning signal for the generator.

A B L A T I O N S

Every component contributes



Boundary reward

Removing R_bdy is the most damaging selector ablation — direct evidence for competence-aware selection.

Filtering matters

Without it, trivial / unanswerable queries inject noise into the RL signal and hurt PopQA & HotpotQA.

Both stages required

Removing Stage 2 hurts more than Stage 1: the generator must adapt to the boundary distribution.

Citation reward

Smaller but consistent gains — improves grounding without driving the headline numbers.

Takeaways

*Reranking is not just relevance — it is **evidence design**.*

Boundary-aware selection

Train the selector to find evidence near $\hat{p} = c$ — challenging yet solvable.

Two-stage RL

Iterating selector ⇄ generator closes the train–test evidence-distribution gap.

+10.3 EM avg. gain

Across 3 backbones × 7 QA benchmarks, with no extra inference-time cost.

Code & data:

`https://github.com/GasolSun36/BAR-RAG`

Thank you — questions welcome.