

Teach a Man to Fish: Self-Adaptive Reinforced Distillation for In-Context Reasoning-based Reranker

Wenwei Wu^{1*}, Youhang Li^{1*}, Mingyang Chen^{1,2}, Jie Yang³, Xiaofeng Gu¹,
Duyi Pan¹, Yiwen Wu¹, Kaichen Zhang¹, Wei Wang^{1,2}, Zixuan Yuan^{1†},

¹The Hong Kong University of Science and Technology (Guangzhou)

²The Hong Kong University of Science and Technology

³The Hong Kong Polytechnic University

{wwu681, yhl627, dpan457, ywu240}@connect.hkust-gz.edu.cn

{mchenbt, kzhangbi}@connect.ust.hk, jessica-jie.yang@connect.polyu.hk

guxiaofengu@gmail.com, weiwcs@ust.hk, zixuanyuan@hkust-gz.edu.cn

Abstract

Recent advances in Large Reasoning Models (LRMs) are reshaping reasoning-intensive reranking, where step-by-step reasoning enables multi-faceted relevance judgments. To scale LRMs, existing approaches distill chain-of-thought traces from a teacher LRM into a student reranker via Supervised Fine-Tuning (SFT) followed by Reinforcement Learning (RL). However, this pipeline is misaligned with the distillation objective: teacher traces are too dense for direct imitation, while RL exploration is too sparse on hard cases. To this end, we propose **SARD** (Self-Adaptive Reinforced Distillation), a tri-mode RL framework that recasts teacher reasoning from a fixed supervision target into an adaptive resource the student draws on only when its own reasoning falls short. Triggered by rollout outcomes, SARD organizes training into three modes: *Free Exploration*, where the student reasons independently; *Guided Exploration*, which supplies reusable in-context guidance abstracted from teacher traces when free rollouts fail; and *Teacher-Guided Repair*, where the teacher diagnoses failures and revises, re-selects, or appends guidance, falling back to teacher trajectories only as a last resort. Experiments on reasoning-intensive reranking benchmarks show SARD outperforms strong reranking and SFT-then-RL baselines, with clear gains on difficult instances and under-represented reasoning patterns.

1 Introduction

The role of reranking is shifting as retrieval tasks become increasingly reasoning-intensive. In standard retrieve-then-read pipelines, a first-stage retriever narrows the candidate pool and a second-stage model reorders documents for downstream tasks (Izacard et al., 2023; Wang et al., 2024b; Thakur et al., 2021; Sun et al., 2023). Because

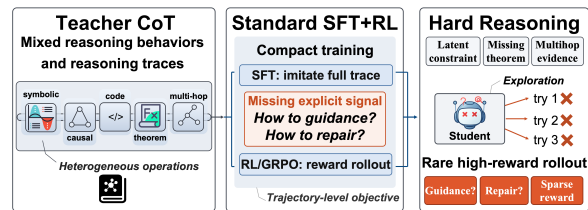


Figure 1: Limitations of SFT+RL distillation. Standard SFT+RL provides useful supervision from teacher traces and rollout-level rewards, but gives limited recovery signal on hard cases with failed exploration. SARD addresses this gap through guidance and failure repair.

modern relevance often hinges on complex latent constraints or multi-hop evidence (SU et al., 2025; Shao et al., 2025), recent approaches deploy Large Reasoning Models (LRMs) to emit explicit chain-of-thought (CoT) traces prior to ranking (Wei et al., 2022; DeepSeek-AI, 2025a). Since direct LRM inference at scale is computationally expensive, the prevailing recipe distills teacher CoT traces into smaller student rerankers via Supervised Fine-Tuning (SFT) and Reinforcement Learning (RL), notably Group Relative Policy Optimization (GRPO) (Shao et al., 2024; DeepSeek-AI, 2025a). Recent reasoning rerankers follow this direction (Weller et al., 2025; Zhang et al., 2025a; Liu et al., 2025), confirming that teacher rationales materially improve ranking quality when tasks demand deep inferential structure.

Despite this progress, existing pipelines still treat complete teacher traces or student rollouts as the primary learning object (Weller et al., 2025; Liu et al., 2025), as illustrated in Figure 1. This trace-centric view ignores a crucial challenge: teacher traces are highly heterogeneous, mixing qualitatively diverse reasoning operations such as symbolic checking, code interpretation, and multi-hop aggregation (SU et al., 2025; Shao et al., 2025). Standard trajectory-level objectives fail to explicitly disentangle *which* behavior is appropriate from

*Equal Contribution

†Corresponding Author

how to execute it. Consequently, on hard or under-represented cases, student exploration under the base prompt often fails repeatedly, leading to severe reward signal scarcity for standard RL, while flat supervised imitation merely provides unconditional demonstrations rather than a failure-aware recovery path (An et al., 2023). This motivates our central question: *how can a student reranker learn from teacher reasoning when its own exploration fails, without reducing distillation to blind trace imitation or introducing teacher dependence at inference time?*

To answer this, we present SARD (Self-Adaptive Reinforced Distillation), a tri-mode RL framework for adaptive recovery from failed exploration. As shown in Figure 2, SARD provides training-time support only when needed. In **Free Exploration**, the student first attempts each instance under the base prompt. Failed cases enter **Guided Exploration**, where compact guidance abstracted from teacher traces redirects exploration through reusable reasoning cues. If guidance still fails, **Teacher-Guided Repair** lets the teacher diagnose the failure and revise, reselect, or append guidance. When guided exploration or teacher-guided repair succeeds, SARD uses the successful assisted trajectory to update the base-prompt policy, encouraging the student to internalize assisted reasoning behavior while keeping deployment guidance-free.

In summary, our contributions are as follows:

- We propose SARD, a self-adaptive reinforced distillation framework for reasoning-intensive reranking, framing training as adaptive failure recovery rather than flat trace imitation.
- We introduce a failure-aware three-mode funnel that escalates support from free exploration to guided exploration and teacher-guided repair only upon failure.
- We design a bridge objective that uses successful trajectories obtained through guided exploration or teacher-guided repair to update the base-prompt policy, allowing training-time assistance to benefit the guidance-free reranker.
- We show that SARD outperforms strong reasoning-distillation baselines on challenging reasoning-intensive benchmarks.

2 Related Work

LLM-based and reasoning-intensive rerankers. LLM-based rerankers recast passage scoring as prompt-conditioned comparison, spanning list-

wise (Sun et al., 2023), pairwise (Qin et al., 2024), setwise (Zhuang et al., 2024), and efficient-decoding variants (Pradeep et al., 2023; Gangi Reddy et al., 2024). Reasoning rerankers further add test-time reasoning, RL, or trace distillation (Weller et al., 2025; Zhuang et al., 2025; Yang et al., 2025; Zhang et al., 2025a), though explicit CoT may hurt calibration or raise inference cost (Lu et al., 2025). Benchmarks such as BRIGHT, ReasonIR, and LIMRANK reveal surface matching’s insufficiency and the need for inferential reasoning (SU et al., 2025; Shao et al., 2025; Song et al., 2025). Closest to ours, ReasonRank combines teacher-synthesized data, cold-start SFT, and RL optimized via a multi-view reward (Liu et al., 2025). SARD shares this objective but targets failed-instance handling via progressive escalation, where the student first explores under the base prompt, receives compact guidance upon failure, and invokes teacher-guided repair only when the initial guidance is insufficient.

Reasoning distillation and post-training. Reasoning distillation transfers teacher reasoning into smaller models through teacher-generated rationales or outputs (Hsieh et al., 2023). As an SFT signal, full teacher traces can be dense and heterogeneous, making flat imitation difficult for smaller students (Li et al., 2025). On-policy distillation reduces teacher–student mismatch but typically requires dense supervision from a separate teacher (Agarwal et al., 2024). Recent self-distillation replaces this teacher with privileged-context self-teaching (Shenfeld et al., 2026; Zhao et al., 2026), but still mainly treats the signal as context-conditioned distribution matching rather than selective recovery support. In parallel, RL-based post-training methods such as GRPO and DeepSeek-R1 learn from outcome rewards rather than fixed traces (Shao et al., 2024; DeepSeek-AI, 2025a), yet unguided RL provides little corrective signal when hard instances repeatedly fail. Recent scaffolded or off-policy methods, such as LUFFY (Yan et al., 2025), Critique-GRPO (Zhang et al., 2025c), and Scaf-GRPO (Zhang et al., 2026), address this exploration bottleneck via external support such as off-policy traces, failure-triggered critiques, or hierarchical hints. SARD extends scaffolded failure recovery to reasoning-intensive reranking, where hard cases require refining evidence selection and candidate comparison. It adaptively repairs compact guidance and distills suc-

successful assisted trajectories into the base-prompt policy, requiring no guidance retrieval or teacher feedback at inference.

3 Method

We present SARD, a self-adaptive reinforced distillation framework for reasoning-intensive listwise reranking (Sun et al., 2023; Pradeep et al., 2023; Liu et al., 2025). The framework uses supervised imitation only as a cold start and then organizes RL into a three-mode adaptive funnel: free exploration, guided exploration, and teacher-guided repair. A key design is a *shaped bridge objective*: for successful trajectories obtained through guided exploration, teacher-guided repair, or teacher fallback, SARD computes token probabilities under the base prompt and applies a bounded policy-shaping function (Yan et al., 2025), turning assisted successes into base-prompt learning signals.

3.1 Problem Setup

A training instance is denoted by $x = (q, \mathcal{D})$, where q is the query and $\mathcal{D} = \{d_1, \dots, d_K\}$ is a candidate document list.

A reranker policy π_θ generates a trajectory $y = (c, \sigma)$, where c denotes the reasoning trace and $\sigma \in \mathfrak{S}_K$ denotes the final permutation of candidate indices (Sun et al., 2023), defined as a bijection from $\{1, \dots, K\}$ to itself, with $\sigma(i)$ specifying the rank of document d_i (Pradeep et al., 2023).

A reward function $R(x, y)$ maps the generated ranking to a scalar reward, such as a combination of $NDCG@k$, recall-oriented terms, and format validity (Liu et al., 2025; Zhang et al., 2025a).

Each instance comprises two data objects: i) a teacher-generated trajectory y^T , from a stronger LRM (Hsieh et al., 2023); ii) an initial reasoning guidance instruction g , assigned from a guidance library $\mathcal{G}$ mined from teacher-generated trajectories; alongside two hyperparameters: i) a success threshold δ , used to decide whether exploration is sufficient; ii) an optional diagnostic budget T , the maximum number of teacher-guided repair rounds.

The guidance library construction and assignment pipeline is an input data-construction component rather than the central training contribution; we summarize it in Appendix A. During inference, the deployed reranker receives only x and never receives g or y^T .

3.2 Overview of SARD

SARD treats reasoning distillation as adaptive failure recovery as illustrated in Figure 2. Instead of applying the same supervision to every instance, it asks how much training-time help the student needs to produce a high-reward reasoning trajectory. To operationalize this, each training instance is dynamically routed through a three-mode reinforcement learning funnel: free exploration, guided exploration, and teacher-guided repair.

The funnel follows a minimal-guidance principle: each instance first uses base-prompt exploration, where successful rollouts are trained with standard GRPO (Shao et al., 2024). Failed instances are retried with compact guidance; if guidance also fails, the teacher repairs the guidance by revising, reselecting, or appending it, with teacher-generated trajectories used only as a last resort. Whenever guidance or repair succeeds, the successful assisted trajectory is bridged back to the base-prompt policy.

This design is related to context-space GRPO, but uses the auxiliary context only as training-time support. Training-Free GRPO shifts a frozen policy from $\pi_\theta(y | q)$ to $\pi_\theta(y | q, E)$ by updating an external prompt-space experience prior E (Cai et al., 2025). In contrast, SARD transiently samples from a guidance-conditioned behavior policy $\pi_{\theta_{\text{old}}}(y | x, g)$ after base-prompt failure, while optimizing the deployable target policy $\pi_\theta(y | x)$ through the shaped bridge objective. Thus, guidance serves as training signal for the base-prompt policy, not as an inference-time token prior.

Algorithm 1 formalizes this mode-switching process. As detailed above, the core mechanism guarantees a strict minimal-guidance progression: the amount of training-time intervention increases only when the previous, less-intrusive mode fails. Crucially, regardless of the intervention level required to rescue a failed trajectory, every successful rollout is ultimately bridged back to optimize the inference-time base prompt, preserving the student’s independent reasoning capability.

Cold-start. Before the adaptive RL stage, we optionally apply a lightweight supervised warm-up to reduce invalid early rollouts and teach the student the output format (DeepSeek-AI, 2025a; Sun et al., 2025). This step is used only for optimization stability and is not the main distillation mechanism. Unless otherwise stated, all adaptive behavior described below refers to the subsequent RL

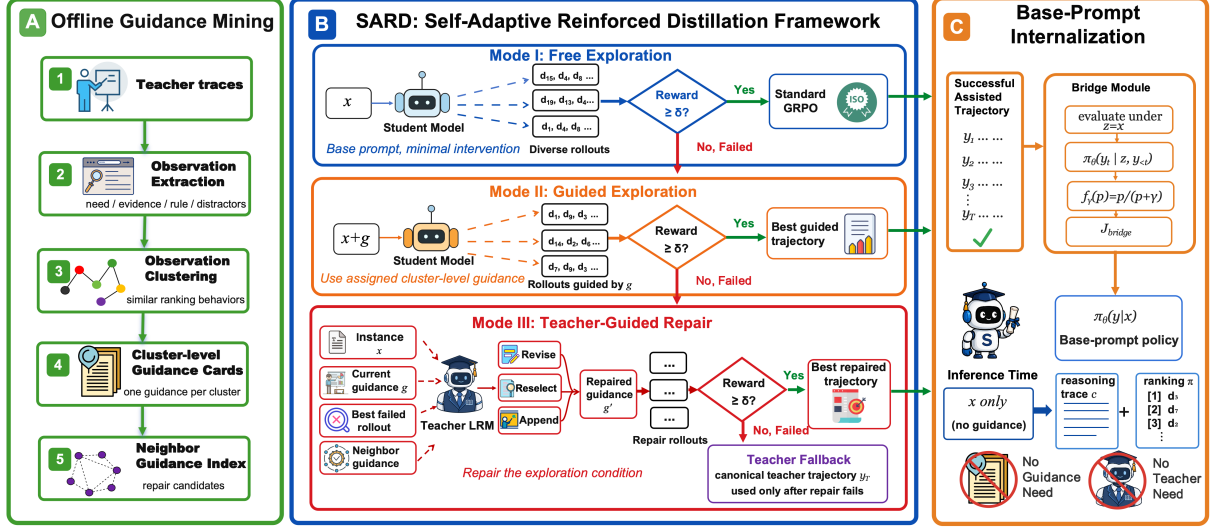


Figure 2: **Overview of SARD.** A tri-mode adaptive reinforcement distillation framework that treats teacher LRM reasoning as training-time failure recovery support.

stage, where teacher trajectories are used only after free exploration, guided exploration, and teacher-guided repair fail. Details of the warm-up objective are deferred to Appendix B.

3.3 Mode I: Free Exploration

At each RL update, SARD first tests whether the current student can solve an instance without guidance. For a prompt x , we sample a group of G trajectories from the old policy,

$$\mathcal{Y}^0 = \{y_1^0, \dots, y_G^0\} \sim \pi_{\theta_{\text{old}}}(\cdot | x), \quad (1)$$

and compute rewards $r_j^0 = R(x, y_j^0)$. If $\max_j r_j^0 \geq \delta$, the instance enters free exploration mode. Following GRPO (Shao et al., 2024), we estimate a group-relative advantage

$$A_j = \frac{r_j - \text{mean}(\{r_\ell\}_{\ell=1}^G)}{\text{std}(\{r_\ell\}_{\ell=1}^G) + \epsilon_A}, \quad (2)$$

and optimize a PPO-style clipped surrogate (Schulman et al., 2017) under the base prompt x . This mode reinforces independent reasoning and is the only mode used for instances the student can already explore successfully.

3.4 Mode II: Guided Exploration

For failed free-exploration instances, SARD changes the student’s exploration condition by retrying the instance with the assigned reasoning guidance. This mode is still student-driven: the teacher does not provide a new trajectory, and successful rollouts are generated by the student with additional guidance (Zhang et al., 2025c).

Guided rescue. If free rollouts fail, we use the assigned guidance instruction g and sample

$$\mathcal{Y}^g = \{y_1^g, \dots, y_G^g\} \sim \pi_{\theta_{\text{old}}}(\cdot | x, g). \quad (3)$$

If $\max_j R(x, y_j^g) \geq \delta$, the instance is guidance-rescueable. Let

$$y^{g,*} = \arg \max_{y \in \mathcal{Y}^g} R(x, y) \quad (4)$$

be the best guided trajectory. We perform two updates. First, we run standard GRPO (Shao et al., 2024) under the guidance-augmented prompt (x, g) to improve the student’s ability to use the guidance. Second, we bridge $y^{g,*}$ back to the base prompt x using the shaped bridge objective in Section 3.6. This second update is crucial: without it, the model may improve only when guidance is provided; with it, the final policy internalizes the rescued trajectory under the base prompt used at inference time.

3.5 Mode III: Teacher-Guided Repair

If the assigned guidance also fails, SARD enters teacher-guided repair. This mode first repairs the guidance-level exploration condition and only then falls back to teacher-generated trajectories.

Teacher-guided repair. If the assigned guidance also fails, SARD asks whether the failure reflects a true student limitation or a misaligned guidance instruction. For at most T rounds, a teacher LRM (DeepSeek-AI, 2025b) receives the ranking instance x , the current guidance $g^{(t-1)}$, the student’s best failed trajectory, and the guidance library $\mathcal{G}$ (Zhang et al., 2025c). The teacher chooses

one of three actions: (i) **Revise**: rewrite the current guidance to make the evidence criteria or comparison rule more precise; (ii) **Reselect**: replace the current guidance with a more suitable existing guidance instruction from $\mathcal{G}$; (iii) **Append**: add a new guidance instruction when no existing guidance covers the observed failure mode.

The resulting guidance $g^{(t)}$ is used for another guided rollout. If any repaired guidance produces a rollout above δ , the instance is updated exactly like a guidance-rescuable instance: GRPO under $(x, g^{(t)})$ and the shaped bridge objective under x . This diagnosis loop lets the guidance library evolve around hard instances while maintaining a bounded teacher budget, but its contents are not queried at inference time.

Teacher-guided fallback. If all T diagnosis rounds fail, we use the teacher-generated trajectory y^T as a final corrective signal (Hinton et al., 2015; Hsieh et al., 2023). Unlike ordinary off-policy correction (Schulman et al., 2017; Espeholt et al., 2018), the shaped bridge does not require token probabilities from the teacher behavior policy. Therefore, a verified teacher trajectory can be inserted into the bridge group under the base prompt x and optimized with the same shaping mechanism used for guided rescues. When the teacher trajectory is not verifiably high reward or when the bridge group has zero reward variance, we additionally use an auxiliary imitation term (Ouyang et al., 2022; Agarwal et al., 2024)

$$\mathcal{L}_{\text{fallback}}(\theta) = -\log \pi_{\theta}(y^T | x), \quad (5)$$

optionally paired with the guidance-augmented view. This fallback is used only as a last resort and is not treated as an on-policy GRPO sample.

3.6 Shaped Off-Policy Bridge

When free exploration fails, SARD may obtain a high-reward trajectory under an auxiliary behavior prompt b , such as a guidance-augmented prompt (x, g) , a repaired prompt $(x, g^{(t)})$, or a teacher-generation context. The bridge update transfers such a trajectory back to the deployable target prompt $z = x$. Thus, the objective is not to make the behavior trajectory appear on-policy, but to internalize a verified rescue trajectory into the base-prompt policy used at inference time.

Let $\mathcal{Y}$ denote the bridge group under target prompt z . For each bridged trajectory $y_j = (y_{j,1}, \dots, y_{j,|y_j|})$, we evaluate each token under

the target prompt rather than the behavior prompt, where $0 < \gamma < 1$:

$$\begin{aligned} p_{j,t}^z(\theta) &= \pi_{\theta}(y_{j,t} | z, y_{j,<t}), \\ s_{j,t}(\theta) &= f_{\gamma}(p_{j,t}^z(\theta)) = \frac{p_{j,t}^z(\theta)}{p_{j,t}^z(\theta) + \gamma}. \end{aligned} \quad (6)$$

The shaping function follows recent off-policy and refinement-based GRPO methods (Yan et al., 2025; Zhang et al., 2025c), but is applied to the target-prompt probability $p_{j,t}^z$ rather than to the behavior-context likelihood. Since $f_{\gamma}(p)$ is bounded and saturates for large p , it reduces over-emphasis on already likely tokens while preserving learning signal for successful but currently unfamiliar reasoning actions.

Rewards are computed with the original instance x , and group-relative advantages are estimated within the bridge group:

$$A_j = \frac{R(x, y_j) - \text{mean}_{y_k \in \mathcal{Y}} R(x, y_k)}{\text{std}_{y_k \in \mathcal{Y}} R(x, y_k) + \epsilon_A}. \quad (7)$$

The bridge objective is

$$\begin{aligned} \mathcal{J}_{\text{bridge}}(\theta) &= -\beta \text{KL}(\pi_{\theta} \| \pi_{\text{ref}}) \\ &+ \frac{1}{|\mathcal{Y}|} \sum_{y_j \in \mathcal{Y}} \frac{\alpha_j}{|y_j|} \sum_{t=1}^{|y_j|} \min \left(s_{j,t}(\theta) A_j, \right. \\ &\quad \left. \text{clip}(s_{j,t}(\theta), 1 - \epsilon, 1 + \epsilon) A_j \right) \end{aligned} \quad (8)$$

where α_j is a sequence-level coefficient, set to 1 for guided or repaired student rollouts and optionally to a confidence or reward-margin weight for teacher fallback trajectories.

Our bridge is related to scaffolded and off-policy RL approaches that use hints, critiques, or external traces to overcome sparse exploration (Yan et al., 2025; Zhang et al., 2025c, 2026). However, unlike hint-guided on-policy optimization, SARD evaluates rescued trajectories under the inference-time target prompt $z = x$, so training-time support is internalized into a guidance-free deployable policy.

3.7 Inference

At inference time, SARD uses only the base-prompt policy $\pi_{\theta}(\cdot | x)$. No guidance retrieval, teacher diagnosis, teacher-generated trajectory, or bridge shaping is needed. This is essential for scalability: reasoning guidance and teacher feedback are training-time aids for closing the teacher-student gap, not additional inference-time dependencies.

4 Experiments

We evaluate SARD as a training-time reinforced distillation framework for reasoning-intensive listwise reranking. Our experiments address four key research questions: **RQ1**, whether SARD improves performance on reasoning-intensive reranking benchmarks; **RQ2**, how reward-driven RL, guided exploration, and teacher-guided repair contribute beyond supervised teacher-trace distillation; **RQ3**, how SARD allocates assistance across training domains; and **RQ4**, whether SARD retains a base-prompt inference interface without requiring explicit test-time guidance.

4.1 Experimental Setup

Benchmarks and retrieval. Following the evaluation protocol of ReasonRank (Liu et al., 2025), we evaluate on two reasoning-intensive reranking benchmarks: **BRIGHT** (SU et al., 2025) and **R2MED** (Zhang et al., 2025b). BRIGHT covers complex QA, coding, mathematics, and theorem retrieval, while R2MED focuses on biomedical and clinical retrieval. We use the top-100 candidate sets released by ReasonRank: BRIGHT candidates are retrieved by ReasonIR (Shao et al., 2025), while R2MED candidates are retrieved by E5-mistral (Wang et al., 2024a). All rerankers use the same listwise sliding-window protocol with 20 passages per window and a stride of 10.

Training and evaluation details. We use the reasoning-intensive reranking training instances from ReasonRank (Liu et al., 2025) as source training data, with domain annotations used only in Section 4.4 to aggregate training-time assistance behavior. We report NDCG@10 as the primary metric. The student is initialized from Qwen2.5-7B-Instruct (Yang et al., 2024); the teacher is DeepSeek-V3.2 (DeepSeek-AI, 2025b) and is used only during training for trace generation, guidance-library construction, and teacher-guided repair. At inference, SARD uses only the trained student with the base prompt. Training is conducted on 8 NVIDIA A800 80GB GPUs.

Baselines. We compare with the first-stage retriever and representative rerankers, including non-reasoning rerankers RankT5 (Zhuang et al., 2023) and RankZephyr (Pradeep et al., 2023), and reasoning rerankers Rank1 (Weller et al., 2025), RankR1 (Zhuang et al., 2025), Rearrank (Zhang et al., 2025a), and ReasonRank (Liu et al., 2025). Published baselines are reported under their original

settings. To isolate SARD’s training design, we use an internally matched SFT+GRPO baseline using the same student initialization, source data, teacher-trace warm-up pipeline, evaluation protocol, and base-prompt deployment condition.

Guidance-library construction is described in Appendix A; reward construction, data allocation, and training details are in Appendix B; prompt templates and case studies are in Appendix C; and additional analyses, including BEIR evaluation with BM25 top-100 candidates (Thakur et al., 2021) and backbone-transfer experiments, are in Appendix D.

4.2 Main Results on Reasoning-Intensive Benchmarks (RQ1)

Table 1 reports results on BRIGHT. SARD achieves the best average NDCG@10 of 36.78, improving over ReasonRank by 1.04 points and Rearrank by 5.03 points. It also improves over the first-stage ReasonIR ranking by 6.19 points. To isolate the effect of SARD’s training design, Section 4.3 further reports internally matched comparisons with SFT+GRPO.

At the subset level, the gains are broad but not uniform. SARD ranks first on 7 of 12 BRIGHT subsets and second on 3 others. The largest gains over ReasonRank appear on sustainable living ($\uparrow 3.03$), LeetCode ($\uparrow 2.65$), earth science ($\uparrow 2.44$), economics ($\uparrow 2.07$), and StackOverflow ($\uparrow 2.06$). At the same time, SARD is second on Psychology, AoPS, and TheoT, and trails more clearly on LeetCode and Pony. Overall, SARD obtains the best average while remaining first or second on most subsets, indicating stronger aggregate performance under the BRIGHT evaluation protocol.

Table 2 evaluates whether the gains transfer to biomedical and medical reranking scenarios. SARD achieves the best average NDCG@10 of 40.50, improving over ReasonRank by $\uparrow 0.97$ points and over Rearrank by $\uparrow 5.44$ points. It ranks first on MedS, MedE, PMCT, PMCC, and IYiC, while ReasonRank remains stronger on Biology, Bioinformatics, and MedD. Thus, SARD achieves the strongest average on R2MED, suggesting that the adaptive training schedule transfers beyond BRIGHT while still leaving subset-level variation.

4.3 Ablation Study (RQ2)

We conduct ablations to assess the contributions of teacher-trace distillation, base-prompt RL, reusable guidance, teacher-guided repair, the bridge update, and bounded shaping (Table 3). The SFT baseline

Model	Econ.	Earth.	Rob.	Bio.	Psy.	Stack.	Sus.	Leet.	Pony	AoPS	TheoQ.	TheoT.	Avg.
<i>First-stage retriever</i>													
ReasonIR (8B)	32.65	43.00	20.82	43.49	39.57	30.96	27.34	31.69	19.55	7.37	33.93	36.68	30.59
<i>Non-reasoning rerankers</i>													
RankT5 (3B)	11.35	22.11	10.94	13.62	11.40	11.35	15.96	27.45	38.05	9.24	18.27	9.46	16.60
RankZephyr (7B)	19.87	17.36	12.35	34.90	24.72	13.35	22.34	<u>29.29</u>	<u>32.37</u>	6.05	28.98	30.07	22.64
<i>Reasoning rerankers</i>													
Rank-R1 (7B)	20.19	27.91	18.13	36.70	30.22	11.32	29.28	17.13	9.35	3.23	14.43	28.98	20.57
Rank1 (7B)	25.34	38.91	16.77	39.91	35.32	24.81	33.47	12.71	28.14	2.58	30.67	38.18	27.23
Rearank (7B)	30.65	40.27	27.16	46.94	40.77	26.12	36.26	30.55	22.84	7.31	32.28	39.86	31.75
ReasonRank (7B)	<u>35.05</u>	<u>47.75</u>	<u>31.22</u>	<u>56.70</u>	47.77	<u>32.47</u>	<u>40.87</u>	23.17	24.95	7.68	<u>39.49</u>	41.80	<u>35.74</u>
SARD (7B)	37.12	50.19	31.66	57.85	<u>47.30</u>	34.53	43.90	25.82	23.74	<u>7.78</u>	40.35	<u>41.12</u>	36.78

Table 1: Main results on BRIGHT. All methods rerank the top-100 passages released by ReasonRank. We report NDCG@10. Among rerankers, best and second-best scores are boldfaced and underlined, respectively.

Models	Biology	Bioin.	MedS.	MedE.	MedD.	PMCT.	PMCC.	IlyiC.	Avg.
<i>First-stage retriever</i>									
E5-mistral (7B)	18.28	41.47	41.01	6.44	11.38	19.81	30.97	21.37	23.84
<i>Non-reasoning reranker</i>									
RankT5 (3B)	13.20	32.84	23.38	2.09	4.18	0.61	14.55	12.39	12.91
RankZephyr (7B)	22.86	43.07	48.24	6.97	10.45	26.64	7.78	14.59	22.58
<i>Reasoning reranker</i>									
Rank-R1 (7B)	34.04	51.63	50.96	12.79	21.96	34.79	31.70	25.08	32.87
Rank1 (7B)	32.64	55.57	54.74	12.78	19.98	34.40	30.17	18.15	32.30
Rearank (7B)	38.35	50.91	59.73	14.06	19.34	37.49	34.29	26.27	35.06
ReasonRank (7B)	46.80	59.70	<u>60.11</u>	<u>16.48</u>	24.92	<u>39.21</u>	<u>39.13</u>	<u>29.85</u>	<u>39.53</u>
SARD (7B)	<u>46.11</u>	<u>57.72</u>	61.56	17.96	<u>24.41</u>	40.47	45.26	30.46	40.50

Table 2: Main results on R2MED. All methods rerank the top-100 passages released by ReasonRank. We report NDCG@10. Among rerankers, best and second-best scores are boldfaced and underlined, respectively.

trains on all non-validation filtered instances under the base prompt, serving as a full-data supervised reference. All RL-based variants share the same cold-start split and adaptive-RL split. **SFT+GRPO** uses base-prompt RL only. **w/o Guided Exploration** removes both reusable guidance and teacher-guided repair, so base-prompt failures fall back directly to teacher trajectories. **w/o Teacher Repair** keeps guided exploration but routes guidance failures directly to teacher fallback. **w/o Bridge** retains assistance routing and trains on successful guided or repaired student trajectories under their respective assisted prompts, but omits the shaped bridge update that transfers them to the base-prompt policy. **w/o Shaping** retains the bridge update under the base-prompt target condition but removes the bounded shaping function, yielding the naive off-policy transfer control. The full SARD model uses the complete adaptive assistance schedule.

The component ablations first support the benefit of the progressive assistance schedule. Removing teacher-guided repair lowers performance to 36.40

Method	BRIGHT Avg.	R2MED Avg.
SFT	35.32	38.07
SFT+GRPO	35.61	39.45
w/o Guided Exploration	35.87	39.34
w/o Teacher Repair	36.40	39.97
w/o Bridge	35.92	40.26
w/o Shaping	35.48	39.57
SARD	36.78	40.50

Table 3: Ablation study on BRIGHT and R2MED.

on BRIGHT and 39.97 on R2MED. Removing both guided exploration and teacher-guided repair, such that base-prompt failures directly use teacher fallback, yields 35.87 on BRIGHT and 39.34 on R2MED, compared with 36.78 and 40.50, respectively, for the full schedule. These results support progressively escalating assistance rather than relying only on base-prompt RL or direct teacher fallback.

The bridge variants further examine whether suc-



Figure 3: Training-time assistance routing of SARD on the adaptive-RL training split. **(a)** Base-prompt success rate by domain, measuring how often instances are resolved through the student’s own rollouts. **(b)** Resolution path after base failure, where each assisted category denotes the first subsequent path that resolves a base-failure instance: guided exploration, teacher-guided repair, or teacher fallback.

successful assisted trajectories are effectively transferred back to the base-prompt policy. Removing the bridge decreases performance by 0.86 points on BRIGHT and 0.24 points on R2MED, while removing bounded shaping decreases performance by 1.30 and 0.93 points, respectively. These results show that the full gains require both transferring assisted successes to the base-prompt policy and bounding the resulting off-policy update.

To examine whether the improvement can be explained by additional unguided sampling alone, we increase the rollout group of the unguided SFT+GRPO baseline from $M = 8$ to $M = 16$. This improves BRIGHT NDCG@10 from 35.61 to 36.35, yet remains below SARD (36.78). The result indicates that additional unguided exploration is beneficial but does not fully account for the gains from selective assistance and bridge-based transfer.

4.4 Training-Time Assistance Routing (RQ3)

Figure 3 analyzes how SARD allocates training-time assistance on the adaptive-RL training split. Overall, base-prompt exploration resolves 6,459 out of 7,664 instances, corresponding to a success rate of 84.3%. The remaining 1,205 instances require additional assistance. Among these assisted cases, *Guided Exploration* resolves the largest share (60.5%), while the rest are split between teacher fallback (21.1%) and *Teacher-Guided Repair* (18.4%). This pattern suggests that SARD primarily relies on the student’s own exploration and invokes teacher-derived support selectively when base-prompt exploration is insufficient.

The domain-level trends further show that assistance is not applied uniformly. Base-prompt

success varies across domains, and after base failure, *Guided Exploration* is the dominant resolution path in most domains. Thus, teacher-derived assistance serves as targeted training-time scaffolding rather than a default input to every example, while the final deployed reranker still uses only the base prompt at inference.

Held-out difficulty analysis. To examine where the final performance difference is concentrated, we partition 500 held-out instances according to the earliest stage at which a shared cold-start checkpoint succeeds, and compare the final SARD and matched SFT+GRPO models under the same base-prompt inference condition. On the 424 base-success instances (84.8%), SARD improves NDCG@10 by 0.0038. Among the 76 instances requiring assistance under this cold-start diagnostic (15.2%), SARD achieves an NDCG@10 of 0.6504, compared with 0.6053 for SFT+GRPO, an improvement of 0.0450. This diagnostic indicates that SARD’s larger gains are concentrated on instances for which base-prompt exploration initially requires assistance.

4.5 Inference-Time Interface (RQ4)

SARD uses guidance retrieval and teacher-guided repair only during training. To examine whether explicit guidance remains useful after training, we compare the trained student with and without guidance on 500 held-out validation instances from the source training data. Each setting uses 8 rollouts with temperature 0.7. Guidance-augmented inference improves mean NDCG@10 from 0.8676 to 0.9142, but the median per-instance gain is only

0.0033, indicating that the benefit is concentrated on a subset of instances. It also increases mean latency from 6.747s to 7.190s, and 36.6% of instances incur higher latency without NDCG@10 improvement. These results suggest that explicit test-time guidance can help some cases, but its benefit is heterogeneous and comes with additional overhead. We therefore use the base-prompt student as the default inference interface, without guidance retrieval, teacher diagnosis, or teacher-generated trajectories at test time. Additional per-instance diagnostics are provided in Appendix D.4.

5 Conclusion

We present SARD, a Self-Adaptive Reinforced Distillation framework that, through a tri-mode reinforcement learning schedule with dynamic training allocation, reformulates reasoning distillation for reasoning-intensive reranking. Experiments on reasoning-intensive reranking benchmarks show that SARD attains gains across multiple datasets, with particularly clear improvements on difficult instances and under-represented reasoning patterns, beyond the trace-imitation limitation of the conventional SFT-then-RL pipeline. Future work may extend SARD to broader reasoning tasks such as multimodal reranking and open-domain question answering, and explore teacher-free self-supervised failure recovery mechanisms that further reduce the computational cost and data dependence of distilling reasoning capabilities from large models.

6 Limitations

While SARD demonstrates clear empirical gains, several aspects of its scope and design are worth noting. i) SARD relies on a strong teacher LRM during training to produce reasoning traces, construct the guidance library, and perform repair, which introduces an upfront training data processing cost even though no teacher access is needed at inference. ii) The effectiveness of the Guided Exploration mode is currently bounded by the diversity and quality of the offline guidance library, which is mined from initial teacher traces.

Acknowledgments

This work is supported by the Guangzhou-HKUST(GZ) Joint Funding Program (No. 2024A03J0630) and NSFC (No. 62402413).

References

- Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos Garea, Matthieu Geist, and Olivier Bachem. 2024. On-policy distillation of language models: Learning from self-generated mistakes. In *The Twelfth International Conference on Learning Representations*.
- Shengnan An, Zexiong Ma, Zeqi Lin, Nanning Zheng, Jian-Guang Lou, and Weizhu Chen. 2023. Learning from mistakes makes llm better reasoner. *arXiv preprint arXiv:2310.20689*.
- Yuzheng Cai, Siqi Cai, Yuchen Shi, Zihan Xu, Lichao Chen, Yulei Qin, Xiaoyu Tan, Gang Li, Zongyi Li, Haojia Lin, Yong Mao, Ke Li, and Xing Sun. 2025. Training-free group relative policy optimization. *arXiv preprint arXiv:2510.08191*.
- DeepSeek-AI. 2025a. [Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning](#). *CoRR*, abs/2501.12948.
- DeepSeek-AI. 2025b. [Deepseek-v3.2: Pushing the frontier of open large language models](#). *CoRR*, abs/2512.02556.
- Lasse Espeholt, Hubert Soyer, Rémi Munos, Karen Simonyan, Volodymyr Mnih, Tom Ward, Yotam Doron, Vlad Firoiu, Tim Harley, Iain Dunning, Shane Legg, and Koray Kavukcuoglu. 2018. [IMPALA: scalable distributed deep-rl with importance weighted actor-learner architectures](#). In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pages 1406–1415. PMLR.
- Revanth Gangi Reddy, JaeHyeok Doo, Yifei Xu, Md Arafat Sultan, Deevya Swain, Avirup Sil, and Heng Ji. 2024. [FIRST: Faster improved listwise reranking with single token decoding](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8642–8652, Miami, Florida, USA. Association for Computational Linguistics.
- Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.
- Cheng-Yu Hsieh, Chun-Liang Li, Chih-kuan Yeh, Hootan Nakhost, Yasuhisa Fujii, Alex Ratner, Ranjay Krishna, Chen-Yu Lee, and Tomas Pfister. 2023. [Distilling step-by-step! outperforming larger language models with less training data and smaller model sizes](#). In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8003–8017, Toronto, Canada. Association for Computational Linguistics.
- Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane Dwivedi-Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. 2023. [Atlas: Few-shot learning with retrieval](#)

- augmented language models. *Journal of Machine Learning Research*, 24(251):1–43.
- Wenyuan Jiang, Wenwei Wu, Le Zhang, Zixuan Yuan, Jian Xiang, Jingbo Zhou, and Hui Xiong. 2024. [Killing two birds with one stone: Cross-modal reinforced prompting for graph and language tasks](#). In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, KDD 2024, Barcelona, Spain, August 25-29, 2024*, pages 1301–1312. ACM.
- Yuetai Li, Xiang Yue, Zhangchen Xu, Fengqing Jiang, Luyao Niu, Bill Yuchen Lin, Bhaskar Ramasubramanian, and Radha Poovendran. 2025. [Small models struggle to learn from strong reasoners](#). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 25366–25394, Vienna, Austria. Association for Computational Linguistics.
- Wenhan Liu, Xinyu Ma, Weiwei Sun, Yutao Zhu, Yuchen Li, Dawei Yin, and Zhicheng Dou. 2025. Reasonrank: Empowering passage ranking with strong reasoning ability. *arXiv preprint arXiv:2508.07050*.
- Xuan Lu, Haohang Huang, Rui Meng, Yaohui Jin, Wenjun Zeng, and Xiaoyu Shen. 2025. Rethinking reasoning in document ranking: Why chain-of-thought falls short. *arXiv preprint arXiv:2510.08985*.
- Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. 2022. [Training language models to follow instructions with human feedback](#). In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.
- Meishu Peng, Ziyue Zhang, Yi Zhang, Pengyang Wang, Zixuan Yuan, and Denghui Zhang. 2026. [Accelerating LLM fine-tuning via embedding knowledge transfer](#). In *Findings of the Association for Computational Linguistics, ACL 2026, San Diego, California, United States, July 2-7, 2026*, pages 90–107. Association for Computational Linguistics.
- Ronak Pradeep, Sahel Sharifymoghaddam, and Jimmy Lin. 2023. Rankzephyr: Effective and robust zero-shot listwise reranking is a breeze! *arXiv preprint arXiv:2312.02724*.
- Zhen Qin, Rolf Jagerman, Kai Hui, Honglei Zhuang, Junru Wu, Le Yan, Jiaming Shen, Tianqi Liu, Jialu Liu, Donald Metzler, Xuanhui Wang, and Michael Bendersky. 2024. [Large language models are effective text rankers with pairwise ranking prompting](#). In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 1504–1518, Mexico City, Mexico. Association for Computational Linguistics.
- John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.
- Rulin Shao, Rui Qiao, Varsha Kishore, Niklas Muennighoff, Xi Victoria Lin, Daniela Rus, Bryan Kian Hsiang Low, Sewon Min, Wen tau Yih, Pang Wei Koh, and Luke Zettlemoyer. 2025. ReasonIR: Training retrievers for reasoning tasks. In *Second Conference on Language Modeling*.
- Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.
- Idan Shenfeld, Mehul Damani, Jonas Hübötter, and Pulkit Agrawal. 2026. Self-distillation enables continual learning. *arXiv preprint arXiv:2601.19897*.
- Tingyu Song, Yilun Zhao, Siyue Zhang, Chen Zhao, and Arman Cohan. 2025. [LimRank: Less is more for reasoning-intensive information reranking](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 20625–20639, Suzhou, China. Association for Computational Linguistics.
- Hongjin SU, Howard Yen, Mengzhou Xia, Weijia Shi, Niklas Muennighoff, Han yu Wang, Liu Haisu, Quan Shi, Zachary S Siegel, Michael Tang, Ruoxi Sun, Jinsung Yoon, Sercan O Arik, Danqi Chen, and Tao Yu. 2025. BRIGHT: A realistic and challenging benchmark for reasoning-intensive retrieval. In *The Thirteenth International Conference on Learning Representations*.
- Weiwei Sun, Lingyong Yan, Xinyu Ma, Shuaiqiang Wang, Pengjie Ren, Zhumin Chen, Dawei Yin, and Zhaochun Ren. 2023. [Is ChatGPT good at search? investigating large language models as re-ranking agents](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14918–14937, Singapore. Association for Computational Linguistics.
- Zengkui Sun, Yijin Liu, Fandong Meng, Yufeng Chen, Jinan Xu, and Jie Zhou. 2025. Warmup-distill: Bridge the distribution mismatch between teacher and student before knowledge distillation. *arXiv preprint arXiv:2502.11766*.
- Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 2021. BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021, December 2021, virtual*.
- Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal

- Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. [Llama: Open and efficient foundation language models](#). *CoRR*, abs/2302.13971.
- Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024a. [Improving text embeddings with large language models](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11897–11916, Bangkok, Thailand. Association for Computational Linguistics.
- Yuhao Wang, Ruiyang Ren, Junyi Li, Wayne Xin Zhao, Jing Liu, and Ji-Rong Wen. 2024b. [REAR: A relevance-aware retrieval-augmented framework for open-domain question answering](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5613–5626, Miami, Florida, USA. Association for Computational Linguistics.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems, NIPS '22*, Red Hook, NY, USA. Curran Associates Inc.
- Orion Weller, Kathryn Ricci, Eugene Yang, Andrew Yates, Dawn Lawrie, and Benjamin Van Durme. 2025. Rank1: Test-time compute for reranking in information retrieval. *arXiv preprint arXiv:2502.18418*.
- Jianhao Yan, Yafu Li, Zican Hu, Zhi Wang, Ganqu Cui, Xiaoye Qu, Yu Cheng, and Yue Zhang. 2025. Learning to reason under off-policy guidance. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.
- An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, and 22 others. 2024. [Qwen2.5 technical report](#). *CoRR*, abs/2412.15115.
- Eugene Yang, Andrew Yates, Kathryn Ricci, Orion Weller, Vivek Chari, Benjamin Van Durme, and Dawn Lawrie. 2025. Rank-k: Test-time reasoning for listwise reranking. *arXiv preprint arXiv:2505.14432*.
- Zixuan Yuan, Yada Zhu, Wei Zhang, and Hui Xiong. 2023. [Earnings call analysis using a sparse attention based encoder and multi-source counterfactual augmentation](#). In *4th ACM International Conference on AI in Finance, ICAIF 2023, Brooklyn, NY, USA, November 27-29, 2023*, pages 331–339. ACM.
- Le Zhang, Bo Wang, Xipeng Qiu, Siva Reddy, and Aishwarya Agrawal. 2025a. [REARANK: Reasoning re-ranking agent via reinforcement learning](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 2458–2471, Suzhou, China. Association for Computational Linguistics.
- Xiangxu Zhang, Lei Li, Xiao Zhou, and Zheng Liu. 2025b. R2med: A benchmark for reasoning-driven medical retrieval. *arXiv preprint arXiv:2505.14558*.
- Xiaoying Zhang, Yipeng Zhang, Hao Sun, Kaituo Feng, Chaochao Lu, Chao Yang, and Helen Meng. 2025c. Critique-grpo: Advancing llm reasoning with natural language and numerical feedback. *arXiv preprint arXiv:2506.03106*.
- Xichen Zhang, Sitong Wu, Yinghao Zhu, Haoru Tan, Shaozuo Yu, Ziyi He, and Jiaya Jia. 2026. Scaff-GRPO: Scaffolded group relative policy optimization for enhancing LLM reasoning. In *The Fourteenth International Conference on Learning Representations*.
- Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. 2025d. [Qwen3 embedding: Advancing text embedding and reranking through foundation models](#). *CoRR*, abs/2506.05176.
- Siyan Zhao, Zhihui Xie, Mengchen Liu, Jing Huang, Guan Pang, Feiyu Chen, and Aditya Grover. 2026. Self-distilled reasoner: On-policy self-distillation for large language models. *arXiv preprint arXiv:2601.18734*.
- Honglei Zhuang, Zhen Qin, Rolf Jagerman, Kai Hui, Ji Ma, Jing Lu, Jianmo Ni, Xuanhui Wang, and Michael Bendersky. 2023. [RankT5: Fine-tuning T5 for text ranking with ranking losses](#). In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2023, Taipei, Taiwan, July 23-27, 2023*, pages 2308–2313. ACM.
- Shengyao Zhuang, Xueguang Ma, Bevan Koopman, Jimmy Lin, and Guido Zuccon. 2025. Rank-r1: Enhancing reasoning in llm-based document rerankers via reinforcement learning. *arXiv preprint arXiv:2503.06034*.
- Shengyao Zhuang, Honglei Zhuang, Bevan Koopman, and Guido Zuccon. 2024. [A setwise approach for effective and highly efficient zero-shot ranking with large language models](#). In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2024, Washington DC, USA, July 14-18, 2024*, pages 38–47. ACM.

A Guidance Library Construction

This appendix describes how we construct the reusable in-context guidance library used in *Guided Exploration* and *Teacher-Guided Repair*. The library is a training-time scaffold: it abstracts successful teacher reasoning into compact guidance cards that can redirect failed student exploration, while leaving the inference-time reranker unchanged.

Teacher rollout collection. We start from the reasoning-intensive reranking instances from ReasonRank. For each instance, we query the teacher with a low-temperature decoding setting and allow a small rollout budget of at most five attempts. We compute NDCG@10 using the original ReasonRank relevance annotations and retain an instance when at least one valid teacher rollout within this budget reaches NDCG@10 above 0.6. Among the valid high-scoring rollouts, we keep a single teacher trajectory for downstream construction. This yields 10,265 strategy-ready instances, each associated with one ranked teacher output. The retained teacher trajectory is used both as a filtered reasoning trace for observation extraction and as the fallback signal if all student-side exploration modes fail during adaptive RL.

Observation extraction. Before clustering, we rewrite each retained teacher trajectory into an abstract observation that summarizes the reusable ranking behavior behind the trace. This step reduces dependence on instance-level wording, passage snippets, and domain-specific vocabulary. Each observation focuses on the latent information need, the evidence type being prioritized, the comparison rule used to distinguish candidates, and common distractors to down-weight. The need to separate task-relevant signals from noisy context also arises in long-document applications such as earnings-call analysis (Yuan et al., 2023). Because our reranking setting allows partial rankings, the observation also records what evidence should be retained and what evidence should be omitted.

Observation clustering. We embed the extracted observations using Qwen3-Embedding-8B (Zhang et al., 2025d) and cluster the embeddings by ranking behavior rather than topic identity. The embedded text contains only transferable strategy content and excludes raw queries, raw passages, teacher traces, final answer lists, and instance metadata. We use average-linkage agglomerative clustering

with cosine distance and set the distance threshold to 0.42. This setting is fixed before adaptive RL based on cluster coherence and qualitative inspection of whether instances inside a cluster can plausibly share one guidance card. In the final construction, 10,265 observation-ready instances are grouped into 772 clusters.

Cluster-level guidance induction. For each observation cluster, we construct a compact cluster packet and ask an LLM to synthesize its shared ranking behavior into one guidance card. Each card describes the objective of the strategy, when it applies, what evidence should be promoted or demoted, how retained candidates should be ranked, and what kinds of candidates should be omitted. The resulting guidance is intentionally more abstract than a per-instance observation: it should be specific enough to redirect failed exploration, but general enough to be reused across instances with similar ranking behavior.

Cluster-induced guidance assignment. Guidance assignment is a deterministic byproduct of the clustering process. After the observations are clustered and each cluster is summarized into a guidance card, the cluster identity of an instance directly determines its assigned guidance. Specifically, each training instance inherits the guidance card associated with the cluster that contains its teacher-derived observation. In this way, the same clustering step both discovers reusable ranking behaviors and provides the assignment key for reusing them.

This design avoids introducing a separate LLM router over the full guidance library. The assigned guidance is injected only when the student fails under the base prompt and enters *Guided Exploration*; at inference time, the deployed reranker receives only the base prompt.

Guidance quality and coverage. Among the 1,205 adaptive-RL training instances for which base exploration fails, the initially assigned guidance directly resolves 729 instances. This conditional resolution rate characterizes the effectiveness of the current guidance library and assignment procedure on escalated instances; it is not an isolated measure of the library’s intrinsic coverage. Guidance quality and coverage may affect how often instances require subsequent teacher-guided repair or teacher fallback. This sensitivity is consistent with reinforced prompting, where indiscrim-

inately incorporating external prompt knowledge can introduce noise and negative transfer (Jiang et al., 2024). The teacher-derived clustering and cluster-inheritance rule used in this work constitutes one concrete implementation for constructing and assigning guidance, rather than an optimal or uniquely required design.

Neighbor guidance candidates for repair. To support *Teacher-Guided Repair*, we also build a local neighbor index over the synthesized guidance cards. We embed each final guidance card using Qwen3-Embedding-8B and retrieve its nearest neighboring cards by cosine similarity. During repair, the teacher receives the failed rollout, the current guidance, and a small local candidate set consisting of the assigned guidance and its nearest neighbors. The teacher may revise the current guidance, re-select a neighboring card, or append an instance-local repair when the observed failure mode is not covered by the existing candidates. These neighbor cards are used only as training-time repair resources and are never queried during inference.

More broadly, prior work in other settings has leveraged auxiliary information through multi-source counterfactual augmentation (Yuan et al., 2023), cross-modal reinforced prompting (Jiang et al., 2024), and embedding knowledge transfer for efficient LLM fine-tuning (Peng et al., 2026).

Examples of mined guidance. Below we show representative guidance prompts from the mined library. For compactness, we present shortened guidance cards rather than the full prompt templates used during training.

Example 1: Causal Mechanism Explanation Prioritization.

Guidance: Causal Mechanism Explanation Prioritization

Use when:
The query asks why or how an outcome occurs, or requires understanding an operational process or causal relationship.

Promote:

- Passages that explicitly connect causes, mechanisms, or processes to concrete effects.
- Passages that explain intermediate steps in a cause-to-outcome chain.
- Passages that describe functional relationships or constraints that determine the observed behavior.

Demote:

- Background passages that only name related concepts.

- Isolated component descriptions without mechanism.
 - Tangential historical or contextual information.
- Core ranking behavior:
Prefer evidence that explains the mechanism behind the query over evidence that is merely topically related.

Example 2: Structural Constraint and Mechanism Proximity Ranking.

Guidance: Structural Constraint and Mechanism Proximity Ranking

Use when:
The query asks why a biological, chemical, or physical process does or does not occur under a specific constraint.

Promote:

- Passages discussing structural, chemical, or physical properties relevant to the query.
- Evidence that links these properties to functional capability or limitation.
- Passages explaining why a standard process fails under the queried condition.

Demote:

- General descriptions of the standard process without explaining the exception.
- Procedural details that do not address the underlying constraint.
- Citations, incomplete references, or unrelated background.

Core ranking behavior:
Prefer passages that explain the constraint causing success or failure, rather than passages that only describe the normal case.

Example 3: Multi-Criteria Evidence Matching.

Guidance: Multi-Criteria Evidence Matching

Use when:
The query requires evidence satisfying multiple conditions, requirement categories, or competing constraints.

Promote:

- Passages that directly satisfy several required criteria.
- Passages that provide concrete evidence across different requirement types.
- Passages that explicitly resolve conflicts between competing factors.

Demote:

- Vague topical matches.
- Passages that satisfy only one minor criterion.
- Background information without direct evidence for the required conditions.

Core ranking behavior:

Rank candidates by completeness and directness of evidence across all major requirements, not by topical overlap alone.

Example 4: Computational Method Prioritization.

Guidance: Computational Method Prioritization

Use when:

The query requires a mathematical, algorithmic, or procedural solution method.

Promote:

- Passages that provide concrete computational steps.
- Passages that explain how to transform the problem into a solvable form.
- Passages that connect definitions or properties to an executable calculation or algorithm.

Demote:

- Definitions without solution procedure.
- Abstract properties that do not help solve the instance.
- Related but different algorithms or formulas.

Core ranking behavior:

Prefer evidence that gives a usable method or reasoning procedure over evidence that only states background concepts.

B Additional Training and Implementation Details

This appendix provides implementation details for the training pipeline. Appendix A describes guidance-library construction and assignment, while Appendix C lists the base, guided, and repair prompts.

B.1 Data Splits and Teacher Traces

We use the reasoning-intensive reranking instances from ReasonRank as the source training data. The training labels released with ReasonRank are used only for consistency filtering of teacher rollouts, not as the reward labels for subsequent RL training. As described in Appendix A, we query the teacher under a low-temperature setting with a small rollout budget and retain instances whose valid teacher rollout achieves sufficiently high NDCG@10 under the original ReasonRank relevance annotations. After this filtering and guidance-library construction process, we obtain 10,265 strategy-ready instances, each associated with a retained teacher trajectory, a ranked output, and an assigned cluster-level guidance card.

We split these 10,265 instances into three parts: 2,101 instances for supervised warm-up, 7,664 in-

stances for adaptive RL, and 500 held-out instances for validation diagnostics, corresponding to approximately 20%, 75%, and 5% of the filtered data, respectively. The supervised warm-up subset is used to initialize the student before rollout-based optimization. The 7,664 adaptive-RL instances are used for the SARD training procedure and for the training-time routing analysis in Section 4.4. The 500 held-out instances are used only for validation diagnostics and the inference-interface analysis in Section 4.5; they are not used for model updates.

The supervised-only SFT baseline is trained on the non-validation portion of the filtered data. For the RL-based variants, the same supervised warm-up split and adaptive-RL split are used, so differences among variants reflect the training schedule rather than differences in the underlying data.

Models and compute. The student is initialized from Qwen2.5-7B-Instruct. The teacher is DeepSeek-V3.2 and is used during training for trace generation, guidance construction, and teacher-guided repair. Training is conducted on 8 NVIDIA A800 80GB GPUs. At inference time, the deployed model uses only the base prompt. Our code and implementation are available at <https://github.com/trail-lab-hkust/SARD>.

B.2 Reward Construction

For RL, we construct training rewards from teacher-derived pseudo-relevance labels. Given an instance x , the teacher receives the candidate document list and is allowed to output an ordered subset of documents that it judges to be relevant or useful for the query, rather than ranking all candidates. Let $S_T(x)$ be this ordered teacher-retained subset. We assign graded pseudo-relevance labels by linearly decreasing weight along the retained order and assigning zero weight to non-retained documents. For example, if the teacher retains five documents, their unnormalized weights are 5, 4, 3, 2, 1, and all remaining candidates receive 0. The weights are normalized before computing NDCG.

For a student rollout y , let π_y be its parsed ranking. We compute

$$R_{\text{rank}}(x, y) = \text{NDCG@10}(\pi_y; S_T(x)) + 0.2 \cdot \text{Recall@10}(\pi_y; S_T(x)).$$

Rollouts that cannot be parsed into a valid ranked list over the candidate identifiers receive zero ranking reward. This reward is used consistently for

free exploration, guided exploration, repaired rollouts, and bridge groups.

B.3 Routing Criterion and Training Modes

During adaptive RL, an instance is considered resolved by a mode when at least one sampled rollout reaches a global training success threshold. Let $C(x, y)$ denote NDCG@10 computed from the teacher-derived pseudo-relevance labels described above. In our implementation, the routing criterion is $C(x, y) \geq \delta$, where $\delta = 0.9$ for all training instances. Instances that satisfy the criterion under the base prompt enter *Free Exploration*. Instances that fall short under the base prompt are retried with the assigned guidance in *Guided Exploration*. If guidance remains insufficient, the instance enters *Teacher-Guided Repair*. If all bounded repair attempts fall short, the teacher trajectory is used as a final fallback signal.

This criterion is used only for training-time routing. Benchmark evaluation follows the ReasonRank protocol and reports standard NDCG@10.

B.4 Training Procedure

For the supervised-only baseline, SFT is trained by supervised trace distillation on the non-validation portion of the filtered data for 5 epochs and does not enter rollout-based optimization. For all RL-based variants, we first apply supervised warm-up on the 2,101 warm-up instances for 3 epochs. This warm-up stage teaches the listwise output format, initializes a basic reasoning policy, and reduces invalid early rollouts before reward-driven training begins.

After warm-up, all RL-based variants continue training on the same 7,664 adaptive-RL instances. SFT+GRPO applies standard GRPO under the base prompt only. The full SARD routes instances according to rollout outcomes: each instance is first attempted through *Free Exploration*; if it fails to reach the routing threshold, it enters *Guided Exploration* with its assigned cluster-level guidance; if the assigned guidance remains insufficient, it enters *Teacher-Guided Repair*; and only instances that remain unresolved after bounded repair attempts fall back to the retained teacher trajectory.

We implement supervised fine-tuning with LLaMA-Factory 0.9.4.dev0 and adaptive RL with verl 0.4.0; vLLM 0.8.5 is used for rollout generation and inference. Adaptive RL uses learning rate 1×10^{-6} , batch size 16, and KL loss coefficient 0.001. Each rollout group contains $M = 8$ sam-

ples generated with temperature 0.7. The repair budget is $T = 3$, and the bridge shaping parameter is $\gamma = 0.1$. Unless otherwise stated, all RL-based results use the checkpoint after 340 adaptive-RL steps.

For the ablation study in Section 4.3, all RL-based variants start from the same supervised warm-up checkpoint and use the same base model, teacher-trace collection pipeline, adaptive-RL split, rollout group size, reward function, and routing threshold. They differ only in which assistance branches are enabled after base-prompt failure. The variant without Guided Exploration removes both reusable guidance and teacher-guided repair, so base-prompt failures are handled directly by teacher fallback. The variant without Teacher Repair keeps Guided Exploration but sends guidance failures directly to teacher fallback. The full SARD enables the complete progression from Free Exploration to Guided Exploration, Teacher-Guided Repair, and teacher fallback.

We additionally evaluate two bridge ablations. The *w/o Bridge* variant retains the same assistance routing and GRPO updates under guidance or repair augmented prompts for guided or repaired student rollouts, but removes the shaped base-prompt bridge update. The *w/o Shaping* variant retains the base-prompt bridge target while removing the bounded shaping function. We also evaluate an SFT+GRPO control on BRIGHT with rollout group size increased from $M = 8$ to $M = 16$, while retaining the remaining training configuration.

Algorithm 1 summarizes the adaptive RL procedure used by the full model. SFT stops after supervised trace distillation, while SFT+GRPO and all SARD variants continue with their corresponding rollout-based training schedules.

B.5 Algorithm

The algorithm of our main method is shown in Algorithm 1.

B.6 Listwise Evaluation Protocol

For BRIGHT and R2MED evaluation, all rerankers operate on the released top-100 candidate passages described in the main experimental setup. Following the ReasonRank protocol, we use listwise sliding-window reranking with 20 passages per window and a stride of 10. The window-level rankings are aggregated into a final ranking, which is evaluated with NDCG@10. For benchmark evalua-

Algorithm 1 SARD: Three-Mode Self-Adaptive Reinforced Distillation

Input: Supervised warm-up set $\mathcal{D}_{\text{SFT}}$; adaptive-RL instances $\{(x_i, g_i, y_i^T)\}_{i=1}^{N_{\text{RL}}}$; guidance library $\mathcal{G}$; initial policy π_θ ; group size $G = 8$; routing score $C(x, y)$; success threshold $\delta = 0.9$; repair budget $T = 3$.

Output: Base-prompt reranker $\pi_\theta(\cdot \mid x)$.

```
1: Cold start. Initialize  $\pi_\theta$  by supervised imitation on  $\mathcal{D}_{\text{SFT}}$ .
2: for each RL iteration do
3:   Set rollout policy  $\pi_{\text{old}} \leftarrow \pi_\theta$ .
4:   for each adaptive-RL instance  $(x_i, g_i, y_i^T)$  do

5:     Mode I: Free Exploration.
6:     Sample  $\mathcal{Y}_i^0 = \{y_{i,j}^0\}_{j=1}^G \sim \pi_{\text{old}}(\cdot \mid x_i)$ .
7:      $s_i^0 \leftarrow \max_{y \in \mathcal{Y}_i^0} C(x_i, y)$ .
8:     if  $s_i^0 \geq \delta$  then
9:       Update  $\pi_\theta$  with GRPO on  $\mathcal{Y}_i^0$  under base prompt  $x_i$ .
10:    continue
11:  end if

12:  Mode II: Guided Exploration.
13:  Sample  $\mathcal{Y}_i^g = \{y_{i,j}^g\}_{j=1}^G \sim \pi_{\text{old}}(\cdot \mid x_i, g_i)$ .
14:   $s_i^g \leftarrow \max_{y \in \mathcal{Y}_i^g} C(x_i, y)$ .
15:  if  $s_i^g \geq \delta$  then
16:     $y_i^* \leftarrow \arg \max_{y \in \mathcal{Y}_i^g} C(x_i, y)$ .
17:    Update  $\pi_\theta$  with GRPO on  $\mathcal{Y}_i^g$  under guidance-augmented prompt  $(x_i, g_i)$ .
18:    Bridge  $y_i^*$  from behavior prompt  $(x_i, g_i)$  to base prompt  $x_i$  using the shaped bridge objective.
19:    continue
20:  end if

21:  Mode III: Teacher-Guided Repair.
22:  Set  $\tilde{g}_i^{(0)} \leftarrow g_i$  and resolved  $\leftarrow$  False.
23:  for  $t = 1, \dots, T$  do
24:     $\tilde{g}_i^{(t)} \leftarrow \text{TeacherRepair}(x_i, \tilde{g}_i^{(t-1)}, \mathcal{Y}_i^g, \mathcal{G})$ .
25:    Sample  $\mathcal{Y}_i^{r,t} = \{y_{i,j}^{r,t}\}_{j=1}^G \sim \pi_{\text{old}}(\cdot \mid x_i, \tilde{g}_i^{(t)})$ .
26:     $s_i^{r,t} \leftarrow \max_{y \in \mathcal{Y}_i^{r,t}} C(x_i, y)$ .
27:    if  $s_i^{r,t} \geq \delta$  then
28:       $y_i^* \leftarrow \arg \max_{y \in \mathcal{Y}_i^{r,t}} C(x_i, y)$ .
29:      Update  $\pi_\theta$  with GRPO on  $\mathcal{Y}_i^{r,t}$  under repaired guidance prompt  $(x_i, \tilde{g}_i^{(t)})$ .
30:      Bridge  $y_i^*$  from behavior prompt  $(x_i, \tilde{g}_i^{(t)})$  to base prompt  $x_i$  using the shaped bridge objective.
31:      resolved  $\leftarrow$  True.
32:      break
33:    end if
34:  end for
35:  if resolved = False then
36:    Use the retained teacher trajectory  $y_i^T$  as fallback supervision.
37:    Bridge or imitate  $y_i^T$  under the base prompt  $x_i$ .
38:  end if
39: end for
40: end for
41: return  $\pi_\theta(\cdot \mid x)$ .
```

Dataset / Resource	License
BRIGHT	CC-BY-4.0
R2MED	MIT License
ReasonRank	MIT License
BEIR	Apache-2.0

Table 4: License information for the datasets and benchmark resources used in this work.

tion, decoding temperature is set to 0. The guidance library, teacher-guided repair, teacher fallback, and bridge objective are used only during training. During benchmark inference, SARD uses the trained student reranker with the base prompt only, without guidance retrieval, teacher diagnosis, or teacher-generated trajectories.

B.7 Datasets, licenses, and responsible use.

Table 4 provides license information for the datasets and benchmark resources used in our experiments. We use all listed resources solely for academic research purposes, in accordance with their designated uses and original access conditions.

C Prompt Template and Qualitative Case Studies

This appendix reports the base, strategy-guided, and repair prompt templates used in SARD, followed by qualitative case studies. The case studies use the original model inputs and evaluated outputs from selected reranking windows; they are meant to illustrate representative ranking behavior alongside the aggregate results in Section 4.

C.1 Prompt Template

All methods are evaluated with the same listwise reranking interface. The system prompt defines the reranking role, the relevance rubric, and the answer format. The user instruction supplies the full query and one local sliding-window block of candidate passages.

System prompt:

```
You are RankLLM, an intelligent assistant specializing in listwise passage reranking.

Your task is to analyze a list of passages provided by the user and rank ONLY the relevant ones based on their relevance to a specific query.

**Core Evaluation Rubrics:**
When evaluating relevance in your internal thought process, adhere to these principles:
1. **Determine Query Type:**
  - **Fact-seeking:** Prioritize direct semantic matches, definitions, or explicit facts.
  - **Reasoning-intensive:** Prioritize passages that provide the logical premise, solution pattern, or underlying mechanism (e.g., inferring relationships, applying logic), even if keywords do not match directly.
2. **Scoring Priority:**
  - **Rank 1 (Solvability):** The document contains the complete solution or answer logic.
  - **Rank 2 (Supportiveness):** The document provides necessary context or partial clues.
  - **Rank 3 (Topicality):** The document discusses the topic but fails to answer the specific question.
  - **Rank 4 (Irrelevance):** The document is unrelated.

**Operational Process:**
1. **Implicit Reasoning:** First, generate a reasoning process enclosed within <thinking> tags. In this block, apply the rubrics above to analyze the query and evaluate the passages. You should compare them based on Solvability and Supportiveness. Finally, decide which passages are relevant and which are irrelevant.
2. **Final Answer:** Generate the final ranked list enclosed within <answer> tags, including ONLY the relevant passages.

**Output Constraints:**
- The output must strictly follow this structure: <thinking> ... </thinking> <answer> ... </answer>
- Do not output any text outside these tags.
- **Answer Format:** Only include relevant passages, listed using identifiers in descending order of relevance. The format of the answer should be [ ] > [ ], e.g., [2] > [1].
- **Recall Bias:** When in doubt about whether a passage is relevant, include it. Only exclude obvious noise/distractors.
```

User instruction template:

```
<user_query>
{full user query}
</user_query>

<retrieved_passages>
<doc_id=1>
{passage text for local candidate 1}
</doc_id=1>
<doc_id=2>
{passage text for local candidate 2}
</doc_id=2>
```

```
...
</retrieved_passages>
```

Perform the analysis and ranking. Ensure all XML tags are closed.

Strategy-guided prompt template. When an instance enters *Guided Exploration*, SARD uses the same system prompt as above, but prepends a structured <strategy_guidance> block to the user instruction. The strategy is used as a decision guideline rather than as a passage-level answer.

```
<strategy_guidance>
Use the following ranking strategy as a decision guideline. Apply it to decide which passages are relevant, which passages should be omitted as irrelevant, and how the relevant passages should be ordered.

Core relevance objective:
{objective}

When to apply:
{when_to_apply}

Query signals:
- {query_signal_1}
- {query_signal_2}
...

Evidence to promote:
- {evidence_to_promote_1}
- {evidence_to_promote_2}
...

Evidence to demote or omit:
- {evidence_to_demote_1}
- {evidence_to_demote_2}
...

Ranking procedure:
1. {step_1}
2. {step_2}
...

Tie-breaking rule:
{tie_breaking_rule}

Fallback behavior:
{fallback_behavior}

Avoid when:
- {avoid_condition_1}
- {avoid_condition_2}
...
</strategy_guidance>

<user_query>
{full user query}
</user_query>

<retrieved_passages>
<doc_id=1>
{passage text for local candidate 1}
</doc_id=1>
...
</retrieved_passages>

Perform the analysis and ranking. Ensure all XML tags are closed.
```

Teacher-guided repair prompt template. When strategy-guided rollouts still fail, the repair model is asked to repair the strategy, not to solve the ranking instance. It receives the ranking problem, the current failed strategy, the top candidate strategies returned by the router, and the student's best failed rollout. The output is a new strategy card in JSON format.

```
System:
You are a strategy repair module for a reinforcement learning ranking system. You are not solving the ranking problem. Your job is to inspect the ranking problem, the current strategy that failed, the student's best failed rollout
```

under that strategy, and a library of candidate strategies, then produce exactly one improved strategy for the next student rollout. Do not output a passage ranking. Do not reveal or guess final passage IDs. Return only valid JSON matching the required schema.

User payload:

```
{
  "task": "repair_ranking_strategy",
  "instruction": {
    "goal": "Produce one improved strategy card for the next student rollout.",
    "allowed_actions": [
      "rewrite_current_strategy",
      "select_and_rewrite_from_library",
      "create_new_strategy"
    ],
    "constraints": [
      "Return JSON only.",
      "Do not output a final passage ranking.",
      "Do not include direct passage-id ordering hints such as '[3] should be first'.",
      "The new strategy must be general instructions for the student, not an answer to this specific instance.",
      "The new strategy must preserve the student's required output format."
    ]
  },
  "ranking_problem": {
    "base_user_instruction": "{base user instruction with query and passages}"
  },
  "current_strategy": "{current strategy card or rendered guidance}",
  "candidate_strategy_library": "{top candidate strategies returned by the router}",
  "student_best_rollout_under_current_strategy": {
    "response": "{student best failed strategy-guided rollout}"
  },
  "output_schema": {
    "diagnosis_type": "strategy_description_bad | wrong_strategy_selected | strategy_missing",
    "action": "rewrite_current_strategy | select_and_rewrite_from_library | create_new_strategy",
    "selected_strategy_id": "string or null",
    "failure_analysis": "short diagnostic explanation",
    "new_strategy": "{strategy card}",
    "confidence": 0.0
  }
}
```

C.2 Case Study Protocol

The NDCG scores reported below are computed after reranking the full top-100 first-stage candidate set with sliding-window listwise reranking. For each qualitative example, we keep the original query, retrieved-passage window, and evaluated `<answer>` output from the corresponding run. The displayed windows are method-specific because different rerankers may visit different local windows during sliding-window reranking; local document identifiers should therefore be interpreted only within the window in which they appear. The model-generated intermediate analysis is omitted, while the final `<answer>` field is preserved as the ranking output consumed by evaluation. Long lines are wrapped only for readability.

Case selection. We include two ablation-style cases and two training-time rescue cases. The ablation cases compare SFT, SFT+GRPO, and SARD on the same top-100 reranking task. The rescue cases show training instances routed from base-prompt exploration to guidance or repair. These examples are selected because the retrieved candidates are topically plausible and the ranking difference is visible from the original input-output traces.

Selected cases. We show (i) data-center waste heat

reuse from `sustainable_living`, where the ranking must separate direct reuse and obstacle evidence from broad cooling context, and (ii) life of a dollar from `economics`, where the ranking must prioritize circular-flow evidence over adjacent banking passages.

C.3 Qualitative Case Studies

C.3.1 Case 1: Data-center waste heat reuse

Dataset / query id. `sustainable_living / 44`.

Qualitative focus. The query asks whether data-center waste heat can be used for nearby heating and why this is not done more often. The key distinction is between passages that directly discuss heat reuse or its obstacles and passages that provide broader data-center cooling context.

Full top-100 NDCG@10. SFT: 0.387; SFT+GRPO: 0.631; SARD: 1.000.

Full query:

Can the heat generated in a data center be used to heat households and offices?

I've just read the article Microsoft sinks a data center off the Scottish coast (The Verge, 6 June 2018)

I understand that the cooling of a data center is a problem. Data centers produce huge amounts of heat, and they have to run at normal temperatures, otherwise the hardware gets damaged. The cooling is much easier when there is a lot of cool water available, so most data centers are built in countries with a cool climate.

But at the same time, the countries with a cool climate consume a lot of energy to warm up water for their heatings.

Why does Microsoft experiment with a data center under sea, instead of using the data center's warm water to heat surrounding buildings? There are similar solutions in other areas, e.g. combined heat and power facilities with district heating schemes.

I've found some articles about data centers involved in district heating, but I'm wondering why this isn't done more often. Are there any obstacles, hidden costs, or other problems (like, keeping the location of the data center a secret)?

SFT. SFT selected input window:

```
<retrieved_passages>
<doc_id=1>
[1] At the Energy Systems Integration Facility (ESIF), heat exchangers in the High-Performance Computing Data Center enable the reuse of energy from computing waste heat. With heat exchangers, heat energy in the energy recovery water (ERW) loop becomes available to heat the facility's process hot water (PHW) loop. Once heated, the PHW loop supplies:
* Active chilled beams to heat the office space
* Air handlers to heat the conference and high bay spaces
* Snow melt loop in the courtyard of the ESIF's main entrance
* District heating loop:
* If additional heat is needed for the building, the PHW loop can draw heat from the campus heating loop
* If surplus heat is available, the PHW loop can provide heat to the campus heating loop. During transition months (April and October), excess heat from the ESIF has been sufficient to provide heat for other buildings, shortening the time boilers are needed to provide campus heat.
</doc_id=1>
<doc_id=2>
[2] I. Adaptive Cooling: Employing AI with cooling systems promises more adaptability than ever. These
```

solutions continually learn and adjust to the specific cooling needs of each data center in real-time — reducing over-cooling and energy waste.

2. Edge Data Centers: We could see even more data centers placed in strategic 'edge' locations in the future. These edge centers ensure low-latency data delivery and introduce innovative cooling possibilities in their unique settings.

3. Sustainable Energy: While it's vital to become more efficient at energy management, it's also possible to use energy from renewable sources. Solar or wind-powered data centers, for instance, are no longer a dream but a reality. Combining renewable energy sources with efficient cooling strategies presents a holistic approach to [sustainability progress](resources/articles/how-digital-reality-tracks-its-sustainability-progress) .

4. These innovations and changes, coupled with advanced cooling techniques, can help make future data centers increasingly energy efficient.

</doc_id=2>

<doc_id=3>

[3] SUSTAINABILITY

Surplus datacenter heat will be repurposed to heat homes in Denmark

View more

</doc_id=3>

<doc_id=4>

[4] * Hot/cold aisle containment: By configuring server racks in alternating rows — with cold air intakes facing one way and hot air output on the other — data centers can create hot and cold 'corridors'. This maximizes efficiency by ensuring cooled air is where it's needed most, while preventing the mixing of hot and cold air.

- * Airflow management: By optimizing airflow through vents and barriers, data centers can minimize areas of stationary air. Proper airflow management lowers cooling costs by reducing the need to compensate for warm spots. It also helps ensure that cooling is constant throughout the facility.
- * Working inside safe limits: Modern IT equipment is actually quite resilient. The American Society of Heating, Refrigerating and Air-Conditioning Engineers (ASHRAE) publish [temperature standards](https://www.ashrae.org/file%20library/technical%20resources/bookstore/supplemental%20files/referencecard_2021thermalguidelines.pdf) that give safe operating conditions for common equipment. In some cases, maintaining a slightly warmer temperature — if safe — can save considerable energy.
- * Preventing downtime: Internet of Things (IoT) devices, machine learning, and automated systems can help data center operators spot problems before they happen. By monitoring systems and installing timely patches and upgrades, facilities can prevent downtime and improve efficiency further.
- * Humidity control: Finally, managing humidity is crucial. Too much can lead to hardware damage, while too little can create damaging static electricity.

</doc_id=4>

<doc_id=5>

[5] * Might not be suitable for all types of data centers

- * Brings potential risks associated with leaks
- * Costly and tricky to implement

</doc_id=5>

<doc_id=6>

[6] At the heart of this challenge lies the implementation of efficient and sustainable cooling solutions. The average modern data center uses around [40 percent](https://www.theodi.org/article/data-centres-cloud-infrastructures-and-the-tangibility-of-internet-power/) of its power for cooling. After all, to provide a secure, high-quality service, processors and other hardware must be kept in optimal conditions. The future of data center cooling techniques must bring methods that remove this excess heat in a more environmentally friendly way.

</doc_id=6>

<doc_id=7>

[7] * Limited to regions with colder climates or cold seasons

- * Variability in external conditions can affect performance
- * May require backup cooling solutions in case of unpredicted changes

</doc_id=7>

<doc_id=8>

[8] * Can use [vast amounts](https://www.washingtonpost.com/climate-environment/2023/04/25/data-centers-drought-water-use/) of the local water supply.

- * Coolant leaks can be environmentally unfriendly if not addressed properly
- * Complex to set up and can have several potential points of failure

</doc_id=8>

<doc_id=9>

[9] views the problem in terms of primary energy saving: he must justify his choices based on the micro-economic context in which they operate. Before deciding to use a cogeneration facility, it is necessary to make an extensive study of energy and thermal needs and their evolution over time. Indeed, the corresponding investments are generally high, and, to amortize them, the facility must operate at its optimum economic point as long as possible when the energy price is justified, and if possible at about 80 – 90% of the maximum engine capacity range leading to the best technical performance. The profitability calculation depends fundamentally on operating conditions and their evolution over time, each facility representing a special case that requires

detailed study. It is particularly important to properly ensure that the statutory criteria authorizing the resale of electricity to utilities will be respected. If security considerations require the use of an independent mechanical power production unit, it is almost certain that its use in cogeneration will be profitable, the overhead being limited to expenses for heat recovery, often well below those corresponding to the engine and alternator.

Moreover, the technical solution choice depends on many factors. For purposes of comparison, Table 8.2 provides approximate values of the different performance indicators for key possible technologies. The first two rows correspond to what is now called micro-cogeneration, based on Stirling engines or small capacity micro

</doc_id=9>

<doc_id=10>

[10] * Performance can decrease in the wrong conditions

- * High dependence on water availability and quality
- * Can require an overhaul of existing data center infrastructure

</doc_id=10>

<doc_id=11>

[11] Incinerators have electric efficiencies of 14–28%. (10) In order to avoid losing the rest of the energy, it can be used for e.g. [district heating](wiki/District_heating "District heating") ([cogeneration](wiki/Cogeneration "Cogeneration")). The total efficiencies of cogeneration incinerators are typically higher than 80% (based on the [lower heating value](wiki/Lower_heating_value "Lower heating value") of the waste).

</doc_id=11>

<doc_id=12>

[12] being used for purposes of either heating or cooling in an absorption machine. An alternator can at times be coupled to the engine instead of the compressor, which Cogeneration and trigeneration 223 allows, according to the tariffs of electricity and refrigeration needs, to modulate the production. The reciprocating engine may finally directly drive a heat pump compressor. The overall efficiency of the system can be very high, given the coefficients of performance of heat pumps. The heat–power ratio CF is quite low, between 0.5 and 1.5. Overall efficiency is generally very good, above 70%. Mechanical efficiency is usually very high, between 30 and 35% for small gas engines, and up to 45% for large diesel and gas engines. Specific equivalent consumption CE is of the order of 1.6 to 2.

8.3.2 GAS TURBINES

In gas turbines, the total residual heat is found in the exhaust. The performance of the cogeneration system is directly related to the recovery of these gases. One solution is to cool exhaust gases in an air–fumes heat exchanger which can heat the air which is then used for many applications. If the turbine is stopped, an auxiliary boiler ensures the supply of heat to meet the plant needs. One generally uses several cascading recovery exchangers on the exhaust, so it can be cooled as much as possible and air is

</doc_id=12>

<doc_id=13>

[13] gases contain a lot of oxygen (16 to 18% O2). It is therefore possible to exploit these gases (usually very clean, especially if the fuel used is natural gas), continuing the combustion processes in furnaces or boilers, or even using them directly as drying fluid. Given these characteristics, it is obvious that possible configurations are very diverse in terms of internal combustion engine cogeneration.

8.3.1 RECIPROCATING ENGINES

The simplest and most common solution is the production of either hot water at a temperature of 100 degrees C, or superheated steam at 110–120 degrees C, as auxiliary to a classical boiler (Figure 8.3.1). Depending on its purity, water can be directly heated in the engine, or must pass through a low temperature heat exchanger. It then recovers the exhaust gas heat in a heat exchanger placed in series. Hundreds of such units of all capacities (a few kW to several MW) are installed worldwide. A second solution is to cool the engine by an air flow which, in series, provides a convective cooling, and then passes through a recovery heat exchanger on oil, the supercharger intercooler if it exists, the classic radiator cooling system, and finally an air/flue gas exhaust. Hot air is then used for drying, its enthalpy being, if necessary, raised by a supplementary burner. The engine can also be used for air conditioning, driving a compressor directly connected on the shaft, to obtain industrial cooling or chilled water, heat recovered

</doc_id=13>

<doc_id=14>

[14] available at different temperatures for various uses. Another solution, widely used today, especially to replace an existing boiler, is to install a heat recovery steam generator (HRSG) at the output of the gas turbine. The problem of optimizing the recovery exchangers is very similar to the one we have discussed in section Figure 8.3.1: Internal combustion engine cogeneration unit Figure 8.3.2: Cogeneration gas turbine, Documentation SECC 224 Energy systems 7.3: the best configuration must indeed both cool at best the GT exhaust and provide heat at the highest temperature level possible as needed. Figures 8.3.2 and 8.3.3 show the two main components of a cogeneration plant for

</doc_id=14>

<doc_id=15>

[15] unusual type given the

level of exhaust gas temperature and composition. It can however be used only if the fuel is desulfurized. Given the cost of light distillates and kerosene, it is generally adapted to natural gas. The afterburner may also be used in combination with a recovery exchanger, and thus allows better control of the thermal power depending on the needs. The gas turbine flow is nearly constant so that its efficiency is the best possible (cf. example in section 8.5.2).

When the gas turbine used is a micro-turbine rated less than 100 kW, it is enough to heat water, either upstream of an existing boiler, or for hot water uses. The recovery exchanger is then simpler and less expensive than a HRSG. Example presented in section 8.5.1 corresponds to this configuration.

Another very efficient scheme is to directly use the exhaust gases as hot fluid in a dryer. As in the case of the reciprocating engine, the turbine itself is placed in the airflow, so that all losses can be recovered, leading to an overall efficiency close to 1. In addition, the output pressure of the exhaust gas is sufficient to avoid any fan. The hot gas temperatures (400 – 500 degrees C) being compatible with many industrial requirements, applications of this method are numerous.

```
</doc_id=15>
<doc_id=16>
[16] # High-Performance Computing Data Center Waste Heat Reuse
</doc_id=16>
</retrieved_passages>
```

SFT evaluated output:

```
<answer>
[5] > [7] > [10] > [8] > [1] > [3] > [9] > [2] > [11] > [12] > [13] > [14] > [15] > [6]
> [4]
</answer>
```

SFT+GRPO. SFT+GRPO selected input window:

```
<retrieved_passages>
<doc_id=1>
[1] At the Energy Systems Integration Facility (ESIF), heat exchangers in the High-Performance Computing Data Center enable the reuse of energy from computing waste heat.
With heat exchangers, heat energy in the energy recovery water (ERW) loop becomes available to heat the facility's process hot water (PHW) loop. Once heated, the PHW loop supplies:
* Active chilled beams to heat the office space
* Air handlers to heat the conference and high bay spaces
* Snow melt loop in the courtyard of the ESIF's main entrance
* District heating loop:
* If additional heat is needed for the building, the PHW loop can draw heat from the campus heating loop
* If surplus heat is available, the PHW loop can provide heat to the campus heating loop. During transition months (April and October), excess heat from the ESIF has been sufficient to provide heat for other buildings, shortening the time boilers are needed to provide campus heat.
</doc_id=1>
<doc_id=2>
[2] 1. Adaptive Cooling: Employing AI with cooling systems promises more adaptability than ever. These solutions continually learn and adjust to the specific cooling needs of each data center in real-time — reducing over-cooling and energy waste.
2. Edge Data Centers: We could see even more data centers placed in strategic 'edge' locations in the future. These edge centers ensure low-latency data delivery and introduce innovative cooling possibilities in their unique settings.
3. Sustainable Energy: While it's vital to become more efficient at energy management, it's also possible to use energy from renewable sources. Solar or wind-powered data centers, for instance, are no longer a dream but a reality. Combining renewable energy sources with efficient cooling strategies presents a holistic approach to [ sustainability progress ](/resources/articles/how-digital-realty-tracks-its-sustainability-progress) .
4. These innovations and changes, coupled with advanced cooling techniques, can help make future data centers increasingly energy efficient.
</doc_id=2>
<doc_id=3>
[3] SUSTAINABILITY
Surplus datacenter heat will be repurposed to heat homes in Denmark
View more
</doc_id=3>
<doc_id=4>
[4] * Hot/cold aisle containment: By configuring server racks in alternating rows — with cold air intakes facing one way and hot air output on the other — data centers can create hot and cold 'corridors'. This maximizes efficiency by ensuring cooled air is where it's needed most, while preventing the mixing of hot and cold air.
* Airflow management: By optimizing airflow through vents and barriers, data centers can minimize areas of
```

stationary air. Proper airflow management lowers cooling costs by reducing the need to compensate for warm spots. It also helps ensure that cooling is constant throughout the facility.

* Working inside safe limits: Modern IT equipment is actually quite resilient. The American Society of Heating, Refrigerating and Air-Conditioning Engineers (ASHRAE) publish [temperature standards](https://www.ashrae.org/file%20library/technical%20resources/bookstore/supplemental%20files/referencecard_2021thermalguidelines.pdf) that give safe operating conditions for common equipment. In some cases, maintaining a slightly warmer temperature — if safe — can save considerable energy.

* Preventing downtime: Internet of Things (IoT) devices, machine learning, and automated systems can help data center operators spot problems before they happen. By monitoring systems and installing timely patches and upgrades, facilities can prevent downtime and improve efficiency further.

* Humidity control: Finally, managing humidity is crucial. Too much can lead to hardware damage, while too little can create damaging static electricity.

```
</doc_id=4>
<doc_id=5>
[5] * Might not be suitable for all types of data centers
* Brings potential risks associated with leaks
* Costly and tricky to implement
</doc_id=5>
<doc_id=6>
[6] At the heart of this challenge lies the implementation of efficient and sustainable cooling solutions. The average modern data center uses around [ 40 percent ](https://www.theodi.org/article/data-centres-cloud-infrastructures-and-the-tangibility-of-internet-power/) of its power for cooling. After all, to provide a secure, high-quality service, processors and other hardware must be kept in optimal conditions. The future of data center cooling techniques must bring methods that remove this excess heat in a more environmentally friendly way.
</doc_id=6>
<doc_id=7>
[7] * Can use [ vast amounts ](https://www.washingtonpost.com/climate-environment/2023/04/25/data-centers-drought-water-use/) of the local water supply.
* Coolant leaks can be environmentally unfriendly if not addressed properly
* Complex to set up and can have several potential points of failure
</doc_id=7>
<doc_id=8>
[8] * Limited to regions with colder climates or cold seasons
* Variability in external conditions can affect performance
* May require backup cooling solutions in case of unpredicted changes
</doc_id=8>
<doc_id=9>
[9] * Performance can decrease in the wrong conditions
* High dependence on water availability and quality
* Can require an overhaul of existing data center infrastructure
</doc_id=9>
<doc_id=10>
[10] The major impediments which this industry faces include initial financing of large scale projects, resource development risks like failure of drilling wells and insufficient productivity, lack of skilled labor to work on the projects, issues related to drilling rigs and market and network issues like proximity to the market and network capital costs. Also, there are some environmental problems in producing geothermal energy like release of hydrogen sulfide and disposal of toxic geothermal fluids while extraction.
</doc_id=10>
<doc_id=11>
[11] Waste combustion is particularly popular in countries such as Japan, Singapore and the Netherlands, where land is a scarce resource. Denmark and Sweden have been leaders by using the energy generated from incineration for more than a century, in localised [ combined heat and power ](/wiki/Combined_heat_and_power "Combined heat and power") facilities supporting [ district heating ](/wiki/District_heating "District heating") schemes. (4) In 2005, waste incineration produced 4.8% of the electricity consumption and 13.7% of the total domestic heat consumption in Denmark. (5) A number of other European countries rely heavily on incineration for handling municipal waste, in particular [ Luxembourg ](/wiki/Luxembourg "Luxembourg" ), the Netherlands, Germany, and France. (2)
</doc_id=11>
<doc_id=12>
[12] Incinerators have electric efficiencies of 14–28%. (10) In order to avoid losing the rest of the energy, it can be used for e.g. [ district heating ](/wiki/District_heating "District heating") ( [ cogeneration ](/wiki/Cogeneration "Cogeneration") ). The total efficiencies of cogeneration incinerators are typically higher than 80% (based on the [ lower heating value ](/wiki/Lower_heating_value "Lower heating value") of the waste).
</doc_id=12>
<doc_id=13>
[13] views the problem in terms of primary energy saving: he must justify his choices based on the micro-economic context in which they operate. Before deciding to use a cogeneration facility, it is necessary to make an extensive study of energy and thermal needs and their evolution over time. Indeed, the corresponding investments are generally high, and, to amortize them, the facility must operate at its optimum economic point as long as possible when the energy price is justified, and if possible at about 80 – 90% of the maximum engine capacity , range leading to the best technical performance. The profitability calculation depends fundamentally on operating conditions and
```

their evolution over time, each facility representing a special case that requires detailed study. It is particularly important to properly ensure that the statutory criteria authorizing the resale of electricity to utilities will be respected. If security considerations require the use of an independent mechanical power production unit, it is almost certain that its use in cogeneration will be profitable, the overhead being limited to expenses for heat recovery, often well below those corresponding to the engine and alternator. Moreover, the technical solution choice depends on many factors. For purposes of comparison, Table 8.2 provides approximate values of the different performance indicators for key possible technologies. The first two rows correspond to what is now called micro-cogeneration, based on Stirling engines or small capacity micro

</doc_id=13>
<doc_id=14>
[14] available at different temperatures for various uses.
Another solution, widely used today, especially to replace an existing boiler, is to install a heat recovery steam generator (HRSG) at the output of the gas turbine. The problem of optimizing the recovery exchangers is very similar to the one we have discussed in section Figure 8.3.1: Internal combustion engine cogeneration unit
Figure 8.3.2: Cogeneration gas turbine, Documentation SECC
224 Energy systems
7.3: the best configuration must indeed both cool at best the GT exhaust and provide heat at the highest temperature level possible as needed.
Figures 8.3.2 and 8.3.3 show the two main components of a cogeneration plant for

</doc_id=14>
<doc_id=15>
[15] being used for purposes of either heating or cooling in an absorption machine. An alternator can at times be coupled to the engine instead of the compressor, which Cogeneration and trigeneration 223 allows, according to the tariffs of electricity and refrigeration needs, to modulate the production. The reciprocating engine may finally directly drive a heat pump compressor. The overall efficiency of the system can be very high, given the coefficients of performance of heat pumps. The heat-power ratio CF is quite low, between 0.5 and 1.5. Overall efficiency is generally very good, above 70%. Mechanical efficiency is usually very high, between 30 and 35% for small gas engines, and up to 45% for large diesel and gas engines. Specific equivalent consumption CE is of the order of 1.6 to 2.

8.3.2 GAS TURBINES
In gas turbines, the total residual heat is found in the exhaust. The performance of the cogeneration system is directly related to the recovery of these gases. One solution is to cool exhaust gases in an air-fumes heat exchanger which can heat the air which is then used for many applications. If the turbine is stopped, an auxiliary boiler ensures the supply of heat to meet the plant needs. One generally uses several cascading recovery exchangers on the exhaust, so it can be cooled as much as possible and air is

</doc_id=15>
<doc_id=16>
[16] unusual type given the level of exhaust gas temperature and composition. It can however be used only if the fuel is desulfurized. Given the cost of light distillates and kerosene, it is generally adapted to natural gas. The afterburner may also be used in combination with a recovery exchanger, and thus allows better control of the thermal power depending on the needs. The gas turbine flow is nearly constant so that its efficiency is the best possible (cf. example in section 8.5.2). When the gas turbine used is a micro-turbine rated less than 100 kW, it is enough to heat water, either upstream of an existing boiler, or for hot water uses. The recovery exchanger is then simpler and less expensive than a HRSG. Example presented in section 8.5.1 corresponds to this configuration. Another very efficient scheme is to directly use the exhaust gases as hot fluid in a dryer. As in the case of the reciprocating engine, the turbine itself is placed in the airflow, so that all losses can be recovered, leading to an overall efficiency close to 1. In addition, the output pressure of the exhaust gas is sufficient to avoid any fan. The hot gas temperatures (400 – 500 degrees C) being compatible with many industrial requirements, applications of this method are numerous.

</doc_id=16>
<doc_id=17>
[17] gases contain a lot of oxygen (16 to 18% O2). It is therefore possible to exploit these gases (usually very clean, especially if the fuel used is natural gas), continuing the combustion processes in furnaces or boilers, or even using them directly as drying fluid. Given these characteristics, it is obvious that possible configurations are very diverse in terms of internal combustion engine cogeneration.

8.3.1 RECIPROCATING ENGINES
The simplest and most common solution is the production of either hot water at a temperature of 100 degrees C, or superheated steam at 110–120 degrees C, as auxiliary to a classical boiler (Figure 8.3.1). Depending on its purity, water can be directly heated in the engine, or must pass through a low temperature heat exchanger. It then

recovers the exhaust gas heat in a heat exchanger placed in series. Hundreds of such units of all capacities (a few kW to several MW) are installed worldwide. A second solution is to cool the engine by an air flow which, in series, provides a convective cooling, and then passes through a recovery heat exchanger on oil, the supercharger intercooler if it exists, the classic radiator cooling system, and finally an air/flue gas exhaust. Hot air is then used for drying, its enthalpy being, if necessary, raised by a supplementary burner. The engine can also be used for air conditioning, driving a compressor directly connected on the shaft, to obtain industrial cooling or chilled water, heat recovered

</doc_id=17>
</retrieved_passages>

SFT+GRPO evaluated output:

<answer>
[3] > [1] > [5] > [7] > [8] > [9] > [11] > [12] > [13] > [14] > [15] > [16] > [17] > [2] > [6]
</answer>

SARD. SARD selected input window:

<retrieved_passages>
<doc_id=1>
[1] At the Energy Systems Integration Facility (ESIF), heat exchangers in the High-Performance Computing Data Center enable the reuse of energy from computing waste heat. With heat exchangers, heat energy in the energy recovery water (ERW) loop becomes available to heat the facility's process hot water (PHW) loop. Once heated, the PHW loop supplies:
* Active chilled beams to heat the office space
* Air handlers to heat the conference and high bay spaces
* Snow melt loop in the courtyard of the ESIF's main entrance
* District heating loop:
* If additional heat is needed for the building, the PHW loop can draw heat from the campus heating loop
* If surplus heat is available, the PHW loop can provide heat to the campus heating loop. During transition months (April and October), excess heat from the ESIF has been sufficient to provide heat for other buildings, shortening the time boilers are needed to provide campus heat.

</doc_id=1>
<doc_id=2>
[2] SUSTAINABILITY
Surplus datacenter heat will be repurposed to heat homes in Denmark
View more
</doc_id=2>
<doc_id=3>
[3] Waste combustion is particularly popular in countries such as Japan, Singapore and the Netherlands, where land is a scarce resource. Denmark and Sweden have been leaders by using the energy generated from incineration for more than a century, in localised [combined heat and power](/wiki/Combined_heat_and_power "Combined heat and power") facilities supporting [district heating](/wiki/District_heating "District heating") schemes. (4) In 2005, waste incineration produced 4.8% of the electricity consumption and 13.7% of the total domestic heat consumption in Denmark. (5) A number of other European countries rely heavily on incineration for handling municipal waste, in particular [Luxembourg](/wiki/Luxembourg "Luxembourg"), the Netherlands, Germany, and France. (2)

</doc_id=3>
<doc_id=4>
[4] views the problem in terms of primary energy saving: he must justify his choices based on the micro-economic context in which they operate. Before deciding to use a cogeneration facility, it is necessary to make an extensive study of energy and thermal needs and their evolution over time. Indeed, the corresponding investments are generally high, and, to amortize them, the facility must operate at its optimum economic point as long as possible when the energy price is justified, and if possible at about 80 – 90% of the maximum engine capacity range leading to the best technical performance. The profitability calculation depends fundamentally on operating conditions and their evolution over time, each facility representing a special case that requires detailed study. It is particularly important to properly ensure that the statutory criteria authorizing the resale of electricity to utilities will be respected. If security considerations require the use of an independent mechanical power production unit, it is almost certain that its use in cogeneration will be profitable, the overhead being limited to expenses for heat recovery, often well below those corresponding to the engine and alternator. Moreover, the technical solution choice depends on many factors. For purposes of comparison, Table 8.2 provides approximate values of the different performance indicators for key possible technologies. The first two rows correspond to what is now called micro-cogeneration, based on Stirling engines or small capacity micro

</doc_id=4>
<doc_id=5>
[5] * Might not be suitable for all types of data centers
* Brings potential risks associated with leaks
* Costly and tricky to implement
</doc_id=5>
<doc_id=6>
[6] * Can use [vast amounts](https://www.washingtonpost.com/climate-environment/2023/04/25/data-centers-drought-water-use/) of the local water supply.
* Coolant leaks can be environmentally unfriendly if not addressed properly

```

* Complex to set up and can have several potential points of failure
</doc_id=6>
<doc_id=7>
[7] * Limited to regions with colder climates or cold seasons
* Variability in external conditions can affect performance
* May require backup cooling solutions in case of unpredicted changes
</doc_id=7>
<doc_id=8>
[8] * Performance can decrease in the wrong conditions
* High dependence on water availability and quality
* Can require an overhaul of existing data center infrastructure
</doc_id=8>
<doc_id=9>
[9] Incinerators have electric efficiencies of 14–28%. (10) In order to avoid
losing the rest of the energy, it can be used for e.g. [ district heating
]/wiki/District_heating "District heating") ( [ cogeneration
]/wiki/Cogeneration "Cogeneration") ). The total efficiencies of cogeneration
incinerators are typically higher than 80% (based on the [ lower heating value
]/wiki/Lower_heating_value "Lower heating value") of the waste).
</doc_id=9>
<doc_id=10>
[10] being used for purposes of either heating or cooling in an absorption machine.
An
alternator can at times be coupled to the engine instead of the compressor, which
Cogeneration and trigeneration 223
allows, according to the tariffs of electricity and refrigeration needs, to modulate the
production.
The reciprocating engine may finally directly drive a heat pump compressor. The
overall efficiency of the system can be very high, given the coefficients of
performance of heat pumps.
The heat–power ratio CF is quite low, between 0.5 and 1.5. Overall efficiency is
generally very good, above 70%. Mechanical efficiency is usually very high,
between 30 and 35% for small gas engines, and up to 45% for large diesel and gas
engines. Specific equivalent consumption CE is of the order of 1.6 to 2.
8.3.2 GAS TURBINES
In gas turbines, the total residual heat is found in the exhaust. The performance of
the cogeneration system is directly related to the recovery of these gases.
One solution is to cool exhaust gases in an air–fumes heat exchanger which can
heat
the air which is then used for many applications. If the turbine is stopped, an
auxiliary boiler ensures the supply of heat to meet the plant needs. One generally
uses several cascading recovery exchangers on the exhaust, so it can be cooled as
much as possible and air is
</doc_id=10>
<doc_id=11>
[11] available at different
temperatures for various
uses.
Another solution, widely
used today, especially to
replace an existing boiler, is
to install a heat recovery
steam generator (HRSG) at
the output of the gas turbine.
The problem of optimizing
the recovery exchangers is
very similar to the one we
have discussed in section
Figure 8.3.1: Internal combustion engine cogeneration unit
Figure 8.3.2: Cogeneration gas turbine,
Documentation SECC
224 Energy systems
7.3: the best configuration must indeed both cool at best the GT exhaust and
provide
heat at the highest temperature level possible as needed.
Figures 8.3.2 and 8.3.3 show the two main components of a cogeneration plant for
</doc_id=11>
</retrieved_passages>

```

SARD evaluated output:

```

<answer>
[1] > [2] > [3] > [5] > [6] > [7] > [8] > [4] > [9] > [10] > [11]
</answer>

```

Analysis. All three methods retrieve plausible evidence about data centers, cooling, or heat reuse, so this is not a case where the weaker systems simply leave the topic. SFT emphasizes cooling techniques and operational constraints, which are relevant because the query begins from data-center cooling, but several of its highest-ranked passages are only indirectly useful for the user's main question. SFT+GRPO improves the ranking by moving concrete reuse evidence closer to the top, yet it still interleaves broad sustainability and cooling passages with the direct answer. SARD ranks pas-

sages about actual heat-exchanger reuse, district heating, and homes/offices heated by surplus data-center heat before turning to obstacle and practicality evidence. This ordering better follows the two-part user intent: first establish that reuse is feasible, then explain why it is not universal.

C.3.2 Case 2: Life of a dollar

Dataset / query id. economics / 9.

Qualitative focus. The query asks for a concrete diagrammatic account of how a dollar moves through major economic sectors. The key distinction is between circular-flow evidence and adjacent passages about banking or money creation.

Full top-100 NDCG@10. SFT: 0.261; SFT+GRPO: 0.709; SARD: 0.822.

Full query:

Is there a diagram outlining the life of a dollar?

I would like to understand the economic system concretely. My reading leads quickly into graphs and theory, which is interesting but not what I'm looking for. Instead, I imagine I should be able to trace the path of a dollar from my pocket through the major components of the economic system (the central bank, commercial banks, investors, companies, tax, government spending, government borrowing...)

SFT. SFT selected input window:

```

<retrieved_passages>
<doc_id=1>
[1] Overview [ [ edit
]/w/index.php?title=Circular_flow_of_income&action=edit&section=1 "Edit
section: Overview") ]

[

]/wiki/File:Circular_flow_of_income_and_expenditure.jpg) Model of the
circular flow of income and expenditure

The circular flow of income is a concept for better understanding of the [
economy ](/wiki/Economy "Economy") as a whole and for example the [ National
Income and Product Accounts ](/wiki/National_Income_and_Product_Accounts
"National Income and Product Accounts") (NIPAs). In its most basic form it
considers a simple economy consisting solely of [ businesses ](/wiki/Business
"Business") and individuals, and can be represented in a so-called "circular
flow diagram." In this simple economy, individuals provide the labour that
enables businesses to produce goods and services. These activities are
represented by the green lines in the diagram. (5)

Alternatively, one can think of these transactions in terms of the monetary
flows that occur. Businesses provide individuals with income (in the form of
compensation) in exchange for their labor. That income is spent on the goods
and services businesses produce. These activities are represented by the blue
lines in the diagram above. (5)

The circular flow diagram illustrates the interdependence of the "flows," or
activities, that occur in the economy, such as the [ production
]/wiki/Production_(economics)\ "Production (economics)") of goods and
services (or the "output" of the economy) and the income generated from that
production. The circular flow also illustrates the equality between the income
earned from production and the value of goods and services produced. (5)

Of course, the total economy is much more complicated than the illustration
above. An economy involves interactions between not only individuals and
businesses, but also Federal, state, and local governments and residents of
the rest of the world. Also not shown in this simple illustration of the
economy are other aspects of economic activity such as investment in [ capital
]/wiki/Capital_(economics)\ "Capital (economics)") (produced—or
fixed—
</doc_id=1>
<doc_id=2>
[2] Leakages and injection [ [ edit
]/w/index.php?title=Circular_flow_of_income&action=edit&section=14 "Edit
section: Leakages and injection") ]

```

The five–sector model considers leakages and injections. * 'Leakage' means withdrawal from the flow. When households and firms save part of their incomes it constitutes leakage. They may be in form of savings, tax payments, and [imports](/wiki/Imports "Imports") . Leakages reduce the flow of income. * 'Injection' means the introduction of income into the flow. When households and firms borrow savings, they constitute injections. Injections increase the flow of income. Injections can take the forms of investment, government spending and [exports](/wiki/Exports "Exports") . As long as leakages are equal to injections, the circular flow of income continues indefinitely. Financial institutions or capital market play the role of intermediaries. This means that income individuals receive from businesses and the goods and services that are sold to them do not count as injections or leakages, as no new money is being introduced to the flow and no money is being taken out of the flow. Leakages and injections can occur in the financial sector, government sector and overseas sector: </doc_id=2> <doc_id=3> [3] Five–sector model [[edit](/w/index.php?title=Circular_flow_of_income&action=edit§ion=11 "Edit section: Five–sector model")] The five–sector model adds the financial sector to the four–sector model. (19) Thus, the five–sector model includes (1) households, (2) firms, (3) government, (4) the rest of the world, and (5) the financial sector. The financial sector includes banks and non–bank intermediaries that engage in borrowing (savings from households) and lending (investments in firms). (19) Money facilitates such an exchange smoothly. Residuals from each market enter the [capital market](/wiki/Capital_market "Capital market") as savings, which in turn are invested in firms and the government sector. Technically speaking, so long as lending is equal to borrowing (i.e., leakages are equal to injections), the circular flow will continue indefinitely. However, this job is done by financial institutions in the economy. * [![Five–sector circular flow of income model](//upload.wikimedia.org/wikipedia/commons/thumb/4/4a/Five_Sector_Circular_Flow_of_Income_Model.jpg/111px–Five_Sector_Circular_Flow_of_Income_Model.jpg)](/wiki/File:Five_Sector_Circular_Flow_of_Income_Model.jpg "Five–sector circular flow of income model") Five–sector circular flow of income model * [![Circular flow diagram, five–sectors model](//upload.wikimedia.org/wikipedia/commons/thumb/b/bc/04_circular_flow_diagram.png/82px–04_circular_flow_diagram.png)](/wiki/File:04_circular_flow_diagram.png "Circular flow diagram, five–sectors model") Circular flow diagram, five–sectors model </doc_id=3> <doc_id=4> [4] Three–sector model [[edit](/w/index.php?title=Circular_flow_of_income&action=edit&section=9 "Edit section: Three–sector model")] [![//upload.wikimedia.org/wikipedia/commons/thumb/c/cf/Economics_circular_flow_cartoon.jpg/220px–Economics_circular_flow_cartoon.jpg)](/wiki/File:Economics_circular_flow_cartoon.jpg) Three–sector circular flow diagram The three–sector model adds the government sector to the two–sector model. (17) (18) Thus, the three–sector model includes (1) households, (2) firms, and (3) government. It excludes the financial sector and the foreign sector. The government sector consists of the economic activities of local, state and federal governments. Flows from households and firms to government are in the form of taxes. The income the government receives flows to firms and households in the form of subsidies, transfers, and purchases of goods and services. (17) (18) Every payment has a corresponding receipt; that is, every flow of money has a corresponding flow of goods in the opposite direction. (17) (18) As a result, the [aggregate expenditure](/w/index.php?title=Aggregate_expenditure&action=edit&redlink=1 "Aggregate expenditure \ (page does not exist)") of the economy is identical to its [aggregate income](/wiki/Aggregate_income "Aggregate income") , making a circular flow. </doc_id=4> <doc_id=5> [5] In the government sector [[edit](/w/index.php?title=Circular_flow_of_income&action=edit§ion=16 "Edit section: In the government sector")] The leakage that the government sector provides is through the collection of revenue through taxes (T) that is provided by households and firms to the government. This is a leakage because it is a leakage out of the current income, thus reducing the expenditure on current goods and services. The injection provided by the government sector is government spending (G) that provides collective services and welfare payments to the community. An example of a tax collected by the government as a leakage is [income tax	](/wiki/Income_tax "Income tax") , and an injection into the economy can be when the government redistributes this income in the form of [welfare payments](/wiki/Social_security "Social security") , that is a form of government spending back into the economy. </doc_id=5> <doc_id=6> [6] Further developments [[edit](/w/index.php?title=Circular_flow_of_income&action=edit§ion=6 "Edit section: Further developments")] [![//upload.wikimedia.org/wikipedia/commons/thumb/b/b4/The_competitive_price_system_adapted_from_Samuelson%2C_1961.jpg/220px–The_competitive_price_system_adapted_from_Samuelson%2C_1961.jpg)](/wiki/File:The_competitive_price_system_adapted_from_Samuelson,_1961.jpg) The competitive price system adapted from [Samuelson](/wiki/Paul_Samuelson "Paul Samuelson") , 1961 An important development was [John Maynard Keynes](/wiki/John_Maynard_Keynes "John Maynard Keynes") 's 1933 publication of the _[General Theory of Employment, Interest and Money](/wiki/General_Theory_of_Employment,_Interest_and_Money "General Theory of Employment, Interest and Money") _ . Keynes' assistant [Richard Stone](/wiki/Richard_Stone "Richard Stone") further developed the concept for the [United Nations](/wiki/United_Nations "United Nations") (UN) and the [Organisation for Economic Co–operation and Development](/wiki/Organisation_for_Economic_Co-operation_and_Development "Organisation for Economic Co–operation and Development") to the systems, which is now used internationally. The first to visualize the modern circular flow of income model was [Frank Knight](/wiki/Frank_Knight "Frank Knight") in 1933 publication of _The Economic Organization._ (13) Knight (1933) explained: _[we may view the] economic organization as a system of prize relations. Seen in the large, free enterprise is an organization of production and distribution in which individuals or family units get their real income, their "living," by selling productive power for money to "business units" or "enterprises", and buying with the money income thus obtained the direct goods and services which they consume. This view, it will be remembered, ignores for the sake of simplicity the fact that an appreciable fraction of the productive power in use at any time is not really employed in satisfying current wants but to make provision for increased want–satisfaction in </doc_id=6> <doc_id=7> [7] ![//upload.wikimedia.org/wikipedia/commons/thumb/b/b8/Circular_flow_of_goods_income.png/360px–Circular_flow_of_goods_income.png)](/wiki/File:Circular_flow_of_goods_income.png) Basic diagram of the circular flow of income. The functioning of the free–market economic system is represented with [firms](/wiki/Firm "Firm") and [households](/wiki/Household "Household") and interaction back and forth. (2) The circular flow of income or circular flow is a model](/wiki/Economic_model "Economic model") of the [economy](/wiki/Economy "Economy") in which the major exchanges are represented as flows of [money](/wiki/Money "Money") , [goods](/wiki/Good_(economics) "Good (economics)") and [services](/wiki/Service_(economics) "Service (economics)") , etc. between [economic agents](/wiki/Economic_agents "Economic agents") . The flows of money and goods exchanged in a closed circuit correspond in value, but run in the opposite direction. The circular flow analysis is the basis of [national accounts](/wiki/National_accounts "National accounts") and hence of [macroeconomics](/wiki/Macroeconomics "Macroeconomics") . The idea of the circular flow was already present in the work of [Richard Cantillon](/wiki/Richard_Cantillon "Richard Cantillon") . (3) [Francois Quesnay](/wiki/Fran%C3%A7ois_Quesnay "Francois Quesnay") developed and visualized this concept in the so–called [Tableau economique](/wiki/Tableau_%C3%A9conomique "Tableau economique") . (4) Important developments of Quesnay's tableau were [Karl Marx](/wiki/Karl_Marx "Karl Marx") 's reproduction schemes in the second volume of _[Capital: Critique of Political Economy](/wiki/Capital:_Critique_of_Political_Economy "Capital: Critique of Political Economy") _ , and [John Maynard Keynes](/wiki/John_Maynard_Keynes "John Maynard Keynes") ' _[General Theory of Employment, Interest and </doc_id=7> <doc_id=8> [8] Two–sector model [[edit](/w/index.php?title=Circular_flow_of_income&action=edit§ion=8 "Edit section: Two–sector model")] [![//upload.wikimedia.org/wikipedia/commons/thumb/c/c2/Two_sector_circular_flow_diagram.jpg/220px–Two_sector_circular_flow_diagram.jpg)](/wiki/File:Two_sector_circular_flow_diagram.jpg) Two–sector circular flow diagram In the basic two–sector circular flow of income model, the economy consists of two [sectors](/wiki/Economic_sector "Economic sector") : (1) [households
---	--

[/wiki/Households "Households"\)](#) and (2) [\[firms \]\(/wiki/Theory_of_the_firm "Theory of the firm"\)](#) . (17) (18) (Some sources refer to households as "individuals" (19) or the "public" (20) and to firms as "businesses" (1) (2) or the "productive sector." (21)) The model assumes that there is no [\[financial sector \]\(/wiki/Financial_sector "Financial sector"\)](#) , no [\[government sector \]\(/wiki/Government_sector "Government sector"\)](#) , and no [\[foreign sector \]\(/wiki/External_sector "External sector"\)](#) . In addition, the model assumes that (a) through their expenditures, households spend all of their [\[income \]\(/wiki/Income "Income"\)](#) on [\[goods and services \]\(/wiki/Goods_and_services "Goods and services"\)](#) or [\[consumption \]\(/wiki/Consumption_\(economics\) "Consumption \(economics\)"\)](#) and (b) through their [\[expenditures \]\(/wiki/Expenditure "Expenditure"\)](#) , households purchase all output produced by firms. (18) This means that all household expenditures become income for firms. The firms then spend all of this income on [\[factors of production \]\(/wiki/Factors_of_production "Factors of production"\)](#) such as labor, capital and raw materials, "transferring" all of their income to the factor owners (which are households). The factor owners (households), in turn, spend all of their income on goods, which leads to a circular flow of income. (20) (18) [22]

</doc_id=8>
 <doc_id=9>

[9] While new broad money has been created on the consumer's balance sheet, the first row of Figure 1 shows that this is without — in the first instance, at least — any change in the amount of central bank money or 'base money'. As discussed earlier, the higher stock of deposits may mean that banks want, or are required, to hold more central bank money in order to meet withdrawals by the public or make payments to other banks. And reserves are, in normal times, supplied 'on demand' by the Bank of England to commercial banks in exchange for other assets on their balance sheets. In no way does the aggregate quantity of reserves directly constrain the amount of bank lending or deposit creation.

This description of money creation contrasts with the notion that banks can only lend out pre-existing money, outlined in the previous section. Bank deposits are simply a record of how much the bank itself owes its customers. So they are a liability of the bank, not an asset that could be lent out. A related misconception is that banks can lend out their reserves. Reserves can only be lent between banks, since consumers do not have access to reserves accounts at the Bank of England.(2) Other ways of creating and destroying deposits

Just as taking out a new loan creates money, the repayment of bank loans destroys money.(3) For example, suppose a consumer has spent money in the supermarket throughout the month by using a credit card. Each purchase made using the Figure 1 Money creation by the aggregate banking sector making additional loans(a)

(a) Balance sheets are highly stylised for ease of exposition: the quantities of each type of money shown do not correspond to the quantities actually held on each sector's balance sheet.

(b) Central bank balance sheet only shows base money liabilities and the corresponding assets.

In practice the central bank holds other non-money liabilities. Its non-monetary assets are mostly made up of government debt. Although that government debt is actually held by the Bank of England Asset Purchase Facility, so does not appear directly on the balance sheet.

(c) Commercial banks' balance sheets only show money assets and liabilities before any loans are made.

(d) Consumers represent the private sector of households and companies. Balance sheet only shows broad money assets and corresponding liabilities — real assets such as the house being transacted are not shown. Consumers' non-money liabilities include existing secured and unsecured loans.

credit card will
 </doc_id=9>
 <doc_id=10>

[10] Reserves

Currency
 New
 deposits
 Deposits
 Assets Liabilities
 Reserves
 Currency
 Deposits
 Consumers(d)
 Assets Liabilities
 New
 deposits
 Deposits
 Currency
 New loans
 Non-money
 Assets Liabilities Broad
 money
 Non-money
 money
 Broad Deposits
 Currency

16 Quarterly Bulletin 2014 Q1

Commercial banks create money, in the form of bank deposits, by making new loans. When a bank makes a loan, for example to someone taking out a mortgage to buy a house, it does not typically do so by giving them thousands of pounds worth of banknotes. Instead, it credits their bank account with a bank deposit of the size of the mortgage. At that moment, new money is created. For this reason, some economists have referred to bank deposits as 'fountain pen money', created at the stroke of bankers' pens when they approve loans.(1)

This process is illustrated in Figure 1, which shows how new lending affects the balance sheets of different sectors of the economy (similar balance sheet diagrams are introduced in 'Money in the modern economy: an introduction'). As shown in the third row of Figure 1, the new deposits increase the assets of the consumer (here taken to represent households and companies) — the extra red bars — and the new loan increases their liabilities — the extra white bars. New broad money has been created. Similarly, both sides of the commercial banking sector's balance sheet increase as new money and loans are created. It is important to note that although the simplified diagram of Figure 1 shows the amount of new money created as being identical to the amount of new lending, in practice there will be several factors that may subsequently cause the amount of deposits to be different from the amount of lending. These are discussed in detail in the next section.

</doc_id=10>
 <doc_id=11>

[11] 14 Quarterly Bulletin 2014 Q1

Money creation in the modern economy

By Michael McLeay, Amar Radia and Ryland Thomas of the Bank's Monetary Analysis Directorate.(1)

* This article explains how the majority of money in the modern economy is created by commercial banks making loans.

* Money creation in practice differs from some popular misconceptions — banks do not act simply as intermediaries, lending out deposits that savers place with them, and nor do they 'multiply up' central bank money to create new loans and deposits.

* The amount of money created in the economy ultimately depends on the monetary policy of the central bank. In normal times, this is carried out by setting interest rates. The central bank can also affect the amount of money directly through purchasing assets or 'quantitative easing'.

Overview

In the modern economy, most money takes the form of bank low and stable inflation. In normal times, the Bank of deposits. But how those bank deposits are created is often England implements monetary policy by setting the interest rate on central bank reserves. This then influences a range of banks making loans. Whenever a bank makes a loan, it interest rates in the economy , including those on bank loans.

simultaneously creates a matching deposit in the

In exceptional circumstances, when interest rates are at their borrower's bank account, thereby creating new money.

effective lower bound, money creation and spending in the

The reality of how money is created today differs from the economy may still be too low to be consistent

with the description found in some economics textbooks: central bank's monetary policy objectives. One possible response is to undertake a series of asset purchases, or * Rather than banks receiving deposits when households 'quantitative easing' (QE). QE is intended to boost the save and then lending them out, bank lending creates amount of money in the economy directly by purchasing deposits. assets, mainly from non-bank financial companies. * In normal times, the central bank does not fix the amount of money in circulation, nor is central bank money QE initially increases the amount of bank deposits those 'multiplied up' into more loans and deposits. companies hold (in place of the assets they sell). Those companies will then wish to rebalance their portfolios of Although commercial banks create money through lending, assets by buying higher-yielding assets, raising the price of they cannot do so freely without

limit
 </doc_id=11>
 <doc_id=12>

[12] much central bank money banks want to hold in reserve (to meet withdrawals by the public, make payments to other banks, or meet regulatory liquidity requirements), which is then, in normal times, supplied on demand by the Bank of England. The rest of this article discusses these practices in more detail.

Money creation in reality

Lending creates deposits — broad money

determination at the aggregate level

As explained in 'Money in the modern economy: an

introduction', broad money is a measure of the total amount of money held by households and companies in the economy. Broad money is made up of bank deposits — which are essentially IOUs from commercial banks to households and companies — and currency — mostly IOUs from the central bank.(4)(5) Of the two types of broad money, bank deposits make up the vast majority — 97% of the amount currently in circulation.(6) And in the modern economy, those bank deposits are mostly created by commercial banks themselves.

(1) Throughout this article, 'banks' and 'commercial banks' are used to refer to banks and building societies together.

(2) See www.youtube.com/watch?v=CvRAqR2pAgw.

(3) There is a long literature that does recognise the 'endogenous' nature of money creation in practice. See, for example, Moore (1988), Howells (1995) and Palley (1996).

(4) The definition of broad money used by the Bank of England, M4ex, also includes a wider range of bank liabilities than regular deposits; see Burgess and Janssen (2007) for more details. For simplicity, this article describes all of these liabilities as deposits.

A box later in this article provides details about a range of popular monetary aggregates in the United Kingdom.

(5) Around 6% of the currency in circulation is made up of coins, which are produced by The Royal Mint. Of the banknotes that circulate in the UK economy, some are issued by some Scottish and Northern Irish commercial banks, although these are fully matched by Bank of England money held at the Bank.

(6) As of December 2013.

Before loans are made After loans are made

Central bank(b)	Assets	Liabilities
Reserves		
Currency		
Base money		
Non-money		

</doc_id=12>
<doc_id=13>
[13] shorthand way of describing the process of increasing lending and deposits until the bank reaches its maximum ratio. As there are no reserve requirements in the United Kingdom the process is less relevant for UK banks.

(3) The fall in bank lending in the United Kingdom since 2008 is an important reason why the growth of money in the economy has been so much lower than in the years leading up to the crisis, as discussed in Bridges, Rossiter and Thomas (2011) and Butt et al (2012).

Topical articles Money creation in the modern economy 17
bill in full at the end of the month, its bank would reduce the amount of deposits in the consumer's account by the value of the credit card bill, thus destroying all of the newly created money.

Banks making loans and consumers repaying them are the most significant ways in which bank deposits are created and destroyed in the modern economy. But they are far from the only ways. Deposit creation or destruction will also occur any time the banking sector (including the central bank) buys or sells existing assets from or to consumers, or, more often, from companies or the government.

Banks buying and selling government bonds is one particularly important way in which the purchase or sale of existing assets by banks creates and destroys money. Banks often buy and hold government bonds as part of their portfolio of liquid assets that can be sold on quickly for central bank money if, for example, depositors want to withdraw currency in large amounts.(1) When banks purchase government bonds from the non-bank private sector they credit the sellers with bank deposits.(2) And, as discussed later in this article, central bank asset purchases, known as quantitative easing (QE), have similar implications for money creation.

Money can also be destroyed through the issuance of long-term debt and equity instruments by banks. In addition to deposits, banks hold other liabilities on their balance sheets. Banks manage their liabilities to ensure that they have at least some capital and longer-term debt liabilities to mitigate certain risks and meet regulatory requirements. Because these 'non-deposit' liabilities represent longer-term investments in the banking system by households and companies, they cannot be exchanged for currency as easily as bank deposits, and therefore increase the resilience of the bank. When banks issue these longer-term debt and equity instruments to non-bank financial

</doc_id=13>
<doc_id=14>
[14] form of a house and a new liability in the form of a new loan. The seller is left with money in the form of bank deposits instead of a house. It is more likely than not that the seller's account will be with a different bank to the buyer's. So when the transaction takes place, the new deposits will be transferred to the seller's bank, as shown in the second row of Figure 2. The buyer's bank would then have fewer deposits than assets. In the first instance, the buyer's bank settles with the seller's bank by transferring reserves. But that would leave

the buyer's bank with fewer reserves and more loans relative to its deposits than before. This is likely to be problematic for the bank since it would increase the risk that it would not be able to meet all of its likely outflows. And, in practice, banks make many such loans every day. So if a given bank financed all of its new loans in this way, it would soon run out of reserves.

Banks therefore try to attract or retain additional liabilities to accompany their new loans. In practice other banks would also be making new loans and creating new deposits, so one way they can do this is to try and attract some of those newly created deposits. In a competitive banking sector, that may involve increasing the rate they offer to households on their savings accounts. By attracting new deposits, the bank can increase its lending without running down its reserves, as shown in the third row of Figure 2. Alternatively, a bank can borrow from other banks or attract other forms of liabilities, at least temporarily. But whether through deposits or other liabilities, the bank would need to make sure it was attracting and retaining some kind of funds in order to keep expanding lending. And the cost of that needs to be measured against the interest the bank expects to earn on the loans it is making, which in turn depends on the level of Bank Rate set by the Bank of England. For example, if a bank continued to attract new borrowers and increase lending by reducing mortgage rates, and sought to attract new deposits by increasing the rates it was paying on its customers' deposits, it might soon find it unprofitable to keep expanding its lending. Competition for loans and deposits, and the desire to make a profit, therefore limit

</doc_id=14>
<doc_id=15>
[15] Significance of study of circular flow of income [[edit](/w/index.php?title=Circular_flow_of_income&action=edit§ion=20 "Edit section: Significance of study of circular flow of income")]

The circular flow of income is significant in four areas: (24)

- [Measurement of national income](/wiki/Measures_of_national_income_and_output "Measures of national income and output")
- Knowledge of Interdependence – Circular flow of income signifies the interdependence of each of activity upon one another. If there is no consumption, there will be no demand and expenditure which in fact restricts the amount of production and income.
- Unending Nature of Economic Activities – It signifies that production, income and expenditure are of unending nature, therefore, economic activities in an economy can never come to a halt. National income is also bound to rise in future.
- Injections and Leakages

</doc_id=15>
<doc_id=16>
[16] Four-sector model [[edit](/w/index.php?title=Circular_flow_of_income&action=edit§ion=10 "Edit section: Four-sector model")]

The four-sector model adds the foreign sector to the three-sector model. (17) (18) (23) (The foreign sector is also known as the "external sector," the "overseas sector," (19) or the "rest of the world.") Thus, the four-sector model includes (1) households, (2) firms, (3) government, and (4) the rest of the world. It excludes the financial sector. The foreign sector comprises (a) foreign trade (imports and exports of goods and services) and (b) inflow and outflow of capital (foreign exchange). (18) Again, each flow of money has a corresponding flow of goods (or services) in the opposite direction. (18) Each of the four sectors receives some payments from the other in lieu of goods and services which makes a regular flow of goods and physical services. The addition of the foreign sector

</doc_id=16>
</retrieved_passages>

SFT evaluated output:

```
<answer>
[9] > [10] > [11] > [12] > [13] > [14] > [3] > [16] > [4] > [5] > [2] > [1] > [7] > [8]
> [6] > [15]
</answer>
```

SFT+GRPO. SFT+GRPO selected input window:

```
<retrieved_passages>
<doc_id=1>
[1] Overview [ [ edit ](/w/index.php?title=Circular_flow_of_income&action=edit&section=1 "Edit section: Overview") ]

[

]/wiki/File:Circular_flow_of_income_and_expenditure.jpg) Model of the
circular flow of income and expenditure
```

The circular flow of income is a concept for better understanding of the [economy](/wiki/Economy "Economy") as a whole and for example the [National Income and Product Accounts](/wiki/National_Income_and_Product_Accounts "National Income and Product Accounts") (NIPAs). In its most basic form it considers a simple economy consisting solely of [businesses](/wiki/Business "Business") and individuals, and can be represented in a so-called "circular flow diagram." In this simple economy, individuals provide the labour that enables businesses to produce goods and services. These activities are represented by the green lines in the diagram. (5)

Alternatively, one can think of these transactions in terms of the monetary flows that occur. Businesses provide individuals with income (in the form of compensation) in exchange for their labor. That income is spent on the goods and services businesses produce. These activities are represented by the blue lines in the diagram above. (5)

The circular flow diagram illustrates the interdependence of the "flows," or activities, that occur in the economy, such as the [production](/wiki/Production_(economics) "Production (economics)") of goods and services (or the "output" of the economy) and the income generated from that production. The circular flow also illustrates the equality between the income earned from production and the value of goods and services produced. (5)

Of course, the total economy is much more complicated than the illustration above. An economy involves interactions between not only individuals and businesses, but also Federal, state, and local governments and residents of the rest of the world. Also not shown in this simple illustration of the economy are other aspects of economic activity such as investment in [capital](/wiki/Capital_(economics) "Capital (economics)") (produced—or fixed—

</doc_id=1>
<doc_id=2>
[2] Leakages and injection [[edit](/w/index.php?title=Circular_flow_of_income&action=edit§ion=14 "Edit section: Leakages and injection")]

The five–sector model considers leakages and injections.

* 'Leakage' means withdrawal from the flow. When households and firms save part of their incomes it constitutes leakage. They may be in form of savings, tax payments, and [imports](/wiki/Imports "Imports") . Leakages reduce the flow of income.

* 'Injection' means the introduction of income into the flow. When households and firms borrow savings, they constitute injections. Injections increase the flow of income. Injections can take the forms of investment, government spending and [exports](/wiki/Exports "Exports") . As long as leakages are equal to injections, the circular flow of income continues indefinitely. Financial institutions or capital market play the role of intermediaries. This means that income individuals receive from businesses and the goods and services that are sold to them do not count as injections or leakages, as no new money is being introduced to the flow and no money is being taken out of the flow.

Leakages and injections can occur in the financial sector, government sector and overseas sector:

</doc_id=2>
<doc_id=3>
[3] Five–sector model [[edit](/w/index.php?title=Circular_flow_of_income&action=edit§ion=11 "Edit section: Five–sector model")]

The five–sector model adds the financial sector to the four–sector model. (19) Thus, the five–sector model includes (1) households, (2) firms, (3) government, (4) the rest of the world, and (5) the financial sector. The financial sector includes banks and non–bank intermediaries that engage in borrowing (savings from households) and lending (investments in firms). (19) Money facilitates such an exchange smoothly. Residuals from each market enter the [capital market](/wiki/Capital_market "Capital market") as savings, which in turn are invested in firms and the government sector. Technically speaking, so long as lending is equal to borrowing (i.e., leakages are equal to injections), the circular flow will continue indefinitely. However, this job is done by financial institutions in the economy.

* [!Five–sector circular flow of income model](/upload.wikimedia.org/wikipedia/commons/thumb/4/4a/Five_Sector_Circular_Flow_of_Income_Model.jpg/111px–Five_Sector_Circular_Flow_of_Income_Model.jpg)](/wiki/File:Five_Sector_Circular_Flow_of_Income_Model.jpg "Five–sector circular flow of income model")

Five–sector circular flow of income model

* [!Circular flow diagram, five–sectors model](/upload.wikimedia.org/wikipedia/commons/thumb/b/bc/04_circular_flow_diagram.png/82px–04_circular_flow_diagram.png)](/wiki/File:04_circular_flow_diagram.png "Circular flow diagram, five–sectors model")

Circular flow diagram, five–sectors model

</doc_id=3>
<doc_id=4>
[4] Three–sector model [[edit](/w/index.php?title=Circular_flow_of_income&action=edit§ion=9 "Edit section: Three–sector model")]

[!](/upload.wikimedia.org/wikipedia/commons/thumb/c/cf/Economics_circular_flow_cartoon.jpg/220px–Economics_circular_flow_cartoon.jpg)](/wiki/File:Economics_circular_flow_cartoon.jpg) Three–sector circular flow diagram

The three–sector model adds the government sector to the two–sector model. (17) (18) Thus, the three–sector model includes (1) households, (2) firms, and (3) government. It excludes the financial sector and the foreign sector. The government sector consists of the economic activities of local, state and federal governments. Flows from households and firms to government are in the form of taxes. The income the government receives flows to firms and households in the form of subsidies, transfers, and purchases of goods and services. (17) (18) Every payment has a corresponding receipt; that is, every flow of money has a corresponding flow of goods in the opposite direction. (17) (18) As a result, the [aggregate expenditure](/w/index.php?title=Aggregate_expenditure&action=edit&redlink=1 "Aggregate expenditure (page does not exist)") of the economy is identical to its [aggregate income](/wiki/Aggregate_income "Aggregate income") , making a circular flow.

</doc_id=4>
<doc_id=5>
[5] In the government sector [[edit](/w/index.php?title=Circular_flow_of_income&action=edit§ion=16 "Edit section: In the government sector")]

The leakage that the government sector provides is through the collection of revenue through taxes (T) that is provided by households and firms to the government. This is a leakage because it is a leakage out of the current income, thus reducing the expenditure on current goods and services. The injection provided by the government sector is government spending (G) that provides collective services and welfare payments to the community. An example of a tax collected by the government as a leakage is [income tax](/wiki/Income_tax "Income tax") , and an injection into the economy can be when the government redistributes this income in the form of [welfare payments](/wiki/Social_security "Social security") , that is a form of government spending back into the economy.

</doc_id=5>
<doc_id=6>
[6] Further developments [[edit](/w/index.php?title=Circular_flow_of_income&action=edit§ion=6 "Edit section: Further developments")]

[!](/upload.wikimedia.org/wikipedia/commons/thumb/b/b4/The_competitive_price_system_adapted_from_Samuelson%2C_1961.jpg/220px–The_competitive_price_system_adapted_from_Samuelson%2C_1961.jpg)](/wiki/File:The_competitive_price_system_adapted_from_Samuelson_1961.jpg) The competitive price system adapted from [Samuelson](/wiki/Paul_Samuelson "Paul Samuelson") , 1961

An important development was [John Maynard Keynes](/wiki/John_Maynard_Keynes "John Maynard Keynes") 's 1933 publication of the [General Theory of Employment, Interest and Money](/wiki/General_Theory_of_Employment_Interest_and_Money "General Theory of Employment, Interest and Money") . Keynes' assistant [Richard Stone](/wiki/Richard_Stone "Richard Stone") further developed the concept for the [United Nations](/wiki/United_Nations "United Nations") (UN) and the [Organisation for Economic Co–operation and Development](/wiki/Organisation_for_Economic_Co-operation_and_Development "Organisation for Economic Co–operation and Development") to the systems, which is now used internationally.

The first to visualize the modern circular flow of income model was [Frank Knight](/wiki/Frank_Knight "Frank Knight") in 1933 publication of _The Economic Organization_. (13) Knight (1933) explained:

_ [we may view the] economic organization as a system of prize relations. Seen in the large, free enterprise is an organization of production and distribution in which individuals or family units get their real income, their "living," by selling productive power for money to "business units" or "enterprises", and buying with the money income thus obtained the direct goods and services which they consume. This view, it will be remembered, ignores for the sake of simplicity the fact that an appreciable fraction of the productive power in use at any time is not really employed in satisfying current wants but to make provision for increased want–satisfaction in

</doc_id=6>
<doc_id=7>
[7] Significance of study of circular flow of income [[edit](/w/index.php?title=Circular_flow_of_income&action=edit§ion=20 "Edit section: Significance of study of circular flow of income")]

The circular flow of income is significant in four areas: (24)

- [Measurement of national income](/wiki/Measures_of_national_income_and_output "Measures of national income and output")
- Knowledge of Interdependence – Circular flow of income signifies the interdependence of each of activity

upon one another. If there is no consumption, there will be no demand and expenditure which in fact restricts the amount of production and income.

3. Unending Nature of Economic Activities – It signifies that production, income and expenditure are of unending nature, therefore, economic activities in an economy can never come to a halt. National income is also bound to rise in future.

4. Injections and Leakages

</doc_id=7>

<doc_id=8>

[8] ![/upload.wikimedia.org/wikipedia/commons/thumb/b/b8/Circular_flow_of_goods_income.png/360px–Circular_flow_of_goods_income.png)](/wiki/File:Circular_flow_of_goods_income.png) Basic diagram of the circular flow of income. The functioning of the free–market economic system is represented with [firms](/wiki/Firm "Firm") and [households](/wiki/Household "Household") and interaction back and forth. (2)

The **circular flow of income** or **circular flow** is a [model](/wiki/Economic_model "Economic model") of the [economy](/wiki/Economy "Economy") in which the major exchanges are represented as flows of [money](/wiki/Money "Money") , [goods](/wiki/Good_(economics) "Good \economics)") and [services](/wiki/Service_(economics) "Service \economics)") , etc. between [economic agents](/wiki/Economic_agents "Economic agents") . The flows of money and goods exchanged in a closed circuit correspond in value, but run in the opposite direction. The circular flow analysis is the basis of [national accounts](/wiki/National_accounts "National accounts") and hence of [macroeconomics](/wiki/Macroeconomics "Macroeconomics") .

The idea of the circular flow was already present in the work of [Richard Cantillon](/wiki/Richard_Cantillon "Richard Cantillon") . (3) [Francois Quesnay](/wiki/Fran%C3%A7ois_Quesnay "Francois Quesnay") developed and visualized this concept in the so–called [Tableau economique](/wiki/Tableau_%C3%A9conomique "Tableau economique") . (4) Important developments of Quesnay's tableau were [Karl Marx](/wiki/Karl_Marx "Karl Marx") 's reproduction schemes in the second volume of _[Capital: Critique of Political Economy](/wiki/Capital:_Critique_of_Political_Economy "Capital: Critique of Political Economy") _ , and [John Maynard Keynes](/wiki/John_Maynard_Keynes "John Maynard Keynes") ' _[General Theory of Employment, Interest and

</doc_id=8>

<doc_id=9>

[9] Four–sector model [[edit](/w/index.php?title=Circular_flow_of_income&action=edit§ion=10 "Edit section: Four–sector model")]

The four–sector model adds the foreign sector to the three–sector model. (17) (18) (23) (The foreign sector is also known as the "external sector," the "overseas sector," (19) or the "rest of the world.") Thus, the four–sector model includes (1) households, (2) firms, (3) government, and (4) the rest of the world. It excludes the financial sector. The foreign sector comprises (a) foreign trade (imports and exports of goods and services) and (b) inflow and outflow of capital (foreign exchange). (18) Again, each flow of money has a corresponding flow of goods (or services) in the opposite direction. (18) Each of the four sectors receives some payments from the other in lieu of goods and services which makes a regular flow of goods and physical services. The addition of the foreign sector

</doc_id=9>

<doc_id=10>

[10] While new broad money has been created on the consumer's balance sheet, the first row of Figure 1 shows that this is without — in the first instance, at least — any change in the amount of central bank money or 'base money'. As discussed earlier, the higher stock of deposits may mean that banks want, or are required, to hold more central bank money in order to meet withdrawals by the public or make payments to other banks. And reserves are, in normal times, supplied 'on demand' by the Bank of England to commercial banks in exchange for other assets on their balance sheets. In no way does the aggregate quantity of reserves directly constrain the amount of bank lending or deposit creation. This description of money creation contrasts with the notion that banks can only lend out pre–existing money, outlined in the previous section. Bank deposits are simply a record of how much the bank itself owes its customers. So they are a liability of the bank, not an asset that could be lent out. A related misconception is that banks can lend out their reserves. Reserves can only be lent between banks, since consumers do not have access to reserves accounts at the Bank of England.(2) Other ways of creating and destroying deposits Just as taking out a new loan creates money, the repayment of bank loans destroys money.(3) For example, suppose a consumer has spent money in the supermarket throughout the month by using a credit card. Each purchase made using the Figure 1 Money creation by the aggregate banking sector making additional loans(a)

(a) Balance sheets are highly stylised for ease of exposition: the quantities of each type of money shown do not correspond to the quantities actually held on each sector's balance sheet.

(b) Central bank balance sheet only shows base money liabilities and the corresponding assets.

In practice the central bank holds other non–money liabilities. Its non–monetary assets are mostly made up of government debt. Although that government debt is actually

held by the Bank of England Asset Purchase Facility, so does not appear directly on the balance sheet.

(c) Commercial banks' balance sheets only show money assets and liabilities before any loans are made.

(d) Consumers represent the private sector of households and companies. Balance sheet only shows broad money assets and corresponding liabilities — real assets such as the house being transacted are not shown. Consumers' non–money liabilities include existing secured and unsecured loans. credit card will

</doc_id=10>

<doc_id=11>

[11] Two–sector model [[edit](/w/index.php?title=Circular_flow_of_income&action=edit§ion=8 "Edit section: Two–sector model")]

[![/upload.wikimedia.org/wikipedia/commons/thumb/c/c2/Two_sector_circular_flow_diagram.jpg/220px–Two_sector_circular_flow_diagram.jpg)](/wiki/File:Two_sector_circular_flow_diagram.jpg) Two–sector circular flow diagram

In the basic two–sector circular flow of income model, the economy consists of two [sectors](/wiki/Economic_sector "Economic sector") : (1) [households](/wiki/Households "Households") and (2) [firms](/wiki/Theory_of_the_firm "Theory of the firm") . (17) (18) (Some sources refer to households as "individuals" (19) or the "public" (20) and to firms as "businesses" (1) (2) or the "productive sector." (21)) The model assumes that there is no [financial sector](/wiki/Financial_sector "Financial sector") , no [government sector](/wiki/Government_sector "Government sector") , and no [foreign sector](/wiki/External_sector "External sector") . In addition, the model assumes that (a) through their expenditures, households spend all of their [income](/wiki/Income "Income") on [goods and services](/wiki/Goods_and_services "Goods and services") or [consumption](/wiki/Consumption_(economics) "Consumption \economics)") and (b) through their [expenditures](/wiki/Expenditure "Expenditure") , households purchase all output produced by firms. (18) This means that all household expenditures become income for firms. The firms then spend all of this income on [factors of production](/wiki/Factors_of_production "Factors of production") such as labor, capital and raw materials, "transferring" all of their income to the factor owners (which are households). The factor owners (households), in turn, spend all of their income on goods, which leads to a circular flow of income. (20) (18) [22

</doc_id=11>

<doc_id=12>

[12] key topics covered in this article.(2)

Two misconceptions about money creation

The vast majority of money held by the public takes the form of bank deposits. But where the stock of bank deposits comes from is often misunderstood. One common misconception is that banks act simply as intermediaries, lending out the deposits that savers place with them. In this view deposits are typically 'created' by the saving decisions of households, and banks then 'lend out' those existing deposits to borrowers, for example to companies looking to finance investment or individuals wanting to purchase houses. In fact, when households choose to save more money in bank accounts, those deposits come simply at the expense of deposits that would have otherwise gone to companies in payment for goods and services. Saving does not by itself increase the deposits or 'funds available' for banks to lend. Indeed, viewing banks simply as intermediaries ignores the fact that, in reality in the modern economy, commercial banks are the creators of deposit money. This article explains how, rather than banks lending out deposits that are placed with them, the act of lending creates deposits — the reverse of the sequence typically described in textbooks.(3) Another common misconception is that the central bank determines the quantity of loans and deposits in the economy by controlling the quantity of central bank money — the so–called 'money multiplier' approach. In that view, central banks implement monetary policy by choosing a quantity of reserves. And, because there is assumed to be a constant ratio of broad money to base money, these reserves are then 'multiplied up' to a much greater change in bank loans and deposits. For the theory to hold, the amount of reserves must be a binding constraint on lending, and the central bank must directly determine the amount of reserves. While the money multiplier theory can be a useful way of introducing money and banking in economic textbooks, it is not an accurate description of how money is created in reality. Rather than controlling the quantity of reserves, central banks today typically implement monetary policy by setting the price of reserves — that is, interest rates. In reality, neither are reserves a binding constraint on lending, nor does the central bank fix the amount of reserves that are available. As with the relationship between deposits and loans, the relationship between reserves and loans typically operates in the reverse way to that described in some

e

</doc_id=12>

<doc_id=13>

[13] 14 Quarterly Bulletin 2014 Q1

Money creation in the modern economy By Michael McLeay, Amar Radia and Ryland Thomas of the Bank's Monetary Analysis Directorate.(1) * This article explains how the majority of money in the modern economy is created by commercial banks making loans. * Money creation in practice differs from some popular misconceptions — banks do not act simply as intermediaries, lending out deposits that savers place with them, and nor do they 'multiply up' central bank money to create new loans and deposits. * The amount of money created in the economy ultimately depends on the monetary policy of the central bank. In normal times, this is carried out by setting interest rates. The central bank can also affect the amount of money directly through purchasing assets or 'quantitative easing'. Overview In the modern economy, most money takes the form of bank loan and stable inflation. In normal times, the Bank of deposits. But how those bank deposits are created is often England implements monetary policy by setting the interest misunderstood: the principal way is through commercial rate on central bank reserves. This then influences a range of banks making loans. Whenever a bank makes a loan, it interest rates in the economy , including those on bank loans. simultaneously creates a matching deposit in the In exceptional circumstances, when interest rates are at their borrower's bank account, thereby creating new money. effective lower bound, money creation and spending in the The reality of how money is created today differs from the economy may still be too low to be consistent with the description found in some economics textbooks: central bank's monetary policy objectives. One possible response is to undertake a series of asset purchases, or * Rather than banks receiving deposits when households 'quantitative easing' (QE). QE is intended to boost the save and then lending them out, bank lending creates amount of money in the economy directly by purchasing deposits. assets, mainly from non-bank financial companies. * In normal times, the central bank does not fix the amount of money in circulation, nor is central bank money QE initially increases the amount of bank deposits those 'multiplied up' into more loans and deposits. companies hold (in place of the assets they sell). Those companies will then wish to rebalance their portfolios of Although commercial banks create money through lending, assets by buying higher-yielding assets, raising the price of they cannot do so freely without limit </doc_id=13> <doc_id=14> [14] much central bank money banks want to hold in reserve (to meet withdrawals by the public, make payments to other banks, or meet regulatory liquidity requirements), which is then, in normal times, supplied on demand by the Bank of England. The rest of this article discusses these practices in more detail. Money creation in reality Lending creates deposits — broad money determination at the aggregate level As explained in 'Money in the modern economy: an introduction', broad money is a measure of the total amount of money held by households and companies in the economy. Broad money is made up of bank deposits — which are essentially IOUs from commercial banks to households and companies — and currency — mostly IOUs from the central bank.(4)(5) Of the two types of broad money, bank deposits make up the vast majority — 97% of the amount currently in circulation.(6) And in the modern economy, those bank deposits are mostly created by commercial banks themselves. (1) Throughout this article, 'banks' and 'commercial banks' are used to refer to banks and building societies together. (2) See www.youtube.com/watch?v=CvRAqR2pAgw. (3) There is a long literature that does recognise the 'endogenous' nature of money creation in practice. See, for example, Moore (1988), Howells (1995) and Palley (1996). (4) The definition of broad money used by the Bank of England, M4ex, also includes a wider range of bank liabilities than regular deposits; see Burgess and Janssen (2007) for more details. For simplicity, this article describes all of these liabilities as deposits. A box later in this article provides details about a range of popular monetary aggregates in the United Kingdom. (5) Around 6% of the currency in circulation is made up of coins, which are produced by The Royal Mint. Of the banknotes that circulate in the UK economy, some are issued	by some Scottish and Northern Irish commercial banks, although these are fully matched by Bank of England money held at the Bank. (6) As of December 2013. Before loans are made After loans are made Central bank(b) Assets Liabilities Assets Liabilities Reserves Currency Base money Non-money </doc_id=14> <doc_id=15> [15] Reserves Currency New deposits Deposits Assets Liabilities Reserves Currency Deposits Consumers(d) Assets Liabilities New deposits Deposits Currency New loans Non-money Assets Liabilities Broad money Non-money money Broad Deposits Currency 16 Quarterly Bulletin 2014 Q1 Commercial banks create money, in the form of bank deposits, by making new loans. When a bank makes a loan, for example to someone taking out a mortgage to buy a house, it does not typically do so by giving them thousands of pounds worth of banknotes. Instead, it credits their bank account with a bank deposit of the size of the mortgage. At that moment, new money is created. For this reason, some economists have referred to bank deposits as 'fountain pen money', created at the stroke of bankers' pens when they approve loans.(1) This process is illustrated in Figure 1, which shows how new lending affects the balance sheets of different sectors of the economy (similar balance sheet diagrams are introduced in 'Money in the modern economy: an introduction'). As shown in the third row of Figure 1, the new deposits increase the assets of the consumer (here taken to represent households and companies) — the extra red bars — and the new loan increases their liabilities — the extra white bars. New broad money has been created. Similarly, both sides of the commercial banking sector's balance sheet increase as new money and loans are created. It is important to note that although the simplified diagram of Figure 1 shows the amount of new money created as being identical to the amount of new lending, in practice there will be several factors that may subsequently cause the amount of deposits to be different from the amount of lending. These are discussed in detail in the next section. </doc_id=15> <doc_id=16> [16] First, central banks provide money for our everyday lives. People use Federal Reserve notes—paper money—to purchase goods and services. Banks use reserves—deposits they hold at the Fed—to settle payments and as a safe and liquid asset. The U.S. Treasury has a "checking account" at the Fed, called the Treasury General Account (TGA), which it uses to manage its cash flows. These and other Fed liabilities are the foundation for our monetary and banking system—the lifeblood of our economy. </doc_id=16> <doc_id=17> [17] with further analysis of policy options and practical implementation issues.33 The Federal Reserve is also an active participant in workstreams supporting the Financial Stability Board's roadmap for improving cross-border payments.34 In addition, the Federal Reserve Bank of New York has established an Innovation Center to facilitate collaboration with the Bank for International Settlements on a number of financial innovations. 32 Bank for International Settlements, Central Bank Digital Currencies: Foundational Principles and Core Features (Basel: BIS, October 2020), https://www.bis.org/publ/othp33.htm. 33 Bank for International Settlements, Central Banks and the BIS Explore What a Retail CBDC Might Look Like, press release, September 2021, https://www.bis.org/press/p210930.htm. 34 Financial Stability Board, Enhancing Cross-border Payments: Stage 3 Roadmap (Washington: FSB, October 2020), https://www.fsb.org/2020/10/enhancing-cross-border-payments-stage-3-roadmap/. 24 Money and Payments: The U.S. Dollar in the Age of Digital Transformation Appendix B: Types of Money Central Bank Money In the United States, central bank money is a liability of the Federal Reserve and comes in two
--	--

forms: (1) physical currency circulating among the general public (with some held by banks) and (2) digital account balances held by banks and a limited number of other eligible institutions at the Federal Reserve.³⁵ Physical currency has legal tender status in the United States. As of November 2021, there was around \$6.39 trillion of central bank money. Of this, \$4.18 trillion was in the form of balances at the Federal Reserve and \$2.21 trillion was in the form of circulating cash.³⁶ Central bank money is the most trusted and safest form of money because the Federal Reserve, as a central bank issuer, presents no credit risk and no liquidity risk. As the safest form of money, central bank money plays an important role as a settlement asset for critically important interbank payment systems.³⁷ Commercial Bank Money Dollar-denominated balances held by consumers in commercial bank accounts are "commercial bank money." These balances are liabilities of commercial banks, not of the Federal Reserve.³⁸ As of November 2021,

</doc_id=17>
<doc_id=18>
[18] When a decision is made to issue tax rebates, the Department of the Treasury is responsible for taking money out of the Federal Reserve and putting it into the hands of consumers. Consumers, in turn, spend the money. Through their spending, they funnel money into the economy, resulting in increased sales of consumer goods and more jobs to produce and distribute those goods.

</doc_id=18>
<doc_id=19>
[19] In the financial sector [[edit](/w/index.php?title=Circular_flow_of_income&action=edit§ion=15 "Edit section: In the financial sector")]

In terms of the circular flow of income model, the leakage that financial institutions provide in the economy is the option for households to save their money. This is a leakage because the saved money cannot be spent in the economy and thus is an idle asset that means not all output will be purchased. The injection that the financial sector provides into the economy is investment (I) into the business/firms sector. An example of a group in the finance sector includes banks such as [Westpac](/wiki/Westpac "Westpac") or financial institutions such as [Suncorp](/wiki/Suncorp_Bank "Suncorp Bank")

</doc_id=19>
<doc_id=20>
[20] Summary of leakages and injections [[edit](/w/index.php?title=Circular_flow_of_income&action=edit§ion=18 "Edit section: Summary of leakages and injections")]

All leakages and injections in the five-sector model

Leakages	↓	↓	↓	↓	↓
Financial	↓	Saving (S)	↓	Investment (I)	↓
Government	↓	Taxes (T)	↓	Government spending (G)	↓
Overseas	↓	Imports (M)	↓	Exports (X)	↓

</doc_id=20>
</retrieved_passages>

SFT+GRPO evaluated output:

```
<answer>
[1] > [3] > [10] > [13] > [15] > [14] > [9] > [4] > [8] > [2] > [5] > [19] > [20] > [11]
> [17] > [12] > [6]
> [16] > [18]
</answer>
```

SARD. SARD selected input window:

<retrieved_passages>
<doc_id=1>
[1] Overview [[edit](/w/index.php?title=Circular_flow_of_income&action=edit§ion=1 "Edit section: Overview")]

[

](/wiki/File:Circular_flow_of_income_and_expenditure.jpg) Model of the circular flow of income and expenditure

The circular flow of income is a concept for better understanding of the [economy](/wiki/Economy "Economy") as a whole and for example the [National Income and Product Accounts](/wiki/National_Income_and_Product_Accounts "National Income and Product Accounts") (NIPAs). In its most basic form it considers a simple economy consisting solely of [businesses](/wiki/Business "Business") and individuals, and can be represented in a so-called "circular flow diagram." In this simple economy, individuals provide the labour that enables businesses to produce goods and services. These activities are represented by the green lines in the diagram. (5)

Alternatively, one can think of these transactions in terms of the monetary

flows that occur. Businesses provide individuals with income (in the form of compensation) in exchange for their labor. That income is spent on the goods and services businesses produce. These activities are represented by the blue lines in the diagram above. (5)

The circular flow diagram illustrates the interdependence of the "flows," or activities, that occur in the economy, such as the [production](/wiki/Production_(economics) "Production (economics)") of goods and services (or the "output" of the economy) and the income generated from that production. The circular flow also illustrates the equality between the income earned from production and the value of goods and services produced. (5)

Of course, the total economy is much more complicated than the illustration above. An economy involves interactions between not only individuals and businesses, but also Federal, state, and local governments and residents of the rest of the world. Also not shown in this simple illustration of the economy are other aspects of economic activity such as investment in [capital](/wiki/Capital_(economics) "Capital (economics)") (produced—or fixed—

</doc_id=1>
<doc_id=2>
[2] Five-sector model [[edit](/w/index.php?title=Circular_flow_of_income&action=edit§ion=11 "Edit section: Five-sector model")]

The five-sector model adds the financial sector to the four-sector model. (19) Thus, the five-sector model includes (1) households, (2) firms, (3) government, (4) the rest of the world, and (5) the financial sector. The financial sector includes banks and non-bank intermediaries that engage in borrowing (savings from households) and lending (investments in firms). (19) Money facilitates such an exchange smoothly. Residuals from each market enter the [capital market](/wiki/Capital_market "Capital market") as savings, which in turn are invested in firms and the government sector. Technically speaking, so long as lending is equal to borrowing (i.e., leakages are equal to injections), the circular flow will continue indefinitely. However, this job is done by financial institutions in the economy.

* [
](/wiki/File:Five_Sector_Circular_Flow_of_Income_Model.jpg "Five-sector circular flow of income model")

Five-sector circular flow of income model

* [
](/wiki/File:04_circular_flow_diagram.png "Circular flow diagram, five-sectors model")

Circular flow diagram, five-sectors model

</doc_id=2>
<doc_id=3>
[3] Three-sector model [[edit](/w/index.php?title=Circular_flow_of_income&action=edit§ion=9 "Edit section: Three-sector model")]

[

](/wiki/File:Economics_circular_flow_cartoon.jpg) Three-sector circular flow diagram

The three-sector model adds the government sector to the two-sector model. (17) (18) Thus, the three-sector model includes (1) households, (2) firms, and (3) government. It excludes the financial sector and the foreign sector. The government sector consists of the economic activities of local, state and federal governments. Flows from households and firms to government are in the form of taxes. The income the government receives flows to firms and households in the form of subsidies, transfers, and purchases of goods and services. (17) (18) Every payment has a corresponding receipt; that is, every flow of money has a corresponding flow of goods in the opposite direction. (17) (18) As a result, the [aggregate expenditure](/w/index.php?title=Aggregate_expenditure&action=edit&redlink=1 "Aggregate expenditure (page does not exist)") of the economy is identical to its [aggregate income](/wiki/Aggregate_income "Aggregate income"), making a circular flow.

</doc_id=3>
<doc_id=4>
[4] Four-sector model [[edit](/w/index.php?title=Circular_flow_of_income&action=edit§ion=10 "Edit section: Four-sector model")]

The four-sector model adds the foreign sector to the three-sector model. (17) (18) (23) (The foreign sector is also known as the "external sector," the "overseas sector," (19) or the "rest of the world.") Thus, the four-sector model includes (1) households, (2) firms, (3) government, and (4) the rest of the world. It excludes the financial sector. The foreign sector comprises (a) foreign trade (imports and exports of goods and services) and (b) inflow and outflow of capital (foreign exchange). (18) Again, each flow of money has a corresponding flow of goods (or services) in the opposite direction. (18) Each of the four sectors receives some payments from the other in lieu of goods and services which makes a regular flow of goods and physical services. The addition of the foreign sector

</doc_id=4>

<doc_id=5>
 [5] Two–sector model [[edit
](/w/index.php?title=Circular_flow_of_income&action=edit§ion=8 "Edit section: Two–sector model")]

[
 !]([/upload.wikimedia.org/wikipedia/commons/thumb/c/c2/
 Two_sector_circular_flow_diagram.jpg/220px–
 Two_sector_circular_flow_diagram.jpg)
](/wiki/File:Two_sector_circular_flow_diagram.jpg) Two–sector circular flow
 diagram

In the basic two–sector circular flow of income model, the economy consists of
 two [sectors](/wiki/Economic_sector "Economic sector") : (1) [households
](/wiki/Households "Households") and (2) [firms](/wiki/Theory_of_the_firm
 "Theory of the firm") . (17) (18) (Some sources refer to households as
 "individuals" (19) or the "public" (20) and to firms as "businesses" (1)
 (2) or the "productive sector." (21)) The model assumes that there is no [
 financial sector](/wiki/Financial_sector "Financial sector") , no [
 government sector](/wiki/Government_sector "Government sector") , and no [
 foreign sector](/wiki/External_sector "External sector") . In addition, the
 model assumes that (a) through their expenditures, households spend all of
 their [income](/wiki/Income "Income") on [goods and services
](/wiki/Goods_and_services "Goods and services") or [consumption
](/wiki/Consumption_(economics) "Consumption (economics)") and (b) through
 their [expenditures](/wiki/Expenditure "Expenditure") , households purchase
 all output produced by firms. (18) This means that all household
 expenditures become income for firms. The firms then spend all of this income
 on [factors of production](/wiki/Factors_of_production "Factors of
 production") such as labor, capital and raw materials, "transferring" all of
 their income to the factor owners (which are households). The factor owners
 (households), in turn, spend all of their income on goods, which leads to a
 circular flow of income. (20) (18) [22
 </doc_id=5>
 <doc_id=6>
 [6] Leakages and injection [[edit
](/w/index.php?title=Circular_flow_of_income&action=edit§ion=14 "Edit
 section: Leakages and injection")]

The five–sector model considers leakages and injections.

* 'Leakage' means withdrawal from the flow. When households and firms save part
 of their incomes it
 constitutes leakage. They may be in form of savings, tax payments, and [imports](/
 wiki/Imports "Imports")
 . Leakages reduce the flow of income.
 * 'Injection' means the introduction of income into the flow. When households and
 firms borrow savings, they
 constitute injections. Injections increase the flow of income. Injections can take the
 forms of investment,
 government spending and [exports](/wiki/Exports "Exports") . As long as leakages
 are equal to injections,
 the circular flow of income continues indefinitely. Financial institutions or capital
 market play the role
 of intermediaries. This means that income individuals receive from businesses and
 the goods and services
 that are sold to them do not count as injections or leakages, as no new money is
 being introduced to the
 flow and no money is being taken out of the flow.

Leakages and injections can occur in the financial sector, government sector
 and overseas sector:
 </doc_id=6>
 <doc_id=7>
 [7] In the government sector [[edit
](/w/index.php?title=Circular_flow_of_income&action=edit§ion=16 "Edit
 section: In the government sector")]

The leakage that the government sector provides is through the collection of
 revenue through taxes (T) that is provided by households and firms to the
 government. This is a leakage because it is a leakage out of the current
 income, thus reducing the expenditure on current goods and services. The
 injection provided by the government sector is government spending (G) that
 provides collective services and welfare payments to the community. An example
 of a tax collected by the government as a leakage is [income tax
](/wiki/Income_tax "Income tax") , and an injection into the economy can be
 when the government redistributes this income in the form of [welfare
 payments](/wiki/Social_security "Social security") , that is a form of
 government spending back into the economy.
 </doc_id=7>
 <doc_id=8>
 [8] Summary of leakages and injections [[edit
](/w/index.php?title=Circular_flow_of_income&action=edit§ion=18 "Edit
 section: Summary of leakages and injections")]

All leakages and injections in the five–sector model Sector | Leakages |
 Injections
 ---|---|---
 Financial | Saving (S) | Investment (I)
 Government | Taxes (T) | Government spending (G)
 Overseas | Imports (M) | Exports (X)
 </doc_id=8>
 <doc_id=9>
 [9] !]([/upload.wikimedia.org/wikipedia/commons/thumb/b/b8/
 Circular_flow_of_goods_income.png/360px–
 Circular_flow_of_goods_income.png)
](/wiki/File:Circular_flow_of_goods_income.png) Basic diagram of the circular
 flow of income. The functioning of the free–market economic system is

represented with [firms](/wiki/Firm "Firm") and [households
](/wiki/Household "Household") and interaction back and forth. (2)

The **circular flow of income** or **circular flow** is a [model
](/wiki/Economic_model "Economic model") of the [economy](/wiki/Economy
 "Economy") in which the major exchanges are represented as flows of [money
](/wiki/Money "Money") , [goods](/wiki/Good_(economics) "Good
 (economics)") and [services](/wiki/Service_(economics) "Service
 (economics)") , etc. between [economic agents](/wiki/Economic_agents
 "Economic agents") . The flows of money and goods exchanged in a closed
 circuit correspond in value, but run in the opposite direction. The circular
 flow analysis is the basis of [national accounts](/wiki/National_accounts
 "National accounts") and hence of [macroeconomics](/wiki/Macroeconomics
 "Macroeconomics") .

The idea of the circular flow was already present in the work of [Richard
 Cantillon](/wiki/Richard_Cantillon "Richard Cantillon") . (3) [Francois
 Quesnay](/wiki/Fran%C3%A7ois_Quesnay "Francois Quesnay") developed and
 visualized this concept in the so–called [Tableau economique
](/wiki/Tableau_%C3%A9conomique "Tableau economique") . (4) Important
 developments of Quesnay's tableau were [Karl Marx](/wiki/Karl_Marx "Karl
 Marx") 's reproduction schemes in the second volume of _[Capital: Critique of
 Political Economy](/wiki/Capital:_Critique_of_Political_Economy "Capital:
 Critique of Political Economy") _ , and [John Maynard Keynes
](/wiki/John_Maynard_Keynes "John Maynard Keynes") ' _[General Theory of
 Employment, Interest and
 </doc_id=9>
 <doc_id=10>
 [10] Further developments [[edit
](/w/index.php?title=Circular_flow_of_income&action=edit§ion=6 "Edit
 section: Further developments")]

[
 !]([/upload.wikimedia.org/wikipedia/commons/thumb/b/b4/
 The_competitive_price_system_adapted_from_Samuelson%2C_1961.jpg
 /220px–
 The_competitive_price_system_adapted_from_Samuelson%2C_1961.jpg)
](/wiki/File:The_competitive_price_system_adapted_from_Samuelson,_1961.jpg)
 The competitive price system adapted from [Samuelson](/wiki/Paul_Samuelson
 "Paul Samuelson") , 1961

An important development was [John Maynard Keynes](/wiki/
 John_Maynard_Keynes
 "John Maynard Keynes") 's 1933 publication of the _[General Theory of
 Employment, Interest and Money
](/wiki/General_Theory_of_Employment,_Interest_and_Money "General Theory of
 Employment, Interest and Money") _ . Keynes' assistant [Richard Stone
](/wiki/Richard_Stone "Richard Stone") further developed the concept for the [
 United Nations](/wiki/United_Nations "United Nations") (UN) and the [
 Organisation for Economic Co–operation and Development
](/wiki/Organisation_for_Economic_Co-operation_and_Development "
 Organisation
 for Economic Co–operation and Development") to the systems, which is now used
 internationally.

The first to visualize the modern circular flow of income model was [Frank
 Knight](/wiki/Frank_Knight "Frank Knight") in 1933 publication of _The
 Economic Organization._ (13) Knight (1933) explained:

_[we may view the] economic organization as a system of prize relations. Seen in
 the large, free enterprise
 is an organization of production and distribution in which individuals or family
 units get their real
 income, their "living," by selling productive power for money to "business units" or
 "enterprises", and
 buying with the money income thus obtained the direct goods and services which
 they consume. This view, it
 will be remembered, ignores for the sake of simplicity the fact that an appreciable
 fraction of the
 productive power in use at any time is not really employed in satisfying current
 wants but to make provision
 for increased want–satisfaction in
 </doc_id=10>
 <doc_id=11>
 [11] Significance of study of circular flow of income [[edit
](/w/index.php?title=Circular_flow_of_income&action=edit§ion=20 "Edit
 section: Significance of study of circular flow of income")]

The circular flow of income is significant in four areas: (24)

- [Measurement of national income](/wiki/
 Measures_of_national_income_and_output "Measures of national
 income and output")
- Knowledge of Interdependence – Circular flow of income signifies the
 interdependence of each of activity
 upon one another. If there is no consumption, there will be no demand and
 expenditure which in fact
 restricts the amount of production and income.
- Unending Nature of Economic Activities – It signifies that production, income
 and expenditure are of
 unending nature, therefore, economic activities in an economy can never come to a
 halt. National income is
 also bound to rise in future.
- Injections and Leakages
 </doc_id=11>
 </retrieved_passages>

SARD evaluated output:

```
<answer>
[2] > [1] > [6] > [7] > [3] > [4] > [8] > [5] > [9]
</answer>
```

Analysis. The query mentions banks, the central bank, borrowing, taxes, and government spending, so passages about money creation or banking are genuinely relevant. The issue is that the user asks for a concrete diagrammatic framing of the *life of a dollar* through economic sectors. SFT therefore remains topically plausible, but its highest-ranked evidence focuses on bank money creation and balance-sheet mechanics rather than the circular-flow model requested by the query. SFT+GRPO identifies circular-flow material earlier, but still mixes in several adjacent money-creation passages high in the order. SARD more cleanly prioritizes the five-sector circular-flow model, the overview diagram, leakages/injections, and government-sector flows, which together form the answer-bearing structure for tracing money across households, firms, financial institutions, and government.

C.4 Training-Time Rescue Examples

The following examples come from the SARD training-time routing pipeline. Unlike the previous ablation case studies, these examples compare successive rescue stages for the same training instance: base-prompt exploration, guided exploration, and, when needed, teacher-guided repair. They are selected to show cases where the output becomes qualitatively more aligned with the labeled evidence after guidance or repair.

C.4.1 Training Case 1: Guidance rescue on ssDNA translation

Route / dataset / query id. guidance_rescue / biology / biology__01773.

Qualitative focus. The base rollout ranks a passage about ribosome translation through mRNA duplex regions above the labeled nucleic-acid chemistry evidence. The guided rollout moves the structurally relevant nucleic-acid passages to the top.

Best NDCG@10. Base prompt: 0.604; guided exploration: 1.000. The corresponding evaluated outputs are shown below.

Top labeled positive documents:

```
- [6] rel=2: [6] Biomolecules possess unique chemical properties that are essential
for their function. Yet, the overabundance of reactive species in both the
environment and within living cells constantly threatens these properties ( 1
, 2 ). Unwanted modifications to the chemical structure ...
```

```
- [7] rel=1: [7] Similar to many other biological molecules, RNA is vulnerable to
chemical insults from endogenous and exogenous sources. Noxious agents such
as reactive oxygen species or alkylating chemicals have the potential to
profoundly affect the chemical properties and hence the functi...
```

Full guided-exploration input window:

```
<user_query>
Can ribosomes read ssDNA? My question is whether translation can be done, either
naturally or artificially, through a ribosome reading (single-stranded) DNA
directly. If not, I would like to know what allows ssRNA to be translated
but not ssDNA..
</user_query>

<retrieved_passages>
<doc_id=1>
[1] The Shine-Dalgarno ( SD ) sequence is, sometimes partially, part of a
ribosomal binding site in bacterial and archaeal messenger RNA . [ 1 ] It is
generally located around 8 bases upstream of the start codon AUG. [ 2 ] The
RNA sequence helps recruit the ribosome to the messenger RNA (mRNA)
to initiate protein synthesis by aligning the ribosome with the start codon.
Once recruited, tRNA may add amino acids in sequence as dictated by the
codons, moving downstream from the translational start site.
</doc_id=1>
<doc_id=2>
[2] 16S ribosomal RNA (rRNA) (that is, the end where translation begins) is
pyrimidine-rich and has the specific sequence Y ACCUCCU UA . They
proposed that these ribosomal nucleotides recognize the complementary
purine-rich sequence AGGAGGU , which is found upstream of the start
codon AUG in a number of mRNAs found in viruses that affect E. coli . [ 2 ]
Many studies have confirmed that base pairing between the Shine-
Dalgarno sequence in mRNA and the 3' end of 16S rRNA is of prime
importance for initiation of translation by bacterial ribosomes. [ 6 ] [ 7 ]
</doc_id=2>
<doc_id=3>
[3] translation in prokaryotes. This change is due to a reduced or increased mRNA-
ribosome pairing efficiency, as evidenced by the fact that compensatory
mutations in the 3'-terminal 16S rRNA sequence can restore translation.
</doc_id=3>
<doc_id=4>
[4] In 1973 Dalgarno and Shine proposed that in eukaryotes , the 3'-end of the
small 18S rRNA may play a role in the termination of protein synthesis by
complementary base pairing with termination codons. [ 9 ] This came from
their observation that the 3' terminal sequences of 18S rRNA from
Drosophila melanogaster , Saccharomyces cerevisiae , and rabbit cells are
identical: GAUCAUUA -3'OH. [ 10 ] The conservation of this sequence
between such distantly related eukaryotes implied that this nucleotide tract
played an important role in the cell. Since this conserved sequence
contained the complement of each of the three eukaryotic termination
codons (UAA, UAG and UGA) it was proposed to have a role in the
termination of protein synthesis in eukaryotes. A similar role for the 3' end
of 16S rRNA in recognising termination triplets in E.coli was proposed in
1974 by Shine and Dalgarno on the basis of complementarity relationships
between the 3'-terminal UUA-OH in 16S rRNA and E.coli termination
codons. [ citation needed ] In F1 phage , a class of viruses that infect
bacteria, the sequence coding for the first few amino acids often contains
termination triplets in the two unused reading frames. [ further explanation
needed ] [ 11 ] In a commentary on this paper, it was noted that
complementary base pairing with the 3'-terminus of 16S rRNA might serve
to abort peptide bond formation after out-of-phase initiation. [ 12 ]
</doc_id=4>
<doc_id=5>
[5] Given the complementary relationship between rRNA and the Shine-Dalgarno
sequence in mRNA, it was proposed that the sequence at the 3'-end of the
rRNA determines the capacity of the prokaryotic ribosome to translate a
particular gene in an mRNA. [ 8 ] Base pairing between the 3'-end of the
rRNA and the Shine-Dalgarno sequence in mRNA is a mechanism by
which the cell can distinguish between initiator AUGs and internal and/or
out-of-frame AUG sequences. The degree of base pairing also plays a role
in determining the rate of initiation at different AUG initiator codons.
</doc_id=5>
<doc_id=6>
[6] Biomolecules possess unique chemical properties that are essential for their
function. Yet, the overabundance of reactive species in both the environment
and within living cells constantly threatens these properties ( 1 , 2 ).
Unwanted modifications to the chemical structure of nucleic acids, proteins,
lipids, or carbohydrates often result in drastic effects on the ability of these
molecules to carry out these functions. Changes to proteins, for example,
are known to cause misfolding and aggregation, and the inability of cells to
degrade these aberrant protein products has been associated with disease
states ( 3 , 4 ). Similarly, lipid peroxidation contributes to the pathogenicity
of several diseases ( 5 ). Because of the reactivity of the oxygen and
nitrogen atoms on the nucleobase, nucleic acids are especially susceptible to
certain types of chemical damage from sources such as reactive oxygen
species (ROS), 2 UV light (UV), and alkylating agents ( Fig. 1 A ) ( 1 ). The
oxygen atoms of the ribose and the phosphodiester backbone are also
vulnerable to chemical damage ( Fig. 1 B ). In the case of DNA, and most
RNA species, small changes have the ability to severely affect their
structure and hence their function ( 6 ). This is because, by-and-large,
nucleic acids require strict Watson-Crick base pairs between the
nucleobases.
</doc_id=6>
<doc_id=7>
[7] Similar to many other biological molecules, RNA is vulnerable to chemical
insults from endogenous and exogenous sources. Noxious agents such as
reactive oxygen species or alkylating chemicals have the potential to
profoundly affect the chemical properties and hence the function of RNA
molecules in the cell. Given the central role of RNA in many fundamental
```

biological processes, including translation and splicing, changes to its chemical composition can have a detrimental impact on cellular fitness, with some evidence suggesting that RNA damage has roles in diseases such as neurodegenerative disorders. We are only just beginning to learn about how cells cope with RNA damage, with recent studies revealing the existence of quality-control processes that are capable of recognizing and degrading or repairing damaged RNA. Here, we begin by reviewing the most abundant types of chemical damage to RNA, including oxidation and alkylation. Focusing on mRNA damage, we then discuss how alterations to this species of RNA affect its function and how cells respond to these challenges to maintain proteostasis. Finally, we briefly discuss how chemical damage to noncoding RNAs such as rRNA, tRNA, small nuclear RNA, and small nucleolar RNA is likely to affect their function.

</doc_id=7>
<doc_id=8>
[8] A ribosome is a macromolecular machine that is responsible for translating the genetic codes in messenger RNA (mRNA) into polypeptide chains. It has been determined that besides translating through the single-stranded region, the ribosome can also translate through the duplex region of mRNA by unwinding the duplex. To understand the mechanism of ribosome translation through the duplex, several models have been proposed to study the dynamics of mRNA unwinding. Here, we present a comprehensive review of these models and also discuss other possible models. We evaluate each model and discuss the consistency and/or inconsistency between the theoretical results that are obtained based on each model and the available experimental data, thus determining which model is the most reasonable one to describe the mRNA unwinding mechanism and dynamics of the ribosome. Moreover, a framework for future studies in this subject is provided.

</doc_id=8>
<doc_id=9>
[9] Using a method developed by Hunt, [4] [5] Shine and Dalgarno showed that the nucleotide tract at the 3' end of E. coli

</doc_id=9>
<doc_id=10>
[10] The Shine–Dalgarno sequence is common in bacteria , but rarer in archaea . [3] It is also present in some chloroplast and mitochondrial transcripts. The six–base consensus sequence is AGGAGG ; in Escherichia coli , for example, the sequence is AGGAGGU, while the shorter GAGG dominates in E. coli virus T4

</doc_id=10>
<doc_id=11>
[11] Malys N (2012). "Shine–Dalgarno sequence of bacteriophage T4: GAGG prevails in early genes". Molecular Biology Reports . 39 (1): 33– 9. doi : 10.1007/s11033-011-0707-4 . PMID

</doc_id=11>
<doc_id=12>
[12] Chen, Hongyun; Bjerknes, Matthew; Kumar, Ravindra; Jay, Ernest (1994). "Determination of the optimal aligned spacing between the Shine – Dalgarno sequence and the translation initiation codon of Escherichia coli m RNAs". Nucleic Acids Research . 22 (23): 4953– 4957. doi : 10.1093/nar/22.23.4953 . ISSN

</doc_id=12>
<doc_id=13>
[13] Arabic Belarusian Catalan Danish German Spanish French Galician Korean Italian Hebrew Hungarian Dutch Japanese Polish Portuguese Russian Thai Turkish Ukrainian Vietnamese Chinese

</doc_id=13>
</retrieved_passages>

Perform the analysis and ranking. Ensure all XML tags are closed.

Base-prompt best output:

<answer>[8] > [6] > [2] > [10] > [5] > [4] > [1] > [3] > [7]</answer>

Strategy-guided best output:

<answer>[6] > [7] > [1] > [2] > [3] > [5]</answer>

Analysis. The base-prompt rollout ranks passage [8] first because it is strongly topical to ribosome translation, but it mainly discusses mRNA translation through duplex regions. The labeled positives are [6] and [7], which discuss chemical and structural properties of nucleic acids and RNA. The guidance redirects the model from standard translation procedure toward the structural/chemical constraint needed by the query, yielding a ranking that places [6] and [7] first.

C.4.2 Training Case 2: Teacher-guided repair on Mendelian height

Route / dataset / query id. repair_rescue / biology / biology__01976.

Qualitative focus. The base and initially guided rollouts rank general incomplete-dominance or blending evidence above the passage that directly states the pea-plant height outcome. Teacher-guided repair shifts the guidance toward trait-specific evidence.

Best NDCG@10. Base prompt: 0.681; guided exploration: 0.736; teacher-guided repair: 1.000. The corresponding evaluated outputs are shown below.

Top labeled positive documents:

- [12] rel=3: [12] Often expression levels in an individual can only reach a certain intensity regardless of whether the individual is a homozygote or heterozygote. For example, pea plants heterozygote for the tall/dwarf allelic pair are the same size as the homozygous tall parent. But expression is not considere...
- [2] rel=2: [2] Genetics is the study of how you became the way you are. Everyone inherits traits, or qualities, from their parents, but before Gregor Mendel, no one was quite sure how that happened. Even smart scientists like Charles Darwin had it all wrong. They thought traits were blended like paint. They be...
- [7] rel=1: [7] Incomplete dominance – the F 1 produces a phenotype quantitatively intermediate between the two homozygous parents; if the product is exactly intermediate between the two homozygous parents the relationship is termed no dominance (although some have tried to substitute the term no dominance for ...

Full repair input window:

<user_query>
Can there be medium height(neither tall nor short) pea plants in Mendel's experiment? Can there be medium height(neither tall nor short) pea plants in Mendel's experiment?
All textbooks I have read seem to imply that pea plants have to be either tall or short, nothing in between..
</user_query>

<retrieved_passages>
<doc_id=1>
[1] That brings us to a question: when someone says two people are as alike as peas in a pod, what does that truly mean? How alike are two peas in a pod? Mendel found out they can be very different. He identified eight different traits that followed the pattern he devised, although it turns out there are many more.

</doc_id=1>
<doc_id=2>
[2] Genetics is the study of how you became the way you are. Everyone inherits traits, or qualities, from their parents, but before Gregor Mendel, no one was quite sure how that happened. Even smart scientists like Charles Darwin had it all wrong. They thought traits were blended like paint. They believed that if a tall person and a short person had a baby, the baby would be of medium height.

</doc_id=2>
<doc_id=3>
[3] Abstract A controversy arose over Mendel's pea crossing experiments after the statistician R.A. Fisher proposed how these may have been performed and criticised Mendel's interpretation of his data. Here we re-examine Mendel's experiments and investigate Fisher's statistical criticisms of bias. We describe pea varieties available in Mendel's time and show that these could readily provide all the material Mendel needed for his experiments; the characters he chose to follow were clearly described in catalogues at the time. The combination of character states available in these varieties, together with Eichling's report of crosses Mendel performed, suggest that two of his F3 progeny test experiments may have involved the same F2 population, and therefore that these data should not be treated as independent variables in statistical analysis of Mendel's data. A comprehensive re-examination of Mendel's segregation ratios does not support previous suggestions that they differ remarkably from expectation. The χ^2 values for his segregation ratios sum to a value close to the expectation and there is no deficiency of extreme segregation ratios. Overall the χ values for Mendel's segregation ratios deviate slightly from the standard normal distribution; this is probably because of the variance associated with phenotypic rather than genotypic ratios and because Mendel excluded some data sets with small numbers of progeny, where he noted the ratios "deviate not insignificantly" from expectation.

```

</doc_id=3>
<doc_id=4>
[4] As an example, let assume that a gene of interest resides on a DNA fragment
that is 3.0 kb in size in parent one and 2.0 kb in size in a second parent. (See
figure below.) When we cross the two parents, one chromosome carrying the
gene comes from each parent. Since our technique recognizes both
copies in the parents, the parental signal will be twice as strong as the F 1
signal that contains one chromosome, and thus one copy of the particular
DNA size, from each parent. The F 2 generation will segregate for the three
different genotypes in a 1:2:1 ratio.
</doc_id=4>
<doc_id=5>
[5] The true test of any theory of science is its ability to explain results that at first
glance appear to be a clear exception to the theory. But, if the exception can
be explained by the theory, then the theory is further validated. One such
genetic example that challenged Mendel's first law was the relationship
between two alleles that do not express a typical dominance/recessive
relationship. That is, the F 1 does not exhibit the genotype of one of the two
pure line parents. This type of allelic relationship was termed codominance.
</doc_id=5>
<doc_id=6>
[6] These purplish-flowered plants are not just pretty to look at. Plants like these
led to a huge leap forward in biology. The plants are common garden peas,
and they were studied in the mid-1800s by an Austrian monk named Gregor
Mendel. With his careful experiments, Mendel uncovered the secrets of
heredity, or how parents pass characteristics to their offspring. You may not
care much about heredity in pea plants, but you probably care about your
own heredity. Mendel's discoveries apply to people as well as to peas —
and to all other living things that reproduce sexually. In this concept, you
will read about Mendel's experiments and the secrets of heredity that he
discovered.
</doc_id=6>
<doc_id=7>
[7] Incomplete dominance – the F 1 produces a phenotype quantitatively
intermediate between the two homozygous parents; if the product is exactly
intermediate between the two homozygous parents the relationship is
termed no dominance (although some have tried to substitute the term no
dominance for codominance, it has not been widely accepted)
</doc_id=7>
<doc_id=8>
[8] S1 Ep10 What Do Pea Plants Have To Do With Your Eye Color? Premiere:
8/3/2023 | 00:13:25 | NR | All of the different plants on Earth—from mango
trees to marigolds—have come about thanks to the simple rules of genetic
inheritance, which determine how traits are passed on from one generation
to the next. In this episode of Crash Course Botany, we'll explore the
remarkable story of how a botanist and mathematician named Gregor
Mendel laid the groundwork for modern genetics.
</doc_id=8>
<doc_id=9>
[9] About the Episode All of the different plants on Earth—from mango trees to
marigolds—have come about thanks to the simple rules of genetic
inheritance, which determine how traits are passed on from one generation
to the next. In this episode of Crash Course Botany, we'll explore the
remarkable story of how a botanist and mathematician named Gregor
Mendel laid the groundwork for modern genetics.
</doc_id=9>
<doc_id=10>
[10] 16: Inheritance and Biotechnology Human Biology (Gabor Gyurkovics) { } {
"16.01:_Case_Study_-_Genes_and_Inheritance" : "property get [Map
MindTouch.Deki.Logic.ExtensionProcessorQueryProvider+<
c__DisplayClass230_0.<PageSubPageProperty>b__1]()", "16.02:_Mendel's
_Experiments_and_Laws_of_Inheritance" : "property get [Map
MindTouch.Deki.Logic.ExtensionProcessorQueryProvider+<
c__DisplayClass230_0.<PageSubPageProperty>b__1]()", "16.03:
_Genetics_of_Inheritance" : "property get [Map MindTouch.Deki.Logic.
ExtensionProcessorQueryProvider+<
c__DisplayClass230_0.<
PageSubPageProperty>b__1]()", "16.04:_Mendelian_Inheritance" : "
property get [Map MindTouch.Deki.Logic.
ExtensionProcessorQueryProvider+<
c__DisplayClass230_0.<
PageSubPageProperty>b__1]()", "16.05:_Non-Mendelian_Inheritance" : "
property get [Map MindTouch.Deki.Logic.
ExtensionProcessorQueryProvider+<
c__DisplayClass230_0.<
PageSubPageProperty>b__1]()", "16.06:_Genetic_Disorders" : "property
get [Map MindTouch.Deki.Logic.ExtensionProcessorQueryProvider+<
c__DisplayClass230_0.<PageSubPageProperty>b__1]()", "16.07:
_Biotechnology" : "property get [Map MindTouch.Deki.Logic.
ExtensionProcessorQueryProvider+<
c__DisplayClass230_0.<
PageSubPageProperty>b__1]()", "16.08:
_Case_Study_Cancer_Conclusion_and_Chapter_Summary" : "property get
[Map MindTouch.Deki.Logic.ExtensionProcessorQueryProvider+<
c__DisplayClass230_0.<PageSubPageProperty>b__1]()" } { "00:
_Front_Matter" : "property get [Map MindTouch.Deki.Logic.
ExtensionProcessorQueryProvider+<
c__DisplayClass230_0.<
PageSubPageProperty>b__1]()", "01:
_Introduction_to_Human_Biology_and_the_Scientific_Method" : "property
get [Map MindTouch.Deki.Logic.ExtensionProcessorQueryProvider+<
c__
</doc_id=10>
<doc_id=11>
[11] F 1 : We would expect red or white flowers in this generation, depending upon
which allele is dominant. But, the F 1
</doc_id=11>
<doc_id=12>
[12] Often expression levels in an individual can only reach a certain intensity
regardless of whether the individual is a homozygote or heterozygote. For
example, pea plants heterozygote for the tall/dwarf allelic pair are the same
size as the homozygous tall parent. But expression is not considered with
DNA markers because we are monitoring the presence or absence of a

```

```

specific DNA fragment. Therefore, DNA fragments are a true example of
codominance where each allele is equally expressed in the F 1
</doc_id=12>
<doc_id=13>
[13] Biochemical phenotype – a phenotype that is revealed by biochemical
experimentation; examples are DNA markers (RFLPs); protein-size
markers (isozymes); quantity of a metabolite; immunological reaction
</doc_id=13>
<doc_id=14>
[14] Genotypic designations could be given for each of the alleles. For example, if
the 3.0 kb fragment is designated as allele A1, the genotype of parent 1 will
be A1A1. Allele designation A2 will be used for the 2.0 kb fragment and
the genotype of parent 2 will be A2A2. Because the F 1 is heterozygous, its
genotype will be A1A2. Finally, the genotypes of the F 2
</doc_id=14>
</retrieved_passages>

Perform the analysis and ranking. Ensure all XML tags are closed.

```

Base-prompt best output:

```
<answer>[7] > [2] > [12] > [5]</answer>
```

Initial strategy-guided best output:

```
<answer>[7] > [12] > [2] > [3] > [5] > [1] > [6]</answer>
```

Repair decision excerpt:

```

diagnosis_type: strategy_description_bad
action: rewrite_current_strategy
confidence: 0.95
failure_analysis: The current strategy failed because it prioritized general
definitions of genetic concepts (like incomplete dominance in [7]) over
specific factual statements regarding the specific trait in question (pea plant
height). Passage [12] explicitly states that for the tall/dwarf allele pair in pea
plants, heterozygotes are the same size as homozygous tall parents (i.e., no
medium height). The student incorrectly ranked [7] first because it defined '
intermediate phenotype' generally, missing that [12] provided the specific
negative evidence for the query's subject. The strategy needs to emphasize '
trait-specific applicability' over 'conceptual definition'.
new_strategy.name: Trait-Specific Fact Prioritization with Conceptual Context
new_strategy.objective: Rank passages by their ability to provide direct, trait-
specific factual answers to the user's query, prioritizing explicit
confirmation or denial of the phenomenon for the specific organism/trait
mentioned. General conceptual definitions are secondary and only useful if
they explicitly apply to the specific case.
new_strategy.key_promote:
- Explicit statements describing the phenotype of the specific organism/trait in
question (e.g., 'heterozygous tall/dwarf pea plants are the same size as
homozygous tall')
- Direct confirmation or denial of the specific phenomenon for the specific case
- Experimental results from the specific scientist/experiment mentioned that
address the specific trait
new_strategy.key_demote:
- General definitions of genetic concepts (e.g., incomplete dominance,
codominance) that do not explicitly link to the specific trait/organism in the
query
- Passages discussing different traits (e.g., flower color, DNA fragments) unless
used as a direct analogy explicitly validated for the target trait
- Historical background or biographical information without specific experimental
findings

```

Repaired-strategy best output:

```

<answer>
[12] > [2] > [7] > [5] > [3]
</answer>

```

Analysis. The base and initial strategy-guided roll-outs both include relevant evidence, but they over-rank general conceptual passages about inheritance or incomplete dominance. The repair model diagnoses this as a strategy-description failure: the current guidance is too general and should prioritize direct, trait-specific evidence for pea-plant height. After repair, the model ranks [12] first, which explicitly states that heterozygous tall/dwarf pea plants are the same size as homozygous tall plants, followed by broader conceptual context.

D Additional Experimental Analyses

D.1 Paired Analysis of BRIGHT Results

The main BRIGHT results use the macro average over 12 subsets, where each subset contributes equally. To complement this domain-balanced metric, we conduct a query-weighted paired analysis for fixed checkpoints, preserving the system pairing for every evaluation query. Decoding uses temperature 0. We draw 10,000 paired-bootstrap resamples with queries as the resampling unit and perform 100,000 two-sided paired sign-flip randomizations.

For SARD and SFT+GRPO ($M = 8$), the query-weighted pooled BRIGHT NDCG@10 values are 36.4901 and 35.2522, respectively. Table 5 reports this comparison together with the two bridge controls. Across all three comparisons, the paired confidence intervals exclude zero and the paired randomization tests support the improvements of full SARD.

Comparison	Δ	95% CI	p
SARD vs. SFT+GRPO ($M = 8$)	+1.2379	[0.3524, 2.1061]	0.0058
SARD vs. w/o Bridge	+0.9508	[0.1175, 1.7806]	0.0241
SARD vs. w/o Shaping	+1.2886	[0.4337, 2.1752]	0.0032

Table 5: Query-weighted pooled BRIGHT paired analysis for fixed checkpoints. Intervals use paired bootstrap resampling and p -values use two-sided paired sign-flip randomization tests.

D.2 Stronger Rollout and On-Policy Distillation Controls

To test whether SARD’s gains can be explained by a larger unguided sampling budget, we increase the SFT+GRPO rollout group from $M = 8$ to $M = 16$. This raises the BRIGHT macro average from 35.61 to 36.35. Full SARD reaches 36.78, retaining a 0.43-point lead under this stronger unguided-rollout control. This result supports the value of selectively transferring assisted successes beyond uniformly increasing unguided exploration.

To test whether on-policy distillation can provide an alternative to bridge-based transfer, we further compare SARD with two OPSD-style controls. The student generates one rollout under the base prompt, while a frozen copy of the supervised warm-up checkpoint provides full-vocabulary forward-KL supervision on the student-generated prefixes. The teacher receives either pre-assigned guidance or the complete teacher response as privileged context; the student receives no privileged

context. Table 6 reports the resulting BRIGHT and R2MED scores.

D.3 Held-Out Difficulty-Stratified Diagnostic

To identify where SARD gains arise, we use the same 500 held-out validation instances as the inference-time diagnostic. We assign each instance according to whether a shared cold-start policy succeeds through base exploration or requires subsequent assistance. We then evaluate the final SARD and SFT+GRPO models under the same base-prompt condition. Table 7 summarizes the resulting difficulty-stratified comparison.

SARD’s largest gains occur on instances requiring assistance at cold start: the combined assistance-required group improves by 0.0450 NDCG@10 over SFT+GRPO, compared with 0.0038 on base-success instances. This diagnostic supports SARD’s intended role: training-time assistance helps unlock trajectories that base-prompt exploration initially misses.

D.4 Inference-Time Guidance Diagnostic

SARD uses in-context guidance and teacher-guided repair only during training. To examine whether explicit guidance remains useful after training, we conduct a diagnostic on the 500 held-out validation instances from the source training data. For each instance, we compare the trained student under two inference modes: the default base prompt and a guidance-augmented prompt using the pre-assigned guidance. Both settings use 8 rollouts with temperature 0.7, and we compute per-instance averages for NDCG@10 and latency.

As summarized in Table 8, guidance-augmented inference improves mean NDCG@10 from 0.8676 to 0.9142, corresponding to an average gain of +0.0467. The median per-instance gain is +0.0033, indicating that much of the improvement is concentrated on a subset of instances. At the same time, guidance increases mean latency from 6.747s to 7.190s, with an average increase of +0.443s. Guidance improves NDCG@10 on 52.4% of instances, while increasing latency on 76.8% of instances; in 36.6% of instances, latency increases without improving NDCG@10. These results support using the base-prompt student as the default inference interface, while treating guidance-at-inference as an optional diagnostic setting.

Figure 4 visualizes the same comparison with absolute per-instance differences. We use absolute differences because relative percentage changes

Method	Teacher privileged context	BRIGHT	R2MED
OPSD-style	Pre-assigned guidance	31.50	28.25
OPSD-style + FullResponse	Complete teacher response	34.07	32.68
SARD (full)	—	36.78	40.50

Table 6: On-policy distillation controls evaluated under base-prompt inference. Stronger privileged context improves the OPSD-style control, while SARD achieves higher reranking performance on both benchmarks.

Cold-start stratum	Instances	SARD	SFT+GRPO	Difference
Overall	500	0.8676	0.8575	+0.0101
Base-success	424	0.9065	0.9027	+0.0038
Assistance-required	76	0.6504	0.6053	+0.0450

Table 7: Held-out difficulty-stratified NDCG@10. The assistance-required group includes instances whose cold-start policy requires initial guidance, teacher-guided repair, or teacher fallback.

Metric	Base	Guided
Mean NDCG@10	0.8676	0.9142
Median NDCG@10	0.9296	0.9545
Mean latency (s)	6.747	7.190
Median latency (s)	6.623	6.961
Difference	Value	95% CI
Mean Δ NDCG@10	+0.0467	[0.0360, 0.0579]
Median Δ NDCG@10	+0.0033	—
Mean Δ latency (s)	+0.443	[0.381, 0.506]
Median Δ latency (s)	+0.420	—
Condition on Δ	Instances	
NDCG@10 > 0	52.4%	
Latency > 0	76.8%	
Latency > 0, NDCG@10 $\leq$ 0	36.6%	

Table 8: Inference-time guidance diagnostic on the 500-instance held-out validation split. Differences are computed as guidance-augmented prompting minus base prompting. Brackets denote bootstrap 95% confidence intervals over instances.

can be unstable when the base NDCG@10 is small. Most points are near zero NDCG@10 change, while latency shifts are more often positive.

D.5 General-Domain Robustness on BEIR

We further evaluate whether SARD preserves general-domain retrieval quality on BEIR using BM25 top-100 candidates. This experiment is intended as a robustness check rather than the primary target setting, since SARD is designed for reasoning-intensive reranking. Table 9 reports averaged NDCG@10 over seven BEIR datasets. SARD remains comparable to strong reranking baselines and slightly improves over ReasonRank on the av-

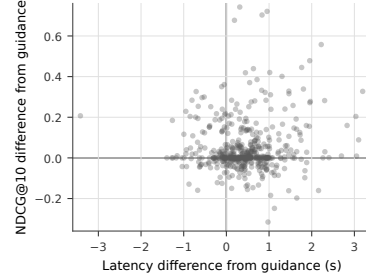


Figure 4: Per-instance quality-latency trade-off of adding explicit guidance at inference on the 500-instance held-out validation split. Each point reports the absolute difference between guidance-augmented and base prompting, computed from average NDCG@10 and latency over 8 rollouts at temperature 0.7. The horizontal and vertical lines mark zero.

Models	BEIR Average	
BM25	43.74	-
RankZephyr (7B)	54.14	+10.40
Rearank (7B)	54.29	+10.54
ReasonRank (7B)	54.35	+10.61
SARD (7B)	54.45	+10.71

Table 9: Averaged NDCG@10 on seven BEIR datasets, including Covid, DBPedia, SciFact, NFCorpus, Signal, Robust04, and News. Improvements are reported relative to BM25.

eraged score. The margin is small, so we interpret this result conservatively: SARD does not show an obvious degradation on these general-domain datasets while improving the reasoning-intensive benchmarks studied in the main experiments.

D.6 Backbone Transfer Analysis

We examine whether the SARD training design transfers beyond a single student initialization. As shown in Table 10, SARD improves the average BRIGHT score over the corresponding ReasonRank results reported under both Qwen2.5-7B-Instruct and Meta-Llama-3.1-8B-Instruct (Touvron et al., 2023). The gains are not uniform across all subsets, suggesting that the effect of adaptive assistance depends on both the backbone and the

Models	Econ.	Earth.	Rob.	Bio.	Psy.	Stack.	Sus.	Leet.	Pony	AoPS	TheoQ.	TheoT.	Avg.
ReasonIR (8B)	32.65	43.00	20.82	43.49	39.57	30.96	27.34	31.69	19.55	7.37	33.93	36.68	30.59
RankT5 (3B)	11.35	22.11	10.94	13.62	11.40	11.35	15.96	27.45	38.05	9.24	18.27	9.46	16.60
RankZephyr (7B)	19.87	17.36	12.35	34.90	24.72	13.35	22.34	29.29	32.37	6.05	28.98	30.07	22.64
Rank-R1 (7B)	20.19	27.91	18.13	36.70	30.22	11.32	29.28	17.13	9.35	3.23	14.43	28.98	20.57
Rank1 (7B)	25.34	38.91	16.77	39.91	35.32	24.81	33.47	12.71	28.14	2.58	30.67	38.18	27.23
Rearank (7B)	30.65	40.27	27.16	46.94	40.77	26.12	36.26	30.55	22.84	7.31	32.28	39.86	31.75
<i>Using Qwen2.5-7B-Instruct as the backbone</i>													
ReasonRank (7B)	35.05	47.75	31.22	56.70	47.77	32.47	40.87	23.17	24.95	7.68	39.49	41.80	35.74
SARD (7B)	37.12	50.19	31.66	57.85	47.30	34.53	43.90	25.82	23.74	7.78	40.35	41.12	36.78
<i>Using Meta-Llama-3.1-8B-Instruct as the backbone</i>													
ReasonRank (8B)	34.47	40.60	27.46	53.44	51.86	29.52	43.58	22.40	23.79	7.69	34.44	39.80	34.08
SARD (8B)	37.95	46.95	30.16	59.80	48.35	34.57	44.51	19.18	21.77	8.24	38.70	39.29	35.79

Table 10: Backbone transfer analysis on BRIGHT. SARD is compared with ReasonRank results reported under the corresponding Qwen2.5-7B-Instruct and Meta-Llama-3.1-8B-Instruct student backbones.

domain. Nevertheless, the averaged results indicate that SARD’s training schedule is not tied to one particular student model.

With Qwen2.5-7B-Instruct, SARD improves the average score from 35.74 to 36.78. With Meta-Llama-3.1-8B-Instruct, it improves the average score from 34.08 to 35.79. These results should be interpreted as a transfer check rather than a comprehensive backbone scaling study.