DVSD: DUAL-VIEW SELF-DISTILLATION FOR LONG-CONTEXT EVIDENCE COMPRESSION AND ANSWERING

Anonymous authors
Paper under double-blind review

ABSTRACT

Long-context reasoning tasks such as question answering and agent memory suffer from poor effective utilization of lengthy inputs, despite continuous expansions to model context windows. A key capability toward solving this problem is enabling models to autonomously compress long contexts into task-relevant compact evidence and leverage the compressed information for downstream tasks. We propose Dual-View Self-Distillation (DVSD), a training framework that jointly optimizes evidence compression and answer generation. A shared self-teacher provides supervision through two complementary views: an answer-privileged view uses reference answers to guide evidence construction without requiring evidence annotations, while an evidence-conditioned view forms answer targets from the question and student-generated evidence, without access to the original context or reference answer. Both views deliver dense token-level guidance along the student's on-policy trajectories, jointly training evidence construction and answer generation. Across three backbone models, DVSD achieves the highest overall scores among all baselines on LongBench v2, RULER, and InfiniteBench. It also generalizes beyond training lengths, improves agent-memory retrieval and test-time learning, and largely preserves the general capabilities.

1 INTRODUCTION

Long-context question answering and agent memory require models to locate and integrate relevant evidence from long documents and accumulated interaction histories Wu et al. (2024); Hu et al. (2026). Despite expanding context windows, effective long-context utilization remains constrained by dispersed evidence and distractor interference Liu et al. (2024); Hsieh et al. (2024); Modarressi et al. (2025). Even retrieval-augmented generation (RAG) often introduce redundant information and suffer from performance degradation as input lengths approach the context limit Jiang et al. (2024). Therefore, improving long-context capability requires two closely related abilities: integrating core information into compact evidence, and using it to answer the question.

Recent long-context post-training methods exploit the stronger performance of models when relevant information is presented in a concise context. LongPO Chen et al. (2025) constructs preferences from short-context responses to optimize long-context generation, while OPSDL Zhang et al. (2026) distills a short-context self-teacher along long-context student trajectories. However, these methods derive supervision from paired long and relevant short contexts, with their correspondence established during data construction. Moreover, they primarily supervise the final answer generation step, leaving the intermediate evidence compression process without explicit training objectives.

Generating evidence before answering allows the model to first select and organize relevant information, but this intermediate step still needs explicit supervision. LongRLVR Chen et al. (2026) observes that evidence recall stagnates under answer-only rewards and introduces a verifiable context reward for predicting evidence chunk identifiers. Effective training therefore calls for supervision of both evidence construction and evidence utilization. However, standard QA data lack evidence annotations, making such supervision difficult to obtain.

On-policy self-distillation (OPSD) offers a way to obtain such supervision. By conditioning a self-teacher on reference answers or demonstrations, OPSD provides dense token-level guidance along the trajectories Zhao et al. (2026). However, prior work on OPSD has observed reference-driven

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

shortcuts Kaur et al. (2026). Some supervisory signals depend more strongly on the reference answer than on the source input Wang et al. (2026). In long-CoT models, teacher supervision can be dominated by reference-induced components, biasing the student toward reference-specific generation patterns while weakening question-conditioned reasoning that transfers to inference Shen et al. (2026). When a single answer-privileged view supervises the entire evidence-answer trajectory, it can produce answer targets from the original context or reference answer even when the generated evidence is incomplete. The central challenge is how answer supervision can guide a model to autonomously construct evidence that is faithful to the long context and sufficient for answering, and how that evidence can become an effective condition for self-distillation.

To address this challenge, we propose Dual-View Self-Distillation (DVSD), a framework that jointly learns evidence construction and answer generation. Given a long context and a question, the student first generates compact evidence and then produces an answer, forming a single structured trajectory. A shared self-teacher supervises the two regions through complementary conditional views. The *answer-privileged evidence view* receives the same task input as the student and additionally receives the reference answer as privileged training information. The reference answer helps the teacher identify which information is relevant to the correct conclusion. This converts answer annotations into dense token-level guidance for evidence construction without requiring a reference evidence sequence. Because supervision is evaluated on student-generated evidence, it follows the compression states visited by the current policy. The *evidence-conditioned answer view* conditions only on the question and the evidence constructed by the student, without access to the original long context or the reference answer. It transfers the model's ability to answer from concise evidence through targets conditioned on the student's own compression.

Both teacher views use teacher forcing to evaluate the corresponding prefixes of the same student-generated trajectory, and separate reverse-KL objectives supervise the evidence and answer regions. As the student evolves, its generated evidence changes the conditions under which answer supervision is produced, coupling evidence construction with answer generation across training. At inference time, the trained model receives the long context and question, constructs compact evidence, and then generates the final answer.

Our main contributions are as follows:

- We formulate long-context evidence construction and answer generation as two jointly trained regions of a single on-policy trajectory, explicitly targeting both the compression and utilization of task-relevant information.
- Our framework does not require ground-truth evidence annotations. Evidence supervision is induced by the model itself through answer-privileged self-distillation and evolves together with the student policy, making the framework applicable to a broad range of long-context tasks without manually annotated supporting spans.
- We evaluate DVSD on long-context question answering and agent-memory benchmarks. Compared with SFT, RL-based methods, and existing OPSD approaches, our method improves long-context reasoning performance while preserving the general capabilities.

2 Related Work

2.1 Long-Context Post-Training

A core challenge in long-context tasks is locating task-relevant evidence. Early work addresses this problem via input compression before generation. LongLLMLingua Jiang et al. (2024) compresses prompts by retaining information relevant to the question, while RECOMP Xu et al. (2024) and SEER Zhao et al. (2024) train compressors or evidence extractors for retrieval-augmented generation. However, these methods decouple compression and answer generation as separate pipeline stages. More recent studies integrate evidence guidance into long-context post-training. LongFaith Yang et al. (2025b) constructs synthetic answers with supporting citations. LongPO Chen et al. (2025) and LongMab Duan et al. (2026) use short or selected contexts to construct training preferences, while OPSDL Zhang et al. (2026) distills a short-context teacher into a long-context student along on-policy trajectories. But they use supervision derived from shorter contexts to improve generation from long inputs. They do not explicitly train the ability to locate and integrate long con-

texts. Other methods directly guide intermediate reasoning. LongRLVR Chen et al. (2026) rewards evidence identification, EAPO Guan et al. (2026) rewards evidence quality, LongPAS Peng et al. (2026) provides feedback on reasoning steps, and MAVEN Gao & Marttinen (2026) rewards edits to an evidence memory. Existing approaches typically separate compression from answering or rely on explicit rewards for intermediate steps. In contrast, our DVSD jointly trains both stages through token-level self-distillation.

2.2 On-Policy Self-Distillation

On-policy distillation trains on responses sampled from the student using teacher predictions Agarwal et al. (2024); Gu et al. (2024). Self-distillation obtains supervision by giving the model additional information, including reference solutions in OPSD Zhao et al. (2026), demonstrations in SDFT Shenfeld et al. (2026), and feedback in SDPO Hübotter et al. (2026). OPCD Ye et al. (2026) similarly distills behavior conditioned on additional context. Purified OPSD Shen et al. (2026) further separates the effects of the reference solution and the question on teacher supervision to address harmful dependence on privileged information. Closely related to our supervision design, ViGOS Wang et al. (2026) assigns different teacher conditions to visual description and subsequent reasoning, using images to supervise the former and reference answers to guide the latter. These studies highlight the importance of tailoring teacher inputs to the supervision needs of different generation stages. However, accurate answer targets from a teacher with access to reference answers do not necessarily reflect the sufficiency of the student's intermediate evidence. DVSD uses reference answers to guide evidence construction, then derives answer supervision from the question and student-generated evidence alone.

3 Method

As illustrated in Figure 1, Dual-View Self-Distillation (DVSD) learns evidence construction and answer generation along a single student-generated trajectory, using two conditional views of a shared self-teacher to supervise the two stages.

3.1 Problem Formulation

We train on $\mathcal{D} = \{(C_i, Q_i, A_i^\star)\}_{i=1}^N$, where C, Q, and $A^\star$ denote a long context, a question, and its reference answer. We omit the instance index below for clarity. Given (C, Q), the objective is to learn a policy π_θ that compresses relevant information from C into evidence $\hat{E} = (\hat{e}_1, \ldots, \hat{e}_M)$ and then generates an answer $\hat{A} = (\hat{a}_1, \ldots, \hat{a}_L)$. The reference answer is used only for teacher supervision during training.

3.2 Joint Student Trajectory and Dual-View Self-Teaching

Joint student trajectory. For each training instance, the current student samples a structured trajectory

$$\hat{\tau} = (\hat{E}, \hat{A}) \sim \pi_\theta(\cdot \mid C, Q), \tag{1}$$

whose distribution factorizes according to the generation order:

$$\pi_\theta(\hat{E}, \hat{A} \mid C, Q) = \pi_\theta(\hat{E} \mid C, Q)\, \pi_\theta(\hat{A} \mid C, Q, \hat{E}). \tag{2}$$

We serialize the trajectory as

<evidence> $\hat{E}$ </evidence><answer> $\hat{A}$ </answer>.

At evidence position j and answer position t, the student next-token distributions are

$$p_{S,j}^E(\cdot) \triangleq \pi_\theta(\cdot \mid C, Q, \hat{e}_{<j}), \tag{3}$$

$$p_{S,t}^A(\cdot) \triangleq \pi_\theta(\cdot \mid C, Q, \hat{E}, \hat{a}_{<t}). \tag{4}$$

Both teacher views evaluate these student-generated prefixes rather than generating separate target sequences.

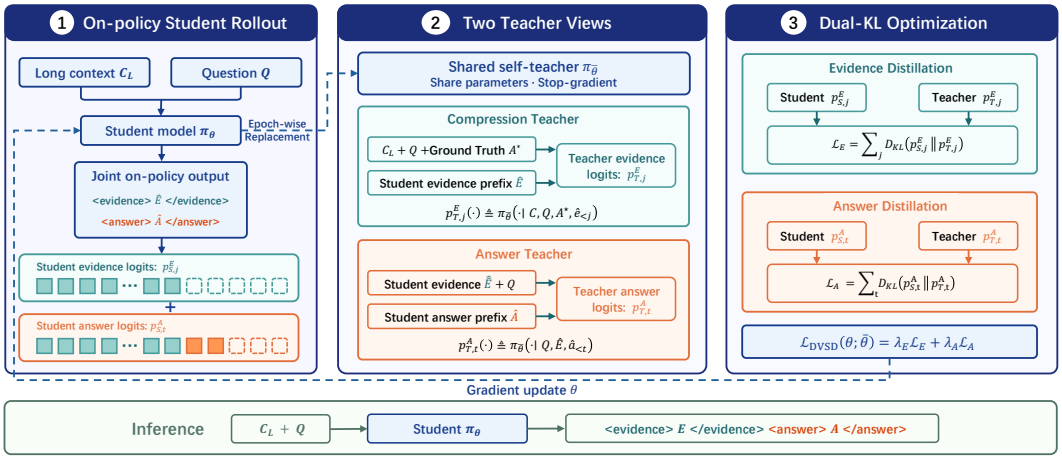


Figure 1: Overview of DVSD. From (C, Q), the student generates an on-policy evidence-answer trajectory $(\hat{E}, \hat{A})$. A shared epoch-level self-teacher evaluates this trajectory under two conditions: $(C, Q, A^\star)$ supervises evidence tokens, whereas $(Q, \hat{E})$ supervises answer tokens. The resulting region-wise reverse-KL losses update only the student, which is used alone at inference.

Two views of a shared self-teacher. Evidence construction and answer generation require different information to form useful targets: the former needs the source context together with the reference answer that determines relevance, whereas the latter must reason from the compressed evidence. DVSD therefore applies two conditional views of the same checkpoint-derived self-teacher $\pi_{\bar{\theta}}$ to the corresponding regions of $\hat{\tau}$. The views share parameters and differ only in their inputs.

The *answer-privileged evidence view* receives the same task input as the student and additionally observes $A^\star$:

$$p^E_{T,j}(\cdot) \triangleq \pi_{\bar{\theta}}(\cdot \mid C, Q, A^\star, \hat{e}_{<j}). \tag{5}$$

The reference answer helps the self-teacher identify which information is relevant to the correct conclusion. The view thereby converts direct answer supervision into an answer-aware distributional target over evidence tokens. Because the teacher evaluates evidence generated by the student, the supervision remains aligned with the compression states produced by the model itself, avoiding the distribution mismatch introduced by externally prescribed compressions.

The *evidence-conditioned answer view* conditions on the question and the evidence constructed by the student:

$$p^A_{T,t}(\cdot) \triangleq \pi_{\bar{\theta}}(\cdot \mid Q, \hat{E}, \hat{a}_{<t}). \tag{6}$$

The teacher derives its answer prediction from information compressed by the student. This view therefore transfers the model's ability to answer from short evidence rather than directly teaching it to reproduce the reference answer.

3.3 Training Objective and Theoretical Motivation

Region-wise self-distillation. Following full-vocabulary token-level self-distillation Zhao et al. (2026); Shenfeld et al. (2026), we instantiate both regional objectives with the student-to-teacher reverse KL:

$$D_{\mathrm{KL}}(p_S \,\|\, p_T) = \sum_{v \in \mathcal{V}} p_S(v) \log \frac{p_S(v)}{p_T(v)}, \tag{7}$$

where $\mathcal{V}$ is the vocabulary. The teacher parameters are held fixed, and gradients are taken only with respect to the student. For a sampled trajectory, the evidence and answer losses are

$$\ell_E(\hat{\tau}) = \frac{1}{M} \sum_{j=1}^{M} D_{\mathrm{KL}}\left(p^E_{S,j} \,\|\, p^E_{T,j}\right), \tag{8}$$

$$\ell_A(\hat{\tau}) = \frac{1}{L} \sum_{t=1}^{L} D_{\mathrm{KL}}\left(p^A_{S,t} \,\|\, p^A_{T,t}\right). \tag{9}$$

Length normalization prevents either region from receiving greater weight solely because it contains more tokens. The dataset-level losses are

$$\mathcal{L}_E(\theta;\bar{\theta}) = \mathbb{E}_{(C,Q,A^\star)\sim\mathcal{D}}\mathbb{E}_{\hat{\tau}\sim\pi_\theta(\cdot|C,Q)}\left[\ell_E(\hat{\tau})\right], \tag{10}$$

$$\mathcal{L}_A(\theta;\bar{\theta}) = \mathbb{E}_{(C,Q,A^\star)\sim\mathcal{D}}\mathbb{E}_{\hat{\tau}\sim\pi_\theta(\cdot|C,Q)}\left[\ell_A(\hat{\tau})\right]. \tag{11}$$

Their weighted combination gives the training objective

$$\mathcal{L}_{\mathrm{DVSD}}(\theta;\bar{\theta}) = \lambda_E\mathcal{L}_E + \lambda_A\mathcal{L}_A, \qquad \lambda_E,\lambda_A \geq 0, \quad \lambda_E+\lambda_A = 1. \tag{12}$$

Each sampled trajectory is treated as fixed within its corresponding update. The two terms optimize distinct stages of one structured generation process. Although ℓ_A is not differentiated through the discrete evidence tokens, the stages remain coupled through shared parameters. Changes in evidence construction alter the compressed condition from which subsequent answer targets are formed. Algorithm 1 summarizes the complete optimization process.

Why the stages require different teacher conditions. We formalize the intended information flow with an ideal evidence-mediated target distribution q over source-grounded evidence and answers:

$$q(E, A \mid C, Q) = q(E \mid C, Q)q(A \mid Q, E), \tag{13}$$

which imposes $A \perp C \mid (Q, E)$. The following result identifies the appropriate information for each stage.

Proposition (stage-compatible conditioning). Under Equation 13, for any observed reference answer $A^\star$ with $q(A^\star \mid C, Q) > 0$, conditioning on $A^\star$ yields

$$q(E \mid C, Q, A^\star) = \frac{q(E \mid C, Q)q(A^\star \mid Q, E)}{\sum_{E'} q(E' \mid C, Q)q(A^\star \mid Q, E')}, \tag{14}$$

whereas the answer-stage target is $q(A \mid Q, E)$.

Proof. By Bayes' rule and Equation 13,

$$q(E \mid C, Q, A^\star) = \frac{q(E, A^\star \mid C, Q)}{q(A^\star \mid C, Q)} = \frac{q(E \mid C, Q)q(A^\star \mid Q, E)}{\sum_{E'} q(E' \mid C, Q)q(A^\star \mid Q, E')}. \tag{15}$$

The answer conditional follows directly from $A \perp C \mid (Q, E)$.

Equation 14 shows that the reference answer serves as a likelihood term: it reweights source-grounded candidate compressions according to whether they support the correct conclusion. The evidence teacher therefore needs $(C, Q, A^\star)$. In contrast, the answer teacher should approximate $q(A \mid Q, E)$. Conditioning it on C would allow the target to bypass E. Because $A^\star$ is the observed value of A, including it in the answer view would instead collapse the ideal conditional to $\delta_{A^\star}(A)$, making the target independent of E. The two DVSD views in Equations 5 and 6 instantiate token-level approximations to these stage-compatible conditionals.

Epoch-wise teacher update. At the beginning of epoch r, the teacher is updated from the student parameters available before that epoch,

$$\bar{\theta}^{(r)} \leftarrow \theta^{(r-1)}, \qquad r = 1, \ldots, R, \tag{16}$$

and is held fixed throughout that epoch. The student obtained at the end of epoch r becomes the teacher for epoch $r + 1$.

4 Experiments

We evaluate DVSD along three complementary dimensions: static long-context understanding, memory in long-context agents, and retention of general capabilities. For each backbone, we compare methods trained on a shared dataset under matched evaluation protocols. Publicly released long-context checkpoints provide additional reference points.

Algorithm 1 Training Procedure for DVSD

Require: $\mathcal{D}$, $\theta^{(0)}$, epochs R, and $\lambda_E, \lambda_A \geq 0$ with $\lambda_E + \lambda_A = 1$
Ensure: Final student policy $\pi_{\theta^{(R)}}$

```
for r = 1, ..., R do
    Set θ ← θ^(r−1) and freeze θ̄^(r) ← θ^(r−1)
    for each mini-batch B_k, k = 0, ..., K_r − 1 do
        for each (C_i, Q_i, A_i^⋆) ∈ B_k do
            Generate τ̂_i = (Ê_i, Â_i) ~ π_θ(· | C_i, Q_i)
            Obtain student distributions p^E_{S,i,j} and p^A_{S,i,t} along τ̂_i
            Teacher-force Ê_i under (C_i, Q_i, A_i^⋆) to obtain p^E_{T,i,j}
            Teacher-force Â_i under (Q_i, Ê_i) to obtain p^A_{T,i,t}
        end for
        Compute the batch average of λ_E ℓ_E + λ_A ℓ_A
        Update θ while holding θ̄^(r) fixed
    end for
    Save the epoch checkpoint θ^(r) ← θ
end for
return π_{θ^(R)}
```

4.1 Experimental Setup

Models and data. We experiment with Llama-3.1-8B-Instruct Grattafiori et al. (2024), Qwen3-4B-Instruct-2507, and Qwen3-8B Yang et al. (2025a). Our training set consists of 10,000 examples sampled from LongRLVR-Data Chen et al. (2026), covering question answering over books, scientific articles, and code. We set the input prompt limit to 32K tokens, which falls within the native context capacity of all backbones. The evidence generation budget is set to 4K tokens, with corresponding ablation studies provided in the Appendix A.5. Answer generation is capped at 128 tokens, which fully covers the answer length requirements.

Baselines. For each backbone, we compare DVSD with the original backbones and with SFT, OPSD Zhao et al. (2026), OPSDL Zhang et al. (2026), LongPO Chen et al. (2025), and LongRLVR Chen et al. (2026) trained on the shared subset. We consider three SFT variants. SFT-A uses the reference answer as its training target. SFT-SA and SFT-SEA use the same evidence-answer output format as DVSD, with reference evidence and answers as targets. The loss is applied only to answer tokens in SFT-SA and to both evidence and answer tokens in SFT-SEA. We also include the released LongMab-PO Duan et al. (2026) and LongFaith-PO Yang et al. (2025b) checkpoints based on Llama-3.1-8B-Instruct as external references with different training data and procedures.

Evaluation. We evaluate long-context understanding on LongBench v2 Bai et al. (2025), RULER Hsieh et al. (2024), and InfiniteBench Zhang et al. (2024) and transfer to agent memory on the LongMemEval S*, EventQA-full, and NLU tasks from MemoryAgentBench Hu et al. (2026). To assess general capability retention after post-training, we evaluate MMLU-Redux Gema et al. (2025), GPQA Diamond Rein et al. (2023), IFEval Zhou et al. (2023), AIME 2025 Zhang & Math-AI (2025), LiveCodeBench v6 Jain et al. (2025), and MMMLU-Lite OpenCompass Contributors (2024), reporting individual scores and their unweighted mean. Within each benchmark and backbone, all methods follow the official evaluation configuration of the corresponding benchmark.

Implementation. We train DVSD for two epochs using LoRA Hu et al. (2021), with an effective batch size of 32 and learning rate 2×10^{-4}. Unless otherwise specified, $(\lambda_E, \lambda_A) = (0.75, 0.25)$, and the teacher is refreshed at each epoch as in Equation 16. Data preparation, training configurations, baseline adaptations, and scoring protocols are provided in Appendix A.

4.2 Main Results

Long-context performance. As shown in Table 1, DVSD achieves the highest overall scores on LongBench v2, RULER, and InfiniteBench within each backbone group. On LongBench v2, it improves over Llama-3.1-8B-Instruct, Qwen3-4B-Instruct-2507, and Qwen3-8B by 9.4, 11.9, and 15.3

Table 1: Long-context performance and general capability retention across three backbones. RULER AVG is the unweighted mean over 32K, 64K, and 128K. General AVG is the unweighted mean of the six benchmarks in Appendix Table 3. Bold and underlining indicate the best and second-best scores within each backbone group.

Models	LongBench v2						RULER				InfiniteBench	General
	Easy	Hard	Short	Medium	Long	Overall	32K	64K	128K	AVG	AVG	AVG
Llama-3.1-8B-Instruct	30.2	28.6	35.0	25.1	27.8	29.2	83.63	80.63	70.14	78.13	35.41	**39.63**
LongMab-PO	34.4	29.6	36.1	27.9	30.6	31.4	**89.12**	82.61	70.81	80.85	36.04	38.30
LongFaith-PO	37.0	26.0	38.9	23.3	29.6	30.2	88.68	82.23	74.53	81.81	38.27	37.94
+SFT-A	38.0	32.8	41.7	31.2	30.6	34.8	84.66	81.57	71.56	79.26	37.90	35.93
+SFT-SA	31.8	36.7	39.4	32.1	32.4	34.8	86.54	82.33	73.08	80.65	34.62	37.70
+SFT-SEA	29.2	28.3	34.4	26.0	24.1	28.6	80.84	77.23	65.21	74.43	32.91	32.61
+OPSD	41.7	35.7	**46.7**	33.5	32.4	38.0	87.01	82.25	71.83	80.36	36.61	37.97
+OPSDL	**42.2**	34.1	43.9	34.4	31.5	37.2	85.05	82.49	73.27	80.27	35.24	37.17
+LongPO	40.6	33.8	42.8	**34.9**	28.7	36.4	86.45	79.79	71.32	79.19	36.13	32.44
+LongRLVR	39.1	34.1	41.7	33.5	31.5	36.0	85.57	82.08	74.38	80.68	36.91	31.24
+DVSD (Ours)	39.6	**37.9**	46.1	**34.9**	**33.3**	**38.6**	89.10	**83.14**	**75.86**	**82.70**	**39.58**	37.52
Qwen3-4B-Instruct-2507	38.0	28.9	36.1	32.1	26.9	32.4	88.55	89.83	80.78	86.39	35.86	**57.24**
+SFT-A	42.7	36.3	43.9	37.2	33.3	38.8	91.53	90.00	83.14	88.22	37.70	41.70
+SFT-SA	41.1	35.7	45.6	35.3	29.6	37.8	91.81	89.92	82.71	88.15	35.06	52.21
+SFT-SEA	40.1	31.5	47.8	31.2	20.4	34.8	91.23	86.09	79.50	85.61	35.97	53.35
+OPSD	**46.9**	38.3	48.9	39.1	34.3	41.6	93.10	**92.08**	84.93	90.04	35.86	57.11
+OPSDL	43.8	34.7	42.2	36.7	34.3	38.2	91.08	90.01	84.97	88.69	35.54	56.33
+LongPO	38.5	30.9	38.9	33.0	26.9	33.8	91.32	90.71	84.92	88.98	34.09	38.98
+LongRLVR	45.8	36.3	47.2	35.8	36.1	40.0	93.10	91.56	84.37	89.68	36.20	56.18
+DVSD (Ours)	**46.9**	**42.8**	**49.4**	**42.8**	**38.9**	**44.3**	**93.55**	92.04	**85.67**	**90.42**	**38.05**	57.16
Qwen3-8B	36.5	29.3	38.9	27.4	29.6	32.0	90.64	80.31	70.52	80.49	34.27	**51.46**
+SFT-A	50.5	39.2	**54.4**	35.8	40.7	43.5	90.59	82.54	73.26	82.13	34.81	45.52
+SFT-SA	39.6	34.7	45.0	29.3	37.0	36.6	89.80	80.49	73.37	81.22	36.47	48.58
+SFT-SEA	22.9	22.2	26.1	19.1	23.1	22.5	87.25	79.01	65.32	77.19	30.16	35.65
+OPSD	47.9	42.8	51.1	40.9	41.7	44.7	92.16	84.06	74.93	83.72	34.33	51.11
+OPSDL	49.0	42.1	53.9	37.2	**44.4**	44.7	93.52	85.06	75.52	84.70	34.10	50.29
+LongPO	43.2	34.7	42.8	33.0	39.8	38.0	93.40	85.73	74.37	84.50	36.12	50.57
+LongRLVR	42.7	37.9	46.7	37.2	33.3	39.8	93.32	85.05	75.92	84.76	35.97	47.17
+DVSD (Ours)	**51.6**	**44.7**	53.3	**44.7**	42.6	**47.3**	**93.80**	**86.22**	**79.57**	**86.53**	**37.52**	51.04

Table 2: Agent-memory performance comparison on three subsets of MemoryAgentBench.

Models	LongMemEval S*		EventQA-full		NLU
	Micro Acc	Task-Avg Acc	Substring EM	Recall	Substring EM
Llama-3.1-8B-Instruct	16.67	19.00	27.0	33.8	9.0
+DVSD (Ours)	**22.33**	**25.41**	**32.6**	**37.4**	**12.0**
Qwen3-4B-Instruct-2507	13.33	15.56	38.0	35.6	17.0
+DVSD (Ours)	**19.00**	**21.37**	**45.0**	**44.4**	**60.0**
Qwen3-8B	14.67	18.15	40.6	40.0	33.0
+DVSD (Ours)	**16.00**	**20.19**	**43.8**	**43.6**	**48.0**

percentage points, respectively, exceeding the strongest competing method in each group by 0.6, 2.7, and 2.6 points. DVSD also achieves the highest accuracy on the Hard subset across all three backbones. At 128K, RULER scores improve by 4.89-9.05 points over the initial checkpoints, extending the gains beyond the 32K training context limit (Appendix Table 4). We attribute these gains in part to jointly optimizing evidence selection and utilization, which can help the model recover relevant information and reduce distraction as the context grows. Appendix Figure 9 illustrates evidence compression and answering on a LongBench v2 example from the Hard and Long subsets.

Transfer to agent memory. We select three tasks from MemoryAgentBench to examine two core memory capabilities Hu et al. (2026). LongMemEval S and EventQA-full assess Accurate Retrieval (AR) in dialogue histories and narratives, respectively, aligning with the evidence retrieval and integration skills targeted by DVSD training. NLU assesses Test-Time Learning (TTL) by requiring the model to infer and apply label mappings from contextual demonstrations, probing generalization beyond evidence-based question answering. As shown in Table 2, DVSD improves micro accuracy on LongMemEval S by 1.33-5.67 percentage points and increases both substring exact match and recall on EventQA-full. All three backbones also improve on NLU, with Qwen3-4B-Instruct-2507



Figure 2: Effect of evidence and answer loss weights on LongBench v2 across three backbones. The weights satisfy $\lambda_E + \lambda_A = 1$. Dashed lines denote base-model performance, and the default configuration $(\lambda_E, \lambda_A) = (0.75, 0.25)$ is highlighted.

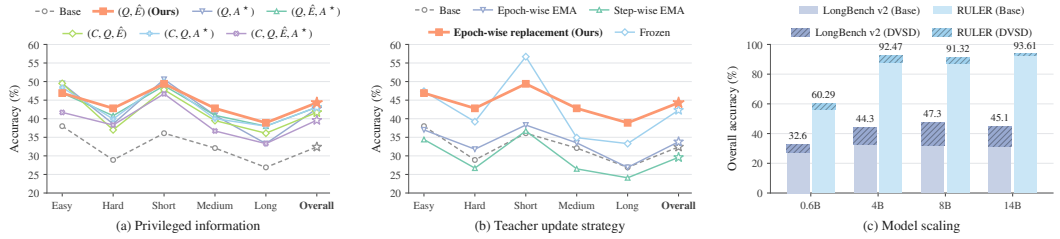


Figure 3: Effects of privileged information for answer-teacher, teacher updates, and model size. (a) LongBench v2 performance with different answer-teacher inputs on Qwen3-4B-Instruct-2507. (b) LongBench v2 performance under four teacher update strategies on the same backbone. (c) LongBench v2 and RULER performance across model sizes. In (c), total bar heights show DVSD scores, while hatched segments indicate gains over the corresponding base models. RULER scores are averaged over six context lengths from 4K to 128K.

rising from 17.0 to 60.0. Together, these results support the transfer of DVSD to evidence retrieval in memory agents and its broader applicability to learning from accumulated interactions.

General capability retention. DVSD improves long-context performance while largely retaining aggregate general capabilities. The six-benchmark average decreases by only 0.08 and 0.42 points on Qwen3-4B-Instruct-2507 and Qwen3-8B, respectively, with both declines smaller than those of any SFT variant (Table 1). Llama-3.1-8B-Instruct likewise retains most of its initial general performance, with a smaller average decline than SFT-A, SFT-SEA, LongPO, and LongRLVR (Appendix Table 3). Such retention is consistent with on-policy self-distillation using the model's own trajectories and conditional predictions to guide adaptation, potentially reducing the behavioral shift induced by post-training Shenfeld et al. (2026).

4.3 Ablation

Evidence and answer loss weights. Figure 2 compares different allocations of the distillation loss between evidence and answer tokens. Across all three backbones, the best joint configuration outperforms either loss alone. Removing the evidence loss ($\lambda_E = 0$) causes a larger degradation than removing the answer loss ($\lambda_E = 1$). Only the evidence teacher receives the reference answer, so removing its loss eliminates direct guidance on which information the compression should preserve. Because the answer teacher conditions on student-generated evidence, omissions or distortions in that evidence can also weaken subsequent answer targets. Evidence supervision therefore shapes both the student's intermediate representation and the context used for answer supervision. When $\lambda_E = 1$, this guidance is retained, but answer tokens receive no direct distillation signal. Adding the answer loss trains the long-context student to match predictions conditioned on concise evidence. The resulting gains support explicit supervision of evidence utilization alongside evidence construction. Setting $\lambda_E = 0.75$ yields the highest overall score on both Qwen models, while $\lambda_E = 0.5$ performs slightly better on Llama-3.1-8B-Instruct. We therefore adopt $(0.75, 0.25)$ as the default, emphasizing the quality of the evidence supplied to the answer teacher while retaining direct supervision of answer generation.

Privileged information for answer supervision. We vary only the answer teacher's input on Qwen3-4B-Instruct-2507, keeping all other training settings unchanged. As shown in Figure 3 (a), $(Q, \hat{E})$ achieves the highest overall accuracy of 44.3. Adding C reduces accuracy to 41.7, supporting the use of student-generated evidence to provide a concise context for answer supervision, while the original context reintroduces redundant and potentially distracting information. Adding $A^\star$ to $(Q, \hat{E})$ also lowers accuracy to 43.1. These suggest that using $\hat{E}$ alone as privileged information prevents the teacher from bypassing the compressed information to construct targets directly from the original context or reference answer. Answer supervision consequently remains tied to the information preserved in the student's compression. The difficulty breakdown further explains the value of this dependence. Compared with $(Q, A^\star)$, $(Q, \hat{E})$ scores 2.6 points lower on Easy but 4.2 and 5.6 points higher on Hard and Long, respectively. This pattern suggests that reference-answer supervision can benefit simpler instances, while more demanding questions benefit from learning to use the student's own evidence.

Teacher update strategy. We compare epoch-wise replacement, a frozen teacher, and step-wise or epoch-wise EMA on Qwen3-4B-Instruct-2507 (Figure 3 (b)). The frozen teacher retains the initial parameters throughout training, while epoch-wise replacement copies the latest student at each epoch boundary and holds the teacher fixed within the epoch. The EMA variants interpolate the teacher and student parameters through

$$\bar{\theta} \leftarrow \rho\bar{\theta} + (1-\rho)\theta, \tag{17}$$

with updates applied at the corresponding frequency. Both EMA variants use the same coefficient, $\rho_{\text{step}} = \rho_{\text{epoch}} = 0.99$, following the update scheme in SDFT Shenfeld et al. (2026). Epoch-wise replacement achieves the highest overall score of 44.3. The frozen teacher remains competitive at 42.3 and performs better on Short, but falls behind on Medium and Long. Relative to the base model's overall score of 32.4, epoch-wise EMA provides a modest improvement to 33.8, whereas step-wise EMA reduces the score to 29.6. A plausible explanation is that step-wise teacher updates introduce additional target drift as the student's evidence distribution evolves, while EMA can delay the incorporation of newly learned behavior. Epoch-wise replacement can incorporate improvements in evidence compression acquired over a full training pass and then provide a fixed supervision policy for the next epoch, combining teacher adaptation with consistent supervision.

Effect of model size. Since DVSD derives both teacher views from the backbone, model capacity may affect the quality of its supervisory distributions. As shown in Figure 3 (c), DVSD improves LongBench v2 and RULER performance at all evaluated sizes from 0.6B to 14B. The largest LongBench v2 improvement occurs at 8B, while the 14B model shows the smallest RULER gain. The latter backbone already achieves the highest baseline RULER score, leaving less headroom for absolute improvement. This suggests that the benefit of self-distillation depends on the backbone's initial proficiency on each benchmark as well as its capacity to learn from the teacher views.

Additional ablations on learning rate and LoRA configuration are reported in Appendix A.5. Appendix C further examines structured evidence and native chain-of-thought reasoning through an OPSD experiment on Qwen3-4B-Thinking-2507.

5 Conclusion

We introduce DVSD, a self-distillation framework that jointly learns evidence compression and answer generation. Its two teacher views use reference answers to guide evidence construction and student-generated evidence to supervise answering, enabling joint training without evidence annotations. Evidence thus serves both as a learned intermediate representation and as the context from which answer supervision is derived. Experiments across three backbones demonstrate consistent gains in long-context performance, generalization beyond the training length, and transfer to agent-memory retrieval and test-time learning. Ablations further support the complementary roles of evidence and answer supervision. Overall, DVSD provides a unified mechanism for learning what information to preserve from a long context and how to use that information for downstream prediction.

AI USE STATEMENT

We used OpenAI large language models to assist with code development, literature search, and drafting and polishing the manuscript. The core research activities, including problem formulation, method development, theoretical analysis, and experimental design and execution, were led and carried out by the authors. The authors take full responsibility for validating the AI-assisted code and text, interpreting the results, and ensuring the accuracy of the scientific claims. The use of an external LLM judge for evaluation is described in Appendix A.4.

REPRODUCIBILITY STATEMENT

Our experiments can be reproduced using the source code provided in the supplementary material and the experimental settings documented in the appendix. The training objective and procedure are specified in Section 3 and Algorithm 1. Appendix A.1 describes the training data and preprocessing. Optimization settings, computational resources, prompt templates, and trajectory decoding details are provided in Appendix A.2. Baseline adaptations and evaluation protocols, including the memory-agent configuration and external-judge settings, are detailed in Appendices A.3 and A.4, respectively. Additional ablation settings are reported in Appendix A.5.

REFERENCES

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos Garea, Matthieu Geist, and Olivier Bachem. On-policy distillation of language models: Learning from self-generated mistakes. In *International Conference on Learning Representations*, volume 2024, pp. 21246–21263, 2024.

Yushi Bai, Shangqing Tu, Jiajie Zhang, Hao Peng, Xiaozhi Wang, Xin Lv, Shulin Cao, Jiazheng Xu, Lei Hou, Yuxiao Dong, et al. Longbench v2: Towards deeper understanding and reasoning on realistic long-context multitasks. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3639–3664, 2025.

Guanzheng Chen, Xin Li, Michael Qizhe Shieh, and Lidong Bing. Longpo: Long context self-evolution of large language models through short-to-long preference optimization. In *International Conference on Learning Representations*, volume 2025, pp. 6638–6654, 2025.

Guanzheng Chen, Michael Qizhe Shieh, and Lidong Bing. Longrlvr: Long-context reinforcement learning requires verifiable context rewards. *arXiv preprint arXiv:2603.02146*, 2026.

Shaohua Duan, Pengcheng Huang, Xinze Li, Zhenghao Liu, Xiaoyuan Yi, Yukun Yan, Shuo Wang, Yu Gu, Ge Yu, and Maosong Sun. Chunks as arms: Multi-armed bandit-guided sampling for long-context llm preference optimization. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3559–3575, 2026.

Ya Gao and Pekka Marttinen. Evidence-state rewards for long-context reasoning. *arXiv preprint arXiv:2607.02073*, 2026.

Aryo Pradipta Gema, Joshua Ong Jun Leang, Giwon Hong, Alessio Devoto, Alberto Carlo Maria Mancino, Rohit Saxena, Xuanli He, Yu Zhao, Xiaotang Du, Mohammad Reza Ghasemi Madani, et al. Are we done with mmlu? In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 5069–5096, 2025.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. Minillm: Knowledge distillation of large language models. In *International Conference on Learning Representations*, volume 2024, pp. 32694–32717, 2024.

Xin Guan, Zijian Li, Shen Huang, Pengjun Xie, Jingren Zhou, and Jiuxin Cao. Evidence-augmented policy optimization with reward co-evolution for long-context reasoning. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 30440–30454, 2026.

Cheng-Ping Hsieh, Simeng Sun, Samuel Kriman, Shantanu Acharya, Dima Rekesh, Fei Jia, Yang Zhang, and Boris Ginsburg. Ruler: What's the real context size of your long-context language models? *arXiv preprint arXiv:2404.06654*, 2024.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

Yuanzhe Hu, Yu Wang, and Julian McAuley. Evaluating memory in llm agents via incremental multi-turn interactions. In *International Conference on Learning Representations*, volume 2026, pp. 156259–156291, 2026.

Jonas Hübotter, Frederike Lübeck, Lejs Behric, Anton Baumann, Marco Bagatella, Daniel Marta, Ido Hakimi, Idan Shenfeld, Thomas Kleine Buening, Carlos Guestrin, et al. Reinforcement learning via self-distillation. *arXiv preprint arXiv:2601.20802*, 2026.

Naman Jain, Alex Gu, Wen-Ding Li, Fanjia Yan, Tianjun Zhang, Sida Wang, Armando Solar-Lezama, Koushik Sen, and Ion Stoica. Livecodebench: Holistic and contamination free evaluation of large language models for code. In *International Conference on Learning Representations*, volume 2025, pp. 58791–58831, 2025.

Huiqiang Jiang, Qianhui Wu, Xufang Luo, Dongsheng Li, Chin-Yew Lin, Yuqing Yang, and Lili Qiu. Longllmlingua: Accelerating and enhancing llms in long context scenarios via prompt compression. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1658–1677, 2024.

Simran Kaur, Narutatsu Ri, Yinghui He, Liam Fowl, and Sanjeev Arora. Rethinking on-policy self-distillation for thinking models. *arXiv preprint arXiv:2607.05184*, 2026.

Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *Transactions of the association for computational linguistics*, 12:157–173, 2024.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Ali Modarressi, Hanieh Deilamsalehy, Franck Dernoncourt, Trung Bui, Ryan A Rossi, Seunghyun Yoon, and Hinrich Schütze. Nolima: Long-context evaluation beyond literal matching. *arXiv preprint arXiv:2502.05167*, 2025.

OpenCompass Contributors. MMMLU-Lite. Hugging Face dataset, 2024. URL `https://huggingface.co/datasets/opencompass/mmmlu_lite`.

Miao Peng, Weizhou Shen, Nuo Chen, Chenliang Li, Ming Yan, and Jia Li. Incentivizing in-depth reasoning over long contexts with process advantage shaping. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 26201–26228, 2026.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R Bowman. Gpqa: A graduate-level google-proof q&a benchmark. *arXiv preprint arXiv:2311.12022*, 2023.

Zhanming Shen, Jintao Tong, Shaotian Yan, Chen Shen, Hao Chen, Wentao Ye, Xiaomeng Hu, Rui Miao, Haobo Wang, Junbo Zhao, Gang Chen, and Jieping Ye. Purified opsd: On-policy self-distillation without losing how to think. *arXiv preprint arXiv:2607.02234*, 2026.

Idan Shenfeld, Mehul Damani, Jonas Hübotter, and Pulkit Agrawal. Self-distillation enables continual learning. *arXiv preprint arXiv:2601.19897*, 2026.

Sihan Wang, Xiyao Liu, Lianqing Liu, and Zhi Han. Seeing before reasoning: Decoupling perception and reasoning for shortcut-resilient multimodal on-policy self-distillation. *arXiv preprint arXiv:2606.19120*, 2026.

Di Wu, Hongwei Wang, Wenhao Yu, Yuwei Zhang, Kai-Wei Chang, and Dong Yu. Longmemeval: Benchmarking chat assistants on long-term interactive memory. *arXiv preprint arXiv:2410.10813*, 2024.

Fangyuan Xu, Weijia Shi, and Eunsol Choi. Recomp: Improving retrieval-augmented lms with context compression and selective augmentation. In *International Conference on Learning Representations*, volume 2024, pp. 43478–43502, 2024.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025a.

Cehao Yang, Xueyuan Lin, Chengjin Xu, Xuhui Jiang, Shengjie Ma, Aofan Liu, Hui Xiong, and Jian Guo. Longfaith: Enhancing long-context reasoning in llms with faithful synthetic data. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 3236–3256, 2025b.

Tianzhu Ye, Li Dong, Xun Wu, Shaohan Huang, and Furu Wei. On-policy context distillation for language models. *arXiv preprint arXiv:2602.12275*, 2026.

Xinrong Zhang, Yingfa Chen, Shengding Hu, Zihang Xu, Junhao Chen, Moo Khai Hao, Xu Han, Zhen Leng Thai, Shuo Wang, Zhiyuan Liu, et al. Infinitebench: Extending long context evaluation beyond 100k tokens. *arXiv preprint arXiv:2402.13718*, 2024.

Xinsen Zhang, Zhenkai Ding, Tianjun Pan, Run Yang, Chun Kang, Xue Xiong, and Jingnan Gu. Opsdl: On-policy self-distillation for long-context language models. *arXiv preprint arXiv:2604.17535*, 2026.

Yifan Zhang and Team Math-AI. American invitational mathematics examination (aime). *Mathematical Association of America (MAA)*, 2025.

Siyan Zhao, Zhihui Xie, Mengchen Liu, Jing Huang, Guan Pang, Feiyu Chen, and Aditya Grover. Self-distilled reasoner: On-policy self-distillation for large language models. *arXiv preprint arXiv:2601.18734*, 2026.

Xinping Zhao, Dongfang Li, Yan Zhong, Boren Hu, Yibin Chen, Baotian Hu, and Min Zhang. Seer: Self-aligned evidence extraction for retrieval-augmented generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 3027–3041, 2024.

Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. Instruction-following evaluation for large language models. *arXiv preprint arXiv:2311.07911*, 2023.

A Experimental Details

A.1 Training Data

We construct the post-training corpus by merging and deduplicating the training splits of LongRLVR-Data and sampling 10,000 examples Chen et al. (2026). The source dataset contains approximately 46K synthetic question-answer pairs over documents of 8K-64K tokens, released as three subsets filtered according to the capabilities of Llama-3.1-8B, Qwen2.5-7B, and Qwen2.5-14B. The sampled corpus is shared across all three backbones and controlled baselines. The 662 examples in the released test splits form the held-out set for epoch-end SFT loss evaluation.

For DVSD, preprocessing removes the original chain-of-thought and chunk-prediction instructions and strips chunk delimiters while preserving document order. Each training instance consists of a context, a question, and a reference answer. Evidence annotations are retained for baselines that require them. We denote the released evidence chunks concatenated in their stored order by E^{ref}.

```
[System]
You are a precise long-document QA reasoner. Follow the requested
XML-like output format exactly.

[User]
Read the long context and answer the question.

First compress the context into concise evidence that preserves every
fact and relationship needed for the answer. Then give a direct,
self-contained answer. Evidence must not copy irrelevant material.

<document>
{CONTEXT}
</document>

<question>
{QUESTION}
</question>

STRICT OUTPUT REQUIREMENT  THIS IS THE FINAL INSTRUCTION:  Return
exactly one sequence in the following format: <evidence>concise,
non-empty supporting evidence</evidence><answer>concise, non-empty
answer</answer>.
```

Figure 4: Student prompt conditioned on (C, Q).

A.2 Training Configuration

Optimization. We train DVSD for two epochs with LoRA applied to all attention and feed-forward linear projections, using rank $r = 8$, scaling $\alpha = 16$, and zero dropout Hu et al. (2021). Optimization uses AdamW with weight decay 0.01 and a peak learning rate of 2×10^{-4}, with linear warmup over the first 10% of steps followed by cosine decay Loshchilov & Hutter (2017). Training runs on four NVIDIA B200 GPUs in bfloat16 with gradient checkpointing. A per-device microbatch size of one and eight gradient accumulation steps give an effective batch size of 32. The distillation temperature is 1.0, and the evidence and answer losses receive weights $(\lambda_E, \lambda_A) = (0.75, 0.25)$. The teacher is copied from the student at each epoch boundary and remains fixed within the epoch.

Input construction. All DVSD views use the backbone's native chat template without a separate thinking stage. The input budget is 32,768 tokens, including instructions and control tokens. We reserve space for the complete question and, in the answer-privileged evidence view, the reference answer. Documents exceeding the remaining budget are truncated to a contiguous prefix followed by a truncation marker.

Trajectory generation. The student generates evidence followed by an answer using the serialization in Section 3.2. The evidence teacher and answer teacher evaluate the corresponding student-generated prefixes under $(C, Q, A^\star)$ and $(Q, \hat{E})$, respectively, through teacher forcing. The prompts for the student and the two teacher views are shown in Figures 4-6.

On-policy rollouts use temperature 0.7, top-$p = 0.9$, and top-$k = 50$. Evidence and answer generation have separate budgets of $B_E = 4{,}096$ and $B_A = 128$ tokens, additional to the input budget. Evidence generation terminates at the closing evidence marker, an end-of-sequence token, or the evidence budget. The trainer completes the `</evidence><answer>` boundary as needed and continues answer generation conditioned on the original input and generated evidence. Input, padding, and trainer-inserted boundary tokens are excluded from the distillation loss.

A.3 Baseline Adaptations

The controlled baselines use the same 10,000 training examples, with supervision specified by their respective objectives.

```
[System]
You are a precise long-document QA reasoner. Follow the requested
XML-like output format exactly.

[User]
Identify the minimal sufficient evidence for the question.

The reference answer is privileged training information. Use it only to
find all supporting facts and relationships in the long context. Under
teacher forcing, continue the supplied student evidence prefix
faithfully.

<document>
{CONTEXT}
</document>

<question>
{QUESTION}
</question>

<reference_answer>
{REFERENCE_ANSWER}
</reference_answer>

STRICT OUTPUT REQUIREMENT  THIS IS THE FINAL INSTRUCTION:  Return
exactly one sequence in the following format: <evidence>concise,
non-empty supporting evidence</evidence>  Do not output any text before
<evidence> or after </evidence>.
```

Figure 5: Answer-privileged evidence teacher prompt conditioned on $(C, Q, A^\star)$.

```
[System]
You are a precise long-document QA reasoner. Follow the requested
XML-like output format exactly.

[User]
Answer the question using the compressed evidence below.
Under teacher forcing, continue the supplied student answer prefix and
give a direct, self-contained answer.

<evidence>
{STUDENT_EVIDENCE}
</evidence>

<question>
{QUESTION}
</question>

STRICT OUTPUT REQUIREMENT  THIS IS THE FINAL INSTRUCTION:  Return
exactly one sequence in the following format: <answer>concise,
non-empty, self-contained answer</answer>  Do not output any text
before <answer> or after </answer>.
```

Figure 6: Evidence-conditioned answer teacher prompt conditioned on $(Q, \hat{E})$.

Supervised fine-tuning. SFT-A minimizes the cross-entropy of $A^\star$ conditioned on (C, Q). SFT-SA and SFT-SEA use the structured target $(E^{\mathrm{ref}}, A^\star)$ with the same evidence and answer delimiters as DVSD. SFT-SA applies the loss only to answer tokens following teacher-forced evidence, whereas SFT-SEA supervises both regions. Input and padding tokens are masked.

Structured targets are limited to 4,096 evidence tokens and 128 answer tokens. Overlength evidence retains its beginning and end, separated by a truncation marker, while overlength answers retain a prefix. At inference, both structured variants generate evidence before answering.

On-policy self-distillation. Our OPSD adaptation conditions a single teacher on $(C, Q, A^\star)$ to supervise the full student response generated from (C, Q) Zhao et al. (2026). OPSDL conditions the student on (C, Q) and the teacher on (E^{ref}, Q), using fixed reference evidence throughout training Zhang et al. (2026). Both adaptations apply token-level student-to-teacher reverse KL at student-generated prefixes, with gradients stopped through teacher predictions.

Preference optimization and reinforcement learning. LongPO constructs preferred responses from reference-evidence contexts and dispreferred responses from full-document contexts. It optimizes the long-context policy using short-context reference probabilities together with the preferred-response likelihood term Chen et al. (2025). LongRLVR retains chunk-labeled inputs and jointly predicts evidence-chunk identifiers and an answer. Its reward combines answer correctness with an F-score-based context reward computed against the reference chunk identifiers Chen et al. (2026).

External checkpoints. We additionally evaluate the released *Llama-3.1-8B-Instruct-LongMab-PO* and *Meta-Llama-3.1-8B-Instruct-LongFaith-PO* checkpoints Duan et al. (2026); Yang et al. (2025b). These serve as external references obtained with their original training data and optimization procedures.

A.4 Evaluation Protocols

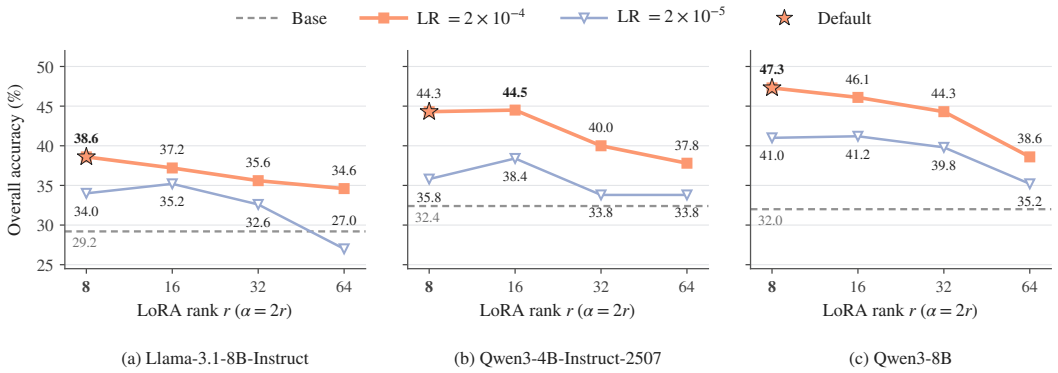
Within each benchmark and backbone, context limits, truncation rules, and decoding settings are shared across methods. All scores are reported as percentages.

Long-context understanding. LongBench v2 is evaluated using multiple-choice accuracy, reported overall and by difficulty and context length Bai et al. (2025). RULER covers all 13 tasks at 4K, 8K, 16K, 32K, 64K, and 128K tokens, with each length-specific score averaged across tasks Hsieh et al. (2024). InfiniteBench uses the benchmark's task-specific scoring functions, and its overall score is the unweighted mean across all 12 tasks Zhang et al. (2024).

Agent memory. We evaluate memory performance using the Letta agent implementation released with MemoryAgentBench in `chat` mode, following its incremental multi-turn interaction protocol Hu et al. (2026). Each evaluated checkpoint serves as the agent's language model for memory management, tool use, and answer generation. Contexts are divided into chunks of up to 32K tokens and delivered sequentially as user messages. During ingestion, the agent can update core memory, which remains in the prompt, and archival memory, which is stored externally. The dialogue history is retained as searchable recall memory. We use Letta 0.7.0 with Qwen3-Embedding-0.6B for archival retrieval and persist agent state in a local SQLite database.

LongMemEval S* contains 300 questions over conversation histories, cross-session integration, temporal reasoning, knowledge updates, and unanswerable queries. We evaluate responses with DeepSeek-V4-Flash-0731 as the external judge LLM. It runs at temperature 0 with thinking disabled and is instructed to output only `Yes` or `No`. We report accuracy over all questions as Micro Acc and the unweighted mean accuracy across the six question types as Task-Avg Acc.

EventQA-full contains 500 questions over five novels. Each question asks the model to identify the subsequent event from six candidates given a sequence of preceding events Hu et al. (2026). Substring EM credits a response containing any complete reference answer after normalization. EventQA Recall averages the per-example indicator that all required answer elements occur in the response under case-insensitive matching.



(a) Llama-3.1-8B-Instruct (b) Qwen3-4B-Instruct-2507 (c) Qwen3-8B

Figure 7: Effect of learning rate and LoRA rank on LongBench v2 overall accuracy across three backbones, with $\alpha = 2r$. Gray dashed lines denote base-model performance. Stars mark the default configuration: learning rate 2×10^{-4} and $(r, \alpha) = (8, 16)$. Bold values indicate the highest score in each panel.

NLU provides utterances paired with numeric labels during memory construction and evaluates label predictions for 100 unseen utterances Hu et al. (2026). We report Substring EM against the reference labels, averaged across the test queries.

General capability retention. We report accuracy on MMLU-Redux, GPQA Diamond, and AIME 2025 Gema et al. (2025); Rein et al. (2023); Zhang & Math-AI (2025), strict prompt-level accuracy on IFEval Zhou et al. (2023), and execution-based pass@1 on LiveCodeBench v6 Jain et al. (2025). MMMLU-Lite scores are averaged equally across its 14 language subsets OpenCompass Contributors (2024).

GPQA Diamond and AIME 2025 use three and eight independent generations per question, respectively, with accuracy averaged across generations. The remaining benchmarks use one generation per question. General AVG is the unweighted mean of the six benchmark scores.

A.5 Additional Ablation

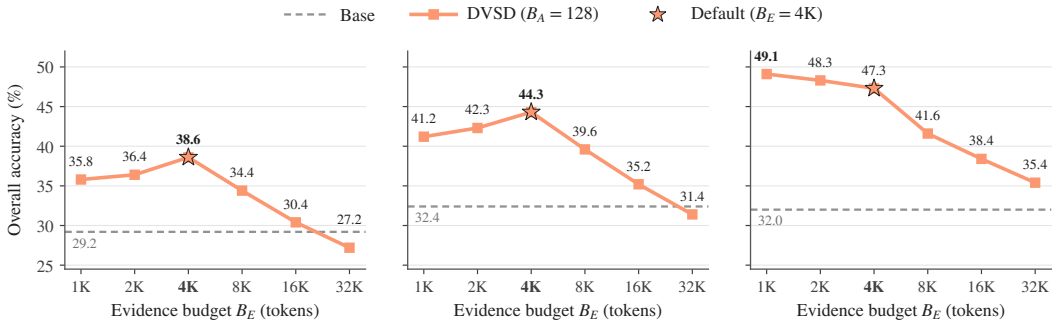
Additional ablations examine optimization hyperparameters and the evidence generation budget on LongBench v2. Across all three backbones, the hyperparameter study compares learning rates $\{2 \times 10^{-5}, 2 \times 10^{-4}\}$ and LoRA rank/scaling pairs $\{(8, 16), (16, 32), (32, 64), (64, 128)\}$, with generation budgets fixed at $(B_E, B_A) = (4{,}096, 128)$. The budget study fixes the answer generation limit at $B_A = 128$ tokens and varies only the evidence generation limit over $B_E \in \{1{,}024, 2{,}048, 4{,}096, 8{,}192, 16{,}384, 32{,}768\}$ tokens. For each backbone, DVSD is retrained separately for each value of B_E. Other settings follow the default configuration in Appendix A.2.

Learning rate and LoRA configuration. Figure 7 reports LongBench v2 overall accuracy for the tested learning rates and LoRA configurations. A learning rate of 2×10^{-4} outperforms 2×10^{-5} at every tested rank across all three backbones. The default $(r, \alpha) = (8, 16)$ achieves the highest score for Llama-3.1-8B-Instruct and Qwen3-8B. Qwen3-4B-Instruct-2507 reaches 44.5 at $(16, 32)$, compared with 44.3 at $(8, 16)$. Ranks 32 and 64 perform worse than rank 16 for all three backbones at both learning rates.

Evidence generation budget. Figure 8 varies the evidence generation limit while holding the answer limit fixed at 128 tokens. Llama-3.1-8B-Instruct and Qwen3-4B-Instruct-2507 achieve their highest scores at the default 4K evidence budget, whereas Qwen3-8B peaks at 1K with 49.1, compared with 48.3 at 2K and 47.3 at 4K. Increasing the evidence budget beyond 4K progressively reduces accuracy across all three backbones. These results show that a larger evidence generation budget does not necessarily improve downstream answering.

Table 3: General capability retention across six benchmarks. Average is the unweighted mean of the six scores. Bold and underlining mark the best and second-best scores within each backbone group.

Models	MMLU-Redux	GPQA	IFEval	AIME	LCB	MMMLU	Average
Llama-3.1-8B-Instruct	**61.70**	30.47	**76.89**	4.27	14.86	49.59	**39.63**
LongMab-PO	59.67	29.63	73.20	1.67	14.86	**50.74**	38.30
LongFaith-PO	56.14	**33.17**	71.53	2.50	15.43	48.89	37.94
+SFT-A	51.44	27.10	71.72	4.58	16.00	44.76	35.93
+SFT-SA	53.81	29.13	73.94	3.33	17.14	48.82	37.70
+SFT-SEA	48.42	23.06	66.17	4.58	12.81	40.61	32.61
+OPSD	57.21	24.41	73.20	**5.42**	**18.26**	49.30	37.97
+OPSDL	56.04	26.93	72.46	5.00	15.33	47.25	37.17
+LongPO	50.17	22.55	67.28	1.25	11.92	41.44	32.44
+LongRLVR	47.63	21.05	65.80	2.92	10.77	39.26	31.24
+DVSD (Ours)	58.46	25.76	73.75	5.00	13.14	48.99	37.52
Qwen3-4B-Instruct-2507	74.53	51.85	83.36	45.84	32.00	55.84	**57.24**
+SFT-A	71.82	7.75	68.21	40.00	7.43	55.01	41.70
+SFT-SA	68.79	39.05	80.96	41.67	27.43	55.36	52.21
+SFT-SEA	68.05	43.60	79.67	43.33	30.86	54.56	53.35
+OPSD	**74.56**	52.19	83.73	44.59	31.43	**56.14**	57.11
+OPSDL	72.47	52.19	**84.10**	**46.25**	30.29	52.65	56.33
+LongPO	28.00	23.40	82.26	45.42	28.00	26.78	38.98
+LongRLVR	72.26	51.69	83.73	45.83	31.43	52.14	56.18
+DVSD (Ours)	71.25	**53.88**	**84.10**	45.42	**33.14**	55.16	57.16
Qwen3-8B	**77.65**	48.99	83.73	18.75	22.29	57.32	**51.46**
+SFT-A	65.11	34.51	**84.66**	19.58	19.43	49.83	45.52
+SFT-SA	69.21	37.88	83.36	17.92	27.43	55.68	48.58
+SFT-SEA	51.32	15.66	28.84	**20.00**	**43.43**	54.64	35.65
+OPSD	75.51	46.63	83.73	18.75	24.57	**57.47**	51.11
+OPSDL	73.58	46.13	82.99	18.75	25.71	54.59	50.29
+LongPO	70.72	50.17	84.29	19.58	24.00	54.68	50.57
+LongRLVR	68.21	43.27	80.22	17.92	20.48	52.89	47.17
+DVSD (Ours)	73.63	**50.34**	82.62	**20.00**	25.71	53.93	51.04



(a) Llama-3.1-8B-Instruct (b) Qwen3-4B-Instruct-2507 (c) Qwen3-8B

Figure 8: Effect of the evidence generation budget on LongBench v2 overall accuracy. The answer generation budget is fixed at 128 tokens. Gray dashed lines denote base-model performance, and stars mark the default 4K evidence budget. Bold values indicate the highest score in each panel.

B Supplementary Results

Table 3 provides the individual benchmark scores underlying General AVG in Table 1. Table 4 extends the RULER comparison to all six context lengths from 4K to 128K. Its Average column aggregates all six lengths, whereas the main table averages 32K, 64K, and 128K. DVSD achieves the highest six-length average within each backbone group and improves over the corresponding base model at every evaluated length.

Table 4: Long-context performance on RULER across six context lengths. Each length-specific score averages all 13 tasks, and Average is the unweighted mean across lengths. Bold and underlining mark the best and second-best scores within each backbone group.

Models	4K	8K	16K	32K	64K	128K	Average
Llama-3.1-8B-Instruct	92.69	93.40	88.84	83.63	80.63	70.14	84.89
LongMab-PO	92.92	96.32	92.78	**89.12**	82.61	70.81	87.43
LongFaith-PO	93.35	95.84	93.49	88.68	82.23	74.53	88.02
+SFT-A	94.12	95.85	92.67	84.66	81.57	71.56	86.74
+SFT-SA	94.06	95.78	90.28	86.54	82.33	73.08	87.01
+SFT-SEA	92.21	93.98	89.12	80.84	77.23	65.21	83.10
+OPSD	94.95	**96.33**	**93.90**	87.01	82.25	71.83	87.71
+OPSDL	94.15	95.43	93.10	85.05	82.49	73.27	87.25
+LongPO	93.82	92.97	93.00	86.45	79.79	71.32	86.23
+LongRLVR	93.67	95.79	92.89	85.57	82.08	74.38	87.40
+DVSD (Ours)	**95.77**	95.22	93.25	89.10	**83.14**	**75.86**	**88.72**
Qwen3-4B-Instruct-2507	90.93	89.01	87.38	88.55	89.83	80.78	87.75
+SFT-A	93.16	93.01	92.47	91.53	90.00	83.14	90.55
+SFT-SA	94.18	93.82	93.27	91.81	89.92	82.71	90.95
+SFT-SEA	93.70	94.08	93.05	91.23	86.09	79.50	89.61
+OPSD	92.54	88.28	90.93	93.10	**92.08**	84.93	90.31
+OPSDL	93.54	88.80	89.51	91.08	90.01	84.97	89.65
+LongPO	93.72	88.85	89.95	91.32	90.71	84.92	89.91
+LongRLVR	93.38	89.39	91.96	93.10	91.56	84.37	90.63
+DVSD (Ours)	**94.68**	**94.37**	**94.52**	**93.55**	92.04	**85.67**	**92.47**
Qwen3-8B	93.64	93.83	92.55	90.64	80.31	70.52	86.92
+SFT-A	93.76	93.71	92.94	90.59	82.54	73.26	87.80
+SFT-SA	92.27	93.57	91.55	89.80	80.49	73.37	86.84
+SFT-SEA	94.42	94.37	89.58	87.25	79.01	65.32	84.99
+OPSD	94.58	94.87	93.76	92.16	84.06	74.93	89.06
+OPSDL	95.99	95.74	**96.38**	93.52	85.06	75.52	90.37
+LongPO	96.29	96.00	94.90	93.40	85.73	74.37	90.12
+LongRLVR	95.99	95.92	95.06	93.32	85.05	75.92	90.21
+DVSD (Ours)	**96.30**	**96.69**	95.36	**93.80**	**86.22**	**79.57**	**91.32**

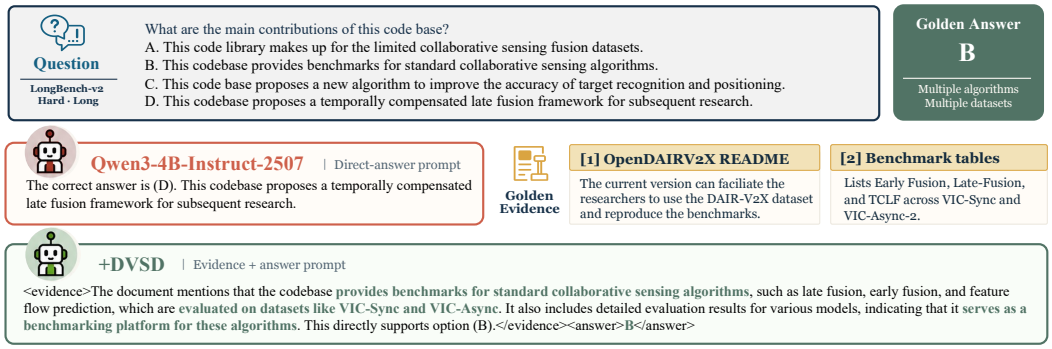


Figure 9: Case Study. Qualitative example from the Hard and Long subsets of LongBench v2. The question asks about the main contribution of the OpenDAIRV2X codebase. Qwen3-4B-Instruct-2507 selects the incorrect option D under a direct-answer prompt. DVSD, prompted to generate evidence before answering, summarizes the benchmarking functionality of the codebase and selects the correct option B. The displayed source excerpts provide supporting context for the reference answer.

C Discussion

Structured evidence versus native reasoning. The evidence sequence generated by DVSD may appear similar to a COT trajectory, but the two intermediate representations serve different roles. To examine whether native reasoning alone provides the same benefit, we apply OPSD to Qwen3-4B-

Table 5: Performance of Qwen3-4B-Thinking-2507 before and after OPSD training.

Models	LongBench v2						RULER			
	Easy	Hard	Short	Medium	Long	Overall	32K	64K	128K	AVG
Qwen3-4B-Thinking-2507	29.7	**33.1**	44.4	**27.9**	18.5	31.8	**94.63**	93.47	86.68	91.59
+OPSD	**34.9**	32.2	**48.3**	27.4	**19.4**	**33.2**	93.90	**94.22**	**87.54**	**91.89**

Thinking-2507 using the same training data. The student generates its native COT trajectory from (C, Q), while an answer-privileged teacher conditioned on $(C, Q, A^\star)$ evaluates the entire trajectory through teacher forcing. As shown in Table 5, OPSD improves LongBench v2 overall accuracy by only 1.4 percentage points and the three-length RULER average by 0.30 points. The within-checkpoint gains in this CoT experiment are smaller than those obtained by DVSD on Qwen3-4B-Instruct-2507. This result suggests that the benefit of DVSD cannot be attributed merely to generating additional intermediate tokens. Native CoT performs sequential reasoning step-by-step from the full context, but does not explicitly target at locating and consolidating scattered evidence amid distractors. When a single answer-privileged teacher supervises the entire CoT trajectory end-to-end, the supervision signal is dominated by the reference answer, encouraging reference-driven shortcuts rather than active evidence extraction from the source. The early evidence-localization steps thus receive only weak indirect guidance, leading to marginal gains. In contrast, DVSD splits the trajectory into two functionally specialized stages with dedicated teacher views. The answer-privileged view focuses exclusively on evidence compression, converting reference answers into dense guidance for selecting and integrating relevant information. The evidence-conditioned view then supervises answering using only the question and the students own generated evidence, which may in turn encourage the model to construct more faithful evidence that better supports answering. This staged, differentiated supervision directly targets the evidence utilization bottleneck, delivering substantially larger improvements than simply adding intermediate reasoning tokens.