SD-Search: Hindsight Self-Distillation for Search-Augmented Reasoning

Zihan Liang[1], Yufei Ma[1], Zhipeng Qian[1], Ben Chen[2], Huangyu Dai, Lingtao Mao, Xuxin Zhang, Chenyi Lei[2], Wenwu Ou

Kuaishou Technology [1] Equal contribution [2] Corresponding authors benchen4395@gmail.com

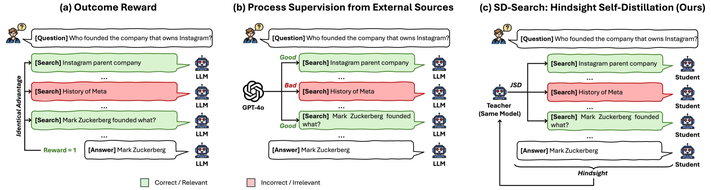
NEURAL INFORMATION PROCESSING SYSTEMS

Motivation

A final-answer reward gives every search decision the same trajectory-level advantage. It cannot directly distinguish a useful query from a redundant one.

Step-level signals from the policy itself

SD-Search uses one model under two contexts: an inference-time student view and a hindsight-conditioned teacher view. It needs no external teacher or extra process annotations.



Hindsight provides query-specific supervision within the existing RL loop.

Training and evaluation

Train on NQ and HotpotQA with GRPO. Evaluate Exact Match on NQ, TriviaQA, PopQA, HotpotQA, 2Wiki, MuSiQue and Bamboogle.

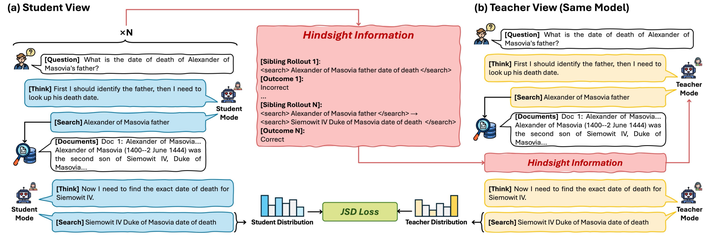
Fixed E5-base-v2 retriever, December 2018 Wikipedia and top-3 passages per query.

Cost and inference

At 3B scale, a 200-step run takes 11.9 hours versus 10.3 hours for AutoRefine on 8 × H800: 15.5% more training time.

Inference uses a single policy with no hindsight block or teacher forward pass.

Hindsight self-distillation



1. Build a hindsight block

Reuse the GRPO rollout group. Keep search spans and outcome labels, and mask reasoning, retrieved documents and answer spans to reduce answer leakage.

2. Match search-token distributions

The teacher and student share parameters. Stop gradients through teacher targets and apply Jensen–Shannon divergence only to tokens inside search-query spans.

GRPO loss + α × query-token JSD

The distillation term averages token-level JSD over search-query positions.

3. Warm up before distillation

Use 50 warmup steps, $\alpha_{SD} = 10^{-3}$ and top-k = 50. Labels use an F1 threshold of 0, so any positive answer overlap receives the “Correct” label.

“Correct” is a training-label convention, not an exact-match claim. Gold answers are required to score training rollouts.

Main results

Average EM across seven benchmarks. Qwen2.5 backbones.

Method	3B	7B
AutoRefine-Base	0.405	0.455
MR-Search-Base	0.414	0.460
Thinker-Instruct	0.430	0.452
SD-Search-Base	**0.428**	**0.471**
SD-Search-Instruct	**0.427**	**0.476**

+2.3 pp at 3B; +2.1 pp at 7B

3B Base and 7B Instruct versus AutoRefine-Base. Thinker uses a 72B teacher and an SFT stage.

Hindsight construction matters

3B-Base configuration	Avg. EM
Full SD-Search	**0.428**
No outcome labels	0.414
No rollout group	0.418
No future masking	0.398
Documents only	0.394

Scope

Training needs reliably scorable reference answers. Hindsight loses its outcome contrast when every rollout in the group has the same label.

Results and ablations are reported in the supplied camera-ready source. pp denotes percentage points.