LEARNING WHEN EVIDENCE-SENSITIVE GUIDANCE HELPS: CALIBRATING COUNTERFACTUAL ON-POLICY DISTILLATION FOR MEMORY-GROUNDED QUESTION ANSWERING

Anonymous authors
Paper under double-blind review

ABSTRACT

Memory-grounded question answering requires language models to use supporting evidence from past interactions. Counterfactual on-policy distillation provides evidence-sensitive supervision by contrasting teacher predictions with and without that evidence. However, the contrast alone does not determine how strongly the teacher target should be adjusted. We propose Mirror-Calibrated Counterfactual On-Policy Distillation (MCC-OPD), which uses a mirror KL budget to calibrate evidence-sensitive teacher corrections at each student-generated prefix without answer-level rewards or a learned value model. MCC-OPD retains the full-context teacher distribution as its base and limits the correction by the KL divergence of an equal-strength move in the opposite direction. We prove that the target divergence is convex and nondecreasing in correction strength, enabling scalar bisection to find the largest feasible coefficient. The student learns from the calibrated target with reverse KL. We also analyze outcome-based calibration through local student utility and compare the two calibration strategies empirically. Trained on only 140 LoCoMo questions, MCC-OPD outperforms strong baselines including full-context OPD, CROP, SA-OPD, and ExOPD in LLM-judge accuracy across LoCoMo, REALTALK, MuSiQue, and LongMemEval, with the largest absolute gain of 6.54 percentage points over CROP on LoCoMo. Matched ablations further show that calibration improves LLM-judge accuracy over uncalibrated counterfactual distillation on all four benchmarks.

1 INTRODUCTION

Long-term conversational assistants need to recall and use information from past interactions to answer users' questions (Maharana et al., 2024; Wu et al., 2025). Providing relevant conversation history does not ensure that a model will use it correctly. We train models to reason and answer using supplied memory (Yin & Tang, 2026; He et al., 2026). We distill this ability from a larger teacher to a smaller student, with the question and evidence supplied to both.

On-policy distillation (OPD) mitigates exposure bias by providing teacher supervision on student-generated rollouts (Agarwal et al., 2024; Gu et al., 2024). However, standard OPD treats all context uniformly in its supervision, without explicitly emphasizing the evidence needed for an answer. Counterfactual supervision compares the teacher's predictions with and without the supporting evidence. The resulting full/counterfactual (FC) target gives more weight to tokens favored by the presence of that evidence, following contrastive objectives in context-aware decoding and visual distillation (Shi et al., 2024; Liang et al., 2026; Qiu et al., 2026).

Evidence sensitivity alone, however, does not tell us how strongly to change the teacher target. A fixed contrast strength can move the target by very different amounts at different prefixes. Both teacher views also share the student's prefix: facts already mentioned there remain visible after their source is removed, and earlier errors can affect both views. In our fixed-prefix diagnostics, removing supporting evidence changes teacher predictions much more than removing the same number of

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.

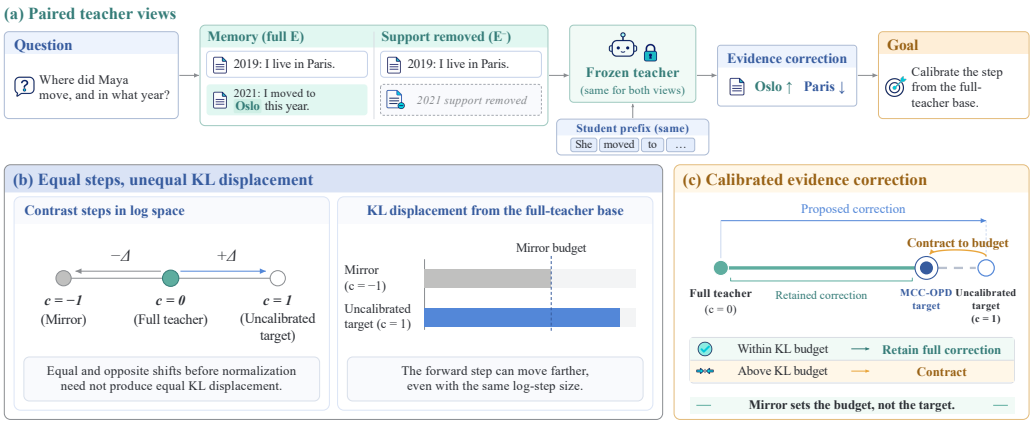

Figure 1: **Evidence contrast proposes a direction; calibration sets its strength.** **(a)** A frozen teacher compares full and support-ablated memory at the same student prefix. Facts in that prefix remain visible to both views. **(b)** Equal and opposite contrast steps can induce unequal KL displacements from the full-context teacher base; the illustrated forward step exceeds its mirror's displacement. **(c)** MCC-OPD uses the mirror displacement as a budget, retaining the uncalibrated target when feasible and otherwise contracting it to the largest feasible strength. The example and distribution glyphs are schematic.

non-support tokens. Yet the uncalibrated target still favors some incorrect answers (Section 4.4). We therefore adapt the correction strength at each student-generated prefix.

How can we calibrate evidence-sensitive teacher corrections without answer-level rewards?

We propose *Mirror-Calibrated Counterfactual On-Policy Distillation* (MCC-OPD), which calibrates evidence-sensitive teacher corrections at each student-generated prefix using only paired teacher predictions. It requires no answer-level rewards or learned value model. Equal-strength shifts from the full-context teacher in opposite log-probability directions can yield different KL divergences after normalization. The reverse shift's divergence sets the mirror KL budget for the proposed correction. MCC-OPD keeps the full correction when it is within budget; otherwise, convexity and monotonicity enable scalar bisection to find the largest feasible coefficient. The student learns from the calibrated target with reverse KL. At inference, only the student and supplied evidence are needed. Figure 1 illustrates the contrast and its calibration.

Our contributions are:

- We introduce a mirror KL budget for calibrating counterfactual teacher corrections in MCC-OPD, without answer-level rewards.
- We justify scalar bisection for finding the largest correction strength within the mirror KL budget by proving that the KL divergence from the full-context teacher is convex and non-decreasing in the correction strength.
- Trained on only 140 LoCoMo questions, MCC-OPD outperforms strong baselines in LLM-judge accuracy across four memory and reasoning benchmarks. Ablations and fixed-prefix diagnostics examine the contribution of calibration and its effects on teacher targets.

2 Related Work

On-policy distillation. Sequence distillation and OPD differ in whether training trajectories come from the teacher or student (Kim & Rush, 2016; Agarwal et al., 2024; Gu et al., 2024). Recent methods condition supervision on privileged context, feedback, or evolving inputs (Ye et al., 2026; Zhang et al., 2026c; Hübotter et al., 2026; Wang et al., 2026), and adapt it to student capacity or accessible information (Yang et al., 2026b; Xue et al., 2026; Zhang, 2026). Evidence contrasts support token selection and target reconstruction (Li et al., 2026; Jiang et al., 2026; Zhang et al.,

2026b; Liang et al., 2026; Qiu et al., 2026); outcomes support steering, trajectory selection, and rollout credit (Qu et al., 2026; Yang et al., 2026c; Gan et al., 2026; Meng & Chen, 2026; Pan et al., 2026b). ExOPD scales the teacher-to-reference log-probability ratio to extrapolate the distillation reward (Yang et al., 2026a). GC-OPD calibrates the discrepancy between group-normalized verifier rewards and OPD scores (Zhang et al., 2026d); it does not construct a full/ablated evidence contrast. MCC-OPD instead controls the displacement of a contrastive teacher target through a mirror KL budget; the outcome rule is included as a separate calibration strategy.

Counterfactual supervision and training. Context manipulations and counterfactual examples train context-faithful behavior, probe knowledge conflicts, and expose insufficient evidence (Longpre et al., 2021; Li et al., 2023; Neeman et al., 2023; Bi et al., 2025; Paranjape et al., 2022; Atanasova et al., 2022; Zhu et al., 2023; Sato et al., 2026). At inference, related interventions support context-aware decoding and source attribution (Shi et al., 2024; Cohen-Wang et al., 2024; Pan et al., 2026a). We use evidence deletion to propose a correction at a fixed student prefix, then separately determine its strength. Attribution of a response to context does not by itself determine the utility of a student update.

Agent memory. Memory systems organize experiences through storage, retrieval, and management policies (Park et al., 2023; Packer et al., 2023; Pan et al., 2025; Xu et al., 2025; Shu et al., 2026; Hu et al., 2025). Learning approaches optimize memory construction, skills, and their use (Yan et al., 2026; Wang et al., 2025; Tu et al., 2026; Zhang et al., 2026a; Zhang & Wang, 2026); memory-state alignment also extends OPD to agents whose context changes over time (Liu et al., 2026). Our scope is the complementary problem of answering from supplied memory and evidence (Zhang et al., 2024; Yin & Tang, 2026; He et al., 2026), without learning retrieval or memory editing.

3 METHOD

3.1 TASK AND PAIRED TEACHER SUPERVISION

We construct an evidence-sensitive target, calibrate its displacement from the full-context teacher base with a mirror KL budget, and distill the result on student-generated trajectories. This defines MCC-OPD. Given a question x and supplied memory E, the student π_θ generates an answer y. A frozen teacher T scores it with full memory and with annotated supporting evidence removed, denoted E^-. At position t, let $s_t = (x, E, y_{<t})$:

$$p_t = \pi_\theta(\cdot \mid s_t), \qquad q_{F,t} = T(\cdot \mid x, E, y_{<t}), \qquad q_{C,t} = T(\cdot \mid x, E^-, y_{<t}). \tag{1}$$

The question, teacher parameters, and student prefix are shared; only the evidence changes. Neither teacher view generates a separate continuation. Deletion need not remove every source of the answer or make the remaining context unanswerable. Evidence construction is specified in Appendix B; reference answers are never student inputs.

3.2 EVIDENCE-SENSITIVE TEACHER TARGET

We compare distributions on a shared partition. It retains each token in $\mathrm{Top}_K(p) \cup \mathrm{Top}_K(q_F)$ as a separate cell and groups the remaining vocabulary mass into one tail cell. Here Top_K selects the K highest-probability tokens. Suppressing position indices, let C assign tokens to cells and write $P = C^\top p$, $Q_F = C^\top q_F$, and $Q_C = C^\top q_C$. All distributions are normalized over the full vocabulary before aggregation (Appendix A.1).

Construct the evidence-sensitive teacher target.

$$\log B_j = \underbrace{\log Q_{F,j}}_{\text{full-context teacher}} + \alpha \underbrace{(\log Q_{F,j} - \log Q_{C,j})}_{\text{evidence contrast}} - \log Z_B. \tag{2}$$

Z_B normalizes B; $\Delta_j = \log B_j - \log Q_{F,j}$ is the resulting target log-ratio correction.

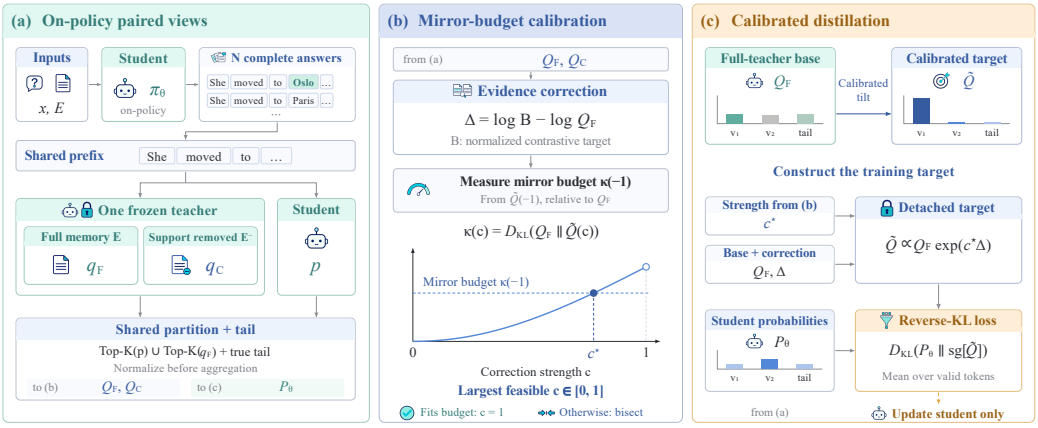

Figure 2: **Mirror-Calibrated Counterfactual On-Policy Distillation (MCC-OPD). (a)** Student answers supply a shared prefix for full and ablated teacher scoring. A shared partition, selected using student and full-context teacher probabilities, retains the remaining mass in a tail cell. **(b)** The auxiliary mirror target defines the budget $\kappa(-1)$, where $\kappa(c) = D_{\mathrm{KL}}(Q_F \| \widetilde{Q}(c))$. MCC-OPD selects the largest feasible $c \in [0, 1]$; the curve illustrates an active budget. **(c)** The calibrated target retains the full-context teacher base and is detached for reverse-KL training. Only the student is updated, without answer-level rewards. The mirror defines the budget, not the training target.

This exponential tilt follows context-aware and visual contrastive supervision (Shi et al., 2024; Liang et al., 2026; Qiu et al., 2026). We calibrate its strength through a scalar c and distill the resulting target:

$$\widetilde{Q}_j(c) = \frac{Q_{F,j} \exp(c\Delta_j)}{Z(c)} \propto Q_{F,j}^{1+c\alpha} Q_{C,j}^{-c\alpha},$$
$$\mathcal{L} = \frac{1}{M} \sum_{(i,t)\in\Omega} D_{\mathrm{KL}}\Big(P_{\theta,it} \| \operatorname{sg} \widetilde{Q}_{it}(c_{it})\Big). \tag{3}$$

Here Ω contains M valid response positions and sg stops gradients. The choices $c = 0$ and $c = 1$ recover OPD (partition KL) and uncalibrated FC, respectively. MCC-OPD chooses a coefficient between these two endpoints. If $Q_F = Q_C$, then $\Delta = 0$ and the target reduces to Q_F for any c, recovering OPD (partition KL). Token selection and all target quantities are detached during student optimization.

3.3 Calibration with a mirror KL budget

A contrast coefficient specifies a move in log-probability space, but normalization determines how far the resulting distribution moves. We therefore calibrate using the divergence from the full-context teacher base, $\kappa(c) = D_{\mathrm{KL}}(Q_F \| \widetilde{Q}(c))$. The target at $c = -1$ mirrors the proposed correction around Q_F:

$$\widetilde{Q}(-1) \propto Q_F^{1-\alpha} Q_C^{\alpha}. \tag{4}$$

At $\alpha = 1$ this is exactly the counterfactual teacher Q_C. For other strengths it is an equal-strength move toward the counterfactual side in log-probability space. Its divergence provides a reference scale for the same evidence contrast at the current prefix, without a learned budget or outcome estimator. When the paired teacher predictions are close, both the proposed correction and its mirror stay close to Q_F. Equal and opposite moves can nevertheless have different KL divergences after normalization. We reduce the forward correction when its divergence exceeds that of the mirror.

Algorithm 1 Mirror-Calibrated Counterfactual On-Policy Distillation (MCC-OPD)

Input: Training examples (x, E, E^{-}), student π_θ, frozen teacher T; top-token count K per distribution, answers per question N, contrast strength α.

for each minibatch $\mathcal{B}$ **do**
  Sample N complete answers per (x, E) from π_θ.
  Collect the valid response positions Ω.
  for each $(i, t) \in \Omega$ **do**
    Evaluate $p_{it}, q_{F,it}, q_{C,it}$ at the same student prefix (Eq. (1)).
    Build C from $\mathrm{Top}_K(p) \cup \mathrm{Top}_K(q_F)$ and a tail; aggregate P, Q_F, Q_C.
    Construct normalized FC target B and correction Δ (Eq. (2)).
    Set $c_{it} = 1$ if $\kappa(1) \le \kappa(-1)$; otherwise bisect $[0, 1]$ (Eq. (5)).
    Set $\widetilde{Q}_{it} = \mathrm{softmax}(\log Q_F + c_{it}\Delta)$; detach it and C.
  end for
  Update θ using the valid-token-mean reverse KL in Eq. (3).
end for
return the trained student π_θ.

Calibrate the correction at each student prefix.

$$c^\star = \max\Big\{ c \in [0,1] : \underbrace{\kappa(c)}_{\text{forward displacement}} \le \underbrace{\kappa(-1)}_{\text{mirror budget}} \Big\}. \tag{5}$$

1. Measure. Compute the budget from the mirror target $\widetilde{Q}(-1)$.
2. Keep or shrink. Keep $c = 1$ if feasible; otherwise bisect $[0, 1]$.
3. Distill. Freeze $\widetilde{Q}(c^\star)$ and update the student with reverse KL.

At each position (i, t), this solution gives c_{it} in Eq. (3) and Algorithm 1. The coefficient stays in $[0, 1]$: calibration neither reverses the contrast nor extrapolates beyond the raw target. The contrast strength α remains a hyperparameter.

Proposition 1 (Convex mirror-budget search)**.** *For positive Q_F and finite Δ, $\kappa(0) = \kappa'(0) = 0$, and κ is convex and nondecreasing on $[0, \infty)$. The feasible set in Eq. (5) is a nonempty closed interval. If the budget is active and Δ is nonconstant, its upper endpoint is the unique solution of $\kappa(c) = \kappa(-1)$ in $(0, 1)$.*

Proposition 1 enables a scalar bisection search for the largest feasible correction coefficient at each student prefix. Writing $\delta = \Delta - \mathbb{E}_{Q_F}\Delta$ gives $\kappa(c) = \log \mathbb{E}_{Q_F} \exp(c\delta)$, whose second derivative is $\mathrm{Var}_{\widetilde{Q}(c)}(\delta) \ge 0$. Appendix A.2 gives the proof and numerical procedure. The resulting constraint controls teacher-target displacement through $D_{\mathrm{KL}}(Q_F \| \widetilde{Q})$; the student learns from this target by minimizing $D_{\mathrm{KL}}(P_\theta \| \widetilde{Q})$. At a fixed prefix and partition the coefficient uses only teacher quantities; both the prefix and part of the partition come from the student. Reward-free here means that calibration consumes no answer reward: it still requires annotated supporting evidence.

Training and inference. For each minibatch, MCC-OPD samples N answers per question from the current student and constructs calibrated teacher targets at their prefixes (Figure 2 and Algorithm 1). We then update the student using Eq. (3), keeping the teacher fixed. At inference, the student answers questions from the supplied memory; teacher scoring and calibration are no longer needed.

4 Experiments

4.1 Experimental Setup

Tasks and evaluation. Our main experiments train on only 140 LoCoMo questions (Maharana et al., 2024), using the question-level split released by He et al. (2026), with corrections to QA errors we identified. Appendix B details the split and QA corrections. We evaluate transfer without

Table 1: **Main results on four benchmarks.** All trained students learn only from LoCoMo. LongMemEval uses the R60 evidence protocol in Appendix E. Judge is LLM-judge accuracy and B1 is BLEU-1; values are percentages. Bold and underlining mark the best and second-best 1.7B scores; the 8B teacher is a reference. Teacher-SFT uses 140 fixed teacher answers. Outcome calibration is compared in Table 5.

	LoCoMo				REALTALK			
Method	**Judge**	**F1**	**EM**	**B1**	**Judge**	**F1**	**EM**	**B1**
Qwen3-1.7B (Student)	56.88	35.32	10.91	21.79	22.53	19.98	1.92	8.09
Qwen3-8B (Teacher)	68.04	52.95	23.32	35.06	34.48	29.15	6.87	12.97
Teacher-SFT	61.58	45.09	18.54	29.66	24.18	22.38	3.30	9.97
OPD (partition KL)	61.66	45.18	17.53	29.61	23.90	22.74	4.26	9.87
OPD (sampled tokens)	59.23	45.15	19.04	29.35	24.73	23.03	5.49	9.96
CROP	56.63	44.74	20.55	30.21	22.39	21.82	5.63	9.43
SA-OPD (paper selector)	60.82	45.18	18.96	30.06	24.45	23.23	5.36	10.07
SA-OPD (code selector)	59.40	45.07	18.54	30.35	24.73	22.69	4.81	9.66
ExOPD	59.90	45.79	20.30	30.80	22.94	23.08	**6.18**	9.94
OPSD	57.05	37.77	13.93	24.24	23.08	21.17	3.98	9.28
GRPO (semantic reward)	59.48	38.84	13.51	25.13	22.53	22.09	3.02	9.02
MCC-OPD (ours)	**63.17**	**47.97**	**20.89**	**32.87**	**26.37**	**23.66**	4.95	**10.76**

	MuSiQue				LongMemEval			
Method	**Judge**	**F1**	**EM**	**B1**	**Judge**	**F1**	**EM**	**B1**
Qwen3-1.7B (Student)	36.79	29.62	17.36	25.31	45.49	18.72	3.41	11.39
Qwen3-8B (Teacher)	54.93	49.12	37.20	45.20	64.53	42.81	25.65	33.45
Teacher-SFT	31.63	28.48	18.51	24.84	41.88	32.11	16.43	24.00
OPD (partition KL)	36.37	31.45	20.26	27.37	45.89	26.56	10.62	18.87
OPD (sampled tokens)	36.65	32.24	21.73	28.84	46.89	29.14	13.63	21.89
CROP	36.42	33.05	22.61	29.83	46.09	28.19	12.83	20.75
SA-OPD (paper selector)	36.97	32.75	21.96	29.22	47.29	28.46	12.22	20.43
SA-OPD (code selector)	37.11	32.75	21.45	29.18	45.89	30.10	15.03	22.49
ExOPD	36.60	33.35	22.88	29.84	46.29	**32.79**	**17.84**	**25.37**
OPSD	39.18	32.69	21.13	28.71	44.29	20.61	5.61	13.38
GRPO (semantic reward)	37.25	31.93	20.58	28.14	44.89	20.05	3.81	12.36
MCC-OPD (ours)	**39.46**	**35.31**	**24.03**	**31.63**	**48.10**	31.64	15.63	23.30

further training on 728 REALTALK questions (Lee et al., 2025), 2,172 MuSiQue questions (Trivedi et al., 2022), and 499 LongMemEval questions (Wu et al., 2025). LoCoMo and REALTALK supply conversations; MuSiQue supplies evidence paragraphs and distractors. LongMemEval uses 60 retrieved history entries per question, with missing evidence restored for five questions (Appendix E). We report LLM-judge accuracy, answer F1, exact match (EM), and BLEU-1. Decoding, answer extraction, and completion of initially empty answers are shared across methods within each benchmark. Appendix D reports results by question category for all four benchmarks.

Training and comparisons. The student is Qwen3-1.7B, and teacher-guided distillation uses frozen Qwen3-8B. MCC-OPD uses $N = 4$ answers per question, three epochs, minibatches of up to 16 questions, and a learning rate of 2×10^{-6}. We report its final checkpoint. The shared partition uses $K = 64$ tokens from each distribution. MCC-OPD uses $\alpha = 1.5$; the outcome-calibration control uses $\alpha = 1$ and $\lambda = 0.02$ at the same $N = 4$. Our full-context baselines are *OPD (partition KL)*, which uses partition reverse KL, and *OPD (sampled tokens)*, which uses a clipped sampled-token objective. We also compare with Teacher-SFT on 140 offline teacher answers, CROP (Li et al., 2026), two SA-OPD selectors (Jiang et al., 2026), ExOPD (Yang et al., 2026a), on-policy self-distillation (OPSD) (Zhao et al., 2026), and GRPO with binary correctness rewards (Shao et al., 2024). All use the same attention backend. Appendix C specifies the baseline adaptations. The $c = 1$ counterfactual OPD control is labeled *Uncalibrated* in the ablation tables.

4.2 Main Results

Trained on only 140 LoCoMo questions, MCC-OPD achieves the highest LLM-judge accuracy on all four benchmarks among the 1.7B models in Table 1, leading 12 of the 16 metric columns. Its equally weighted averages across benchmarks are also highest for Judge (44.27%), F1 (34.65%), and BLEU-1 (24.64%). At the same $N = 4$, MCC-OPD improves every metric on every benchmark over OPD (partition KL), with an average absolute Judge gain of 2.32%. It also surpasses sampled-token OPD on all four Judge scores, by up to 3.94% on LoCoMo.

These gains extend beyond the training benchmark without further training. The same checkpoint transfers to conversational memory on REALTALK and LongMemEval and multi-hop reasoning on MuSiQue. Relative to OPD (partition KL), its absolute Judge gains on MuSiQue and LongMemEval are 3.08% and 2.20%, accompanied by F1 gains of 3.86% and 5.08%. MCC-OPD also outperforms CROP and both SA-OPD variants in Judge accuracy on every benchmark, with the largest gain of 6.54% over CROP on LoCoMo.

ExOPD provides a strong target-extrapolation comparison. MCC-OPD's absolute Judge gains over ExOPD are 3.27%, 3.43%, 2.85%, and 1.80% on LoCoMo, REALTALK, MuSiQue, and LongMemEval, respectively. MCC-OPD also surpasses Teacher-SFT in all four Judge scores and both OPSD and semantic-reward GRPO in all 16 metrics. ExOPD leads in REALTALK EM and LongMemEval's three lexical metrics; Teacher-SFT also exceeds MCC-OPD on the latter three. Appendix D provides category results, and Appendix E reports LongMemEval's task-averaged and abstention accuracies.

4.3 Ablation Study

Table 2: **Component ablations of MCC-OPD.** Each ablation removes one component from MCC-OPD. Removing mirror calibration fixes $c = 1$ (the uncalibrated target). MCC-OPD and this variant use $\alpha = 1.5$. Removing evidence contrast recovers the full-context teacher target and OPD (partition KL). All models use $N = 4$. Values are percentages; bold marks the best score in each column.

	LoCoMo				REALTALK			
Method	**Judge**	**F1**	**EM**	**B1**	**Judge**	**F1**	**EM**	**B1**
MCC-OPD (ours)	**63.17**	**47.97**	**20.89**	**32.87**	**26.37**	23.66	4.95	**10.76**
− mirror calibration	62.42	47.28	20.81	31.77	25.27	**24.31**	**5.22**	10.74
− evidence contrast	61.66	45.18	17.53	29.61	23.90	22.74	4.26	9.87
	MuSiQue				**LongMemEval**			
Method	**Judge**	**F1**	**EM**	**B1**	**Judge**	**F1**	**EM**	**B1**
MCC-OPD (ours)	**39.46**	**35.31**	**24.03**	**31.63**	**48.10**	**31.64**	**15.63**	**23.30**
− mirror calibration	36.51	32.37	21.27	28.99	47.09	28.63	11.82	21.02
− evidence contrast	36.37	31.45	20.26	27.37	45.89	26.56	10.62	18.87

Evidence contrast and calibration. Table 2 compares MCC-OPD with each component removed. Removing the evidence contrast recovers full-context OPD and lowers all four metrics on every benchmark. At the same $\alpha = 1.5$, mirror calibration improves Judge accuracy over the uncalibrated target by 0.76, 1.10, 2.95, and 1.00 percentage points on LoCoMo, REALTALK, MuSiQue, and LongMemEval, respectively. Figure 3(c) shows the paired calibration effects on all four benchmarks. The clearest benefit is on MuSiQue: all four metrics improve by 2.63–2.95 points, with positive paired 95% confidence intervals; the Judge interval is $[1.66, 4.24]$ points. On LongMemEval, F1 improves by 3.00 points, with a paired 95% interval of $[0.90, 5.17]$. The smaller Judge gains on LoCoMo, REALTALK, and LongMemEval have intervals crossing zero, and REALTALK F1 and EM decrease with calibration. Appendix C.1 reports additional component comparisons.

Contrast strength. At $N = 4$, we evaluate $\alpha \in \{0.75, 1, 1.25, 1.5, 2\}$. Figure 6 in Appendix C.3 shows the Judge and F1 curves. The response depends on the benchmark: $\alpha = 1.5$ gives the highest MuSiQue Judge and F1 and LoCoMo F1, while $\alpha = 2$ gives the highest LoCoMo Judge and

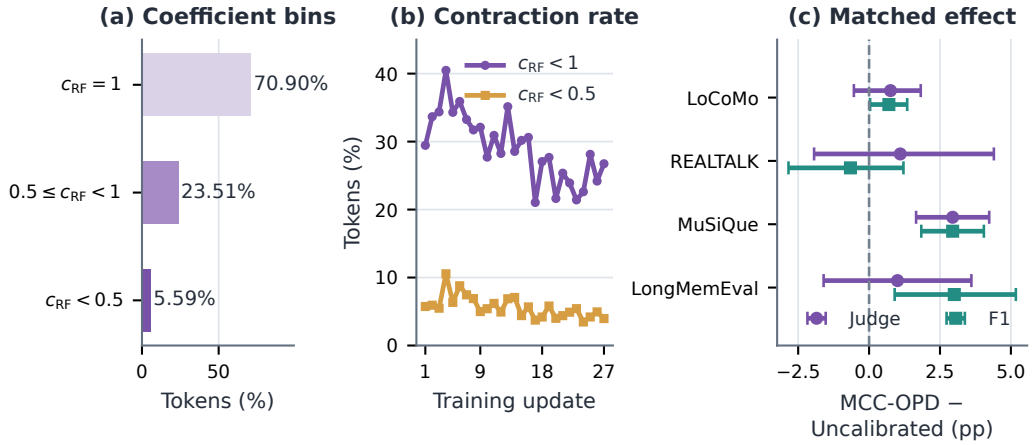

Figure 3: MCC-OPD calibration in training and in a matched component comparison. (a,b) All 124,943 valid response tokens from the 27-update MCC-OPD run; the three coefficient bins exhaust the recorded tokens. The reward-free calibration coefficient is denoted c_{RF}. (c) MCC-OPD minus Uncalibrated on all four benchmarks, with $\alpha = 1.5$ and $N = 4$ for both runs. LongMemEval results cover 499 questions. Bars are paired 95% bootstrap intervals (conversation clusters for LoCoMo/REALTALK, questions for MuSiQue/LongMemEval), conditional on the final checkpoints after 27 updates and the fixed Judge protocol.

$\alpha = 1$ the highest REALTALK Judge and F1. Our main model uses the same $\alpha = 1.5$ across all benchmarks.

Number of sampled answers. MCC-OPD retains its advantage over full-context OPD at a larger rollout count. At $N = 8$ and $\alpha = 1.5$, it exceeds OPD (partition KL) at the same N on all twelve metrics across LoCoMo, REALTALK, and MuSiQue. Increasing MCC-OPD from $N = 4$ to $N = 8$ does not uniformly improve performance: REALTALK lexical scores rise, while LoCoMo and MuSiQue scores fall, lowering mean Judge accuracy across the three benchmarks by 1.21 points. Appendix C.4 reports the results at both contrast strengths.

Calibration signal. MCC-OPD calibrates each student-generated prefix using only the paired teacher predictions. Without answer-level rewards, it achieves higher Judge accuracy than outcome calibration on all four benchmarks. Outcome calibration uses answer rewards to estimate the correction's first-order benefit to the student (*student utility*). Table 5 in Appendix C.1 reports this strategy comparison at $N = 4$, with $\alpha = 1.5$ for MCC-OPD and $\alpha = 1$ for outcome calibration. Appendix C.2 examines reward choice and availability; Appendix A.3 derives the utility estimator.

Teacher and student model sizes. Appendix C.7 compares Qwen3 teachers (4B and 8B) and students (0.6B and 1.7B) at $N = 8$. With an 8B teacher and a 0.6B student, MCC-OPD exceeds outcome calibration on 11 of 12 metrics. Larger students improve Judge accuracy across all three benchmarks, while teacher-size effects and the relative performance of the two calibration rules depend on the model pairing.

4.4 Mechanism Analysis

Evidence sensitivity and answer correctness. At fixed student prefixes, the teacher responds much more strongly to removing supporting evidence than to removing the same number of non-support tokens. We measure this response by replaying eight saved answers per training question from an outcome-calibrated student, keeping both prefixes and token partitions fixed. Mean partition JSD is 0.0913 for support deletion versus 0.0021 for the control across response positions, and 0.0248 versus 0.0009 on answer positions (Figure 4). Yet the uncalibrated target still increases the probability of some incorrect answers. Relative to the control, it promotes fewer correct answers

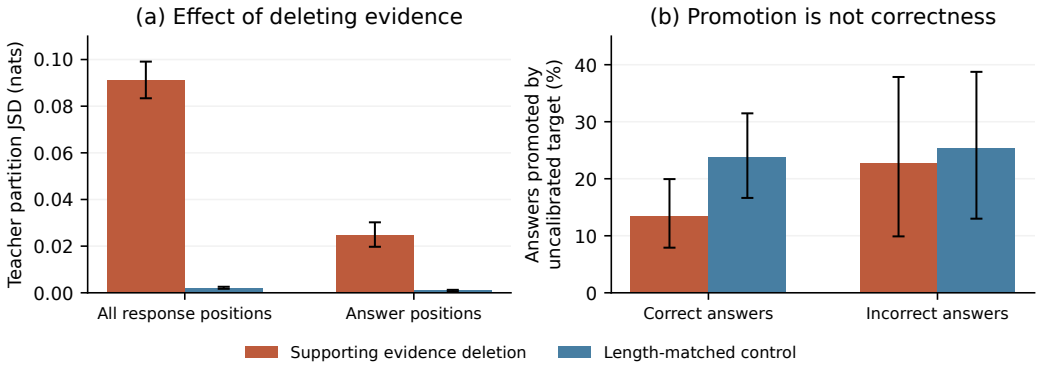

Figure 4: **Evidence sensitivity and target usefulness differ.** Left: teacher partition Jensen–Shannon divergence (JSD) after deleting annotated supporting evidence or an exactly token-length-matched set of non-support turns, at fixed student prefixes and token partitions. Right: the fraction of existing answers promoted by the corresponding uncalibrated FC target, grouped by semantic correctness. Promotion means a mean answer-token log-probability increase greater than 0.01 nat. Values are question-weighted; bars show 95% question bootstrap intervals. These are frozen-trajectory diagnostics, not post-training accuracy changes.

without a statistically clear reduction in incorrect-answer promotion. The stronger response therefore identifies sensitivity to the supplied evidence, but does not by itself identify useful corrections. Appendix F details the replay and the token-length-matched control.

How the mirror budget controls correction strength. Across all 124,943 valid response tokens in the MCC-OPD training run, the budget retains the full correction at 70.90% of positions and reduces it at 29.10% (Figure 3(a, b)). Only 5.59% of positions have a coefficient below 0.5, and the mean coefficient is 0.9121. The mirror budget thus preserves most counterfactual corrections and selectively reduces those that exceed its KL limit.

We also measure how calibration changes the local logit direction defined in Eq. (7). Summed over the training trajectory, the squared norm of this change is 8.02% of the corresponding sum for uncalibrated FC. This complements the coefficient distribution by accounting for the size of the local directions being adjusted. These are logit-level measurements on MCC-OPD's own trajectory. Appendix C.5 reports the training curves, and Appendix F.1 examines the budget on a separate set of fixed prefixes. Appendix F.2 presents paired answer changes and qualitative examples from LoCoMo, REALTALK, and MuSiQue.

5 Conclusion

We presented MCC-OPD, which uses a mirror KL budget to calibrate evidence-sensitive corrections while retaining the full-context teacher base. Convexity of the target divergence enables a scalar search for the largest feasible correction coefficient under the mirror KL constraint. Calibration requires no answer-level rewards or learned value model, and inference uses only the trained student.

Trained on only 140 LoCoMo questions, MCC-OPD achieves the highest Judge accuracy among the 1.7B models in our main comparison on LoCoMo, REALTALK, MuSiQue, and LongMemEval, leading 12 of 16 evaluation metrics. It also achieves the highest average Judge accuracy, F1, and BLEU-1 across these four benchmarks. Matched ablations also show higher Judge accuracy than uncalibrated counterfactual distillation on all four benchmarks. These results support calibrating evidence-sensitive teacher targets as a practical way to improve memory-grounded answering in small language models with limited training data. Limitations are discussed in Appendix H.

Reproducibility statement

We describe the method and training algorithm in Section 3 and Algorithm 1. Appendix A provides the theoretical assumptions and proofs. Appendices B, C, and E document data preparation and corrections, training configurations, baseline implementations, evaluated checkpoints, and evaluation protocols. Appendices F and G detail the diagnostic procedures and their assumptions.

AI use statement

We used generative AI tools to assist with literature review, method development and mathematical derivations, implementation, data review and generation of baseline prompt variants, result analysis, and preparation of text and figures. The authors take responsibility for the final content and reported results.

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*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/5be69a584901a26c521c2b51e40a4c20-Abstract-Conference.html.

Arash Ahmadian, Chris Cremer, Matthias Gallé, Marzieh Fadaee, Julia Kreutzer, Olivier Pietquin, Ahmet Üstün, and Sara Hooker. Back to basics: Revisiting REINFORCE-style optimization for learning from human feedback in LLMs. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12248–12267, 2024. doi: 10.18653/v1/2024.acl-long.662. URL https://aclanthology.org/2024.acl-long.662/.

Pepa Atanasova, Jakob Grue Simonsen, Christina Lioma, and Isabelle Augenstein. Fact checking with insufficient evidence. *Transactions of the Association for Computational Linguistics*, 10: 746–763, 2022. doi: 10.1162/tacl_a_00486. URL https://aclanthology.org/2022.tacl-1.43/.

Baolong Bi, Shaohan Huang, Yiwei Wang, Tianchi Yang, Zihan Zhang, Haizhen Huang, Lingrui Mei, Junfeng Fang, Zehao Li, Furu Wei, Weiwei Deng, Feng Sun, Qi Zhang, and Shenghua Liu. Context-DPO: Aligning language models for context-faithfulness. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 10280–10300, 2025. doi: 10.18653/v1/2025.findings-acl.536. URL https://aclanthology.org/2025.findings-acl.536/.

Benjamin Cohen-Wang, Harshay Shah, Kristian Georgiev, and Aleksander Mądry. ContextCite: Attributing model generation to context. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-3035. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/adbea136219b64db96a9941e4249a857-Abstract-Conference.html.

Siyuan Gan, Yuhan Li, Xiran Wang, Linjian Meng, Boyan Wang, Zhen Zhao, Jing Huo, and Yang Gao. When teacher guidance misleads: Reward-aligned on-policy distillation. *arXiv preprint arXiv:2608.27960*, 2026. URL https://arxiv.org/abs/2608.27960.

Shixiang Gu, Timothy Lillicrap, Zoubin Ghahramani, Richard E. Turner, and Sergey Levine. Q-Prop: Sample-efficient policy gradient with an off-policy critic. In *International Conference on Learning Representations*, 2017. URL https://arxiv.org/abs/1611.02247.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. MiniLLM: Knowledge distillation of large language models. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/8ac015d409635f196f9e3e9dcfb9a94e-Abstract-Conference.html.

Xinjie He, Zhiyuan Lin, Su Liu, Jialun Wu, Qiyang Xie, Weikai Zhou, and Shuai Xiao. What training data teaches RL memory agents: An empirical study of curriculum effects in memory-augmented QA. *arXiv preprint arXiv:2605.23067*, 2026. URL https://arxiv.org/abs/2605.23067.

Yuyang Hu, Shichun Liu, Yanwei Yue, Guibin Zhang, Boyang Liu, Fangyi Zhu, Jiahang Lin, Honglin Guo, Shihan Dou, Zhiheng Xi, Senjie Jin, Jiejun Tan, Yanbin Yin, Jiongnan Liu, Zeyu Zhang, Zhongxiang Sun, Yutao Zhu, Hao Sun, Boci Peng, Zhenrong Cheng, Xuanbo Fan, Jiaxin Guo, Xinlei Yu, Zhenhong Zhou, Zewen Hu, Jiahao Huo, Junhao Wang, Yuwei Niu, Yu Wang, Zhenfei Yin, Xiaobin Hu, Yue Liao, Qiankun Li, Kun Wang, Wangchunshu Zhou, Yixin Liu, Dawei Cheng, Qi Zhang, Tao Gui, Shirui Pan, Yan Zhang, Philip Torr, Zhicheng Dou, Ji-Rong Wen, Xuanjing Huang, Yu-Gang Jiang, and Shuicheng Yan. Memory in the Age of AI Agents. *arXiv preprint arXiv:2512.13564*, 2025. URL https://arxiv.org/abs/2512.13564.

Jonas Hübotter, Frederike Lübeck, Lejs Behric, Anton Baumann, Marco Bagatella, Daniel Marta, Ido Hakimi, Idan Shenfeld, Thomas Kleine Buening, Carlos Guestrin, and Andreas Krause. Reinforcement Learning via Self-Distillation. *arXiv preprint arXiv:2601.20802*, 2026. URL https://arxiv.org/abs/2601.20802.

Yinuo Jiang, Yongjie Ye, Zhou Tao, Xiang Zhuang, Qiang Zhang, Huajun Chen, and Tiankai Li. When Teachers Mislead: Spurious-Signal-Aware On-Policy Distillation. *arXiv preprint arXiv:2608.03632*, 2026. URL https://arxiv.org/abs/2608.03632.

Yoon Kim and Alexander M. Rush. Sequence-level knowledge distillation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 1317–1327, 2016. doi: 10.18653/v1/D16-1139. URL https://aclanthology.org/D16-1139/.

Dong-Ho Lee, Adyasha Maharana, Jay Pujara, Xiang Ren, and Francesco Barbieri. REALTALK: A 21-Day Real-World Dataset for Long-Term Conversation. *arXiv preprint arXiv:2502.13270*, 2025. URL https://arxiv.org/abs/2502.13270.

Daliang Li, Ankit Singh Rawat, Manzil Zaheer, Xin Wang, Michal Lukasik, Andreas Veit, Felix Yu, and Sanjiv Kumar. Large language models with controllable working memory. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 1774–1793, 2023. doi: 10.18653/v1/2023.findings-acl.112. URL https://aclanthology.org/2023.findings-acl.112/.

Enhan Li, Junhao He, and Hongyang Du. CROP: Task relevance via counterfactuals for selective on-policy distillation. *arXiv preprint arXiv:2608.13387v4*, 2026. URL https://arxiv.org/abs/2608.13387v4.

Yijun Liang, Yunjie Tian, Yijiang Li, Yuqi Jia, Furong Huang, Tianyi Zhou, and Di Fu. Visual contrastive self-distillation. *arXiv preprint arXiv:2607.21556*, 2026. URL https://arxiv.org/abs/2607.21556.

Zhiyuan Liu, Tinghong Ye, Chenghao Liu, Yizhuo Li, and Songfang Huang. MemOPD: On-Policy Distillation through Memory State Alignment for Long-Horizon Agents. *arXiv preprint arXiv:2608.07068*, 2026. URL https://arxiv.org/abs/2608.07068.

Shayne Longpre, Kartik Perisetla, Anthony Chen, Nikhil Ramesh, Chris DuBois, and Sameer Singh. Entity-based knowledge conflicts in question answering. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 7052–7063, 2021. doi: 10.18653/v1/2021.emnlp-main.565. URL https://aclanthology.org/2021.emnlp-main.565/.

Adyasha Maharana, Dong-Ho Lee, Sergey Tulyakov, Mohit Bansal, Francesco Barbieri, and Yuwei Fang. Evaluating very long-term conversational memory of LLM agents. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13851–13870, 2024. doi: 10.18653/v1/2024.acl-long.747. URL https://aclanthology.org/2024.acl-long.747/.

Zibin Meng and Kani Chen. CRAFT: Counterfactual credit assignment from free sibling rollouts for self-distilled agentic reinforcement learning. *arXiv preprint arXiv:2606.29476v1*, 2026. URL https://arxiv.org/abs/2606.29476v1.

Ella Neeman, Roee Aharoni, Or Honovich, Leshem Choshen, Idan Szpektor, and Omri Abend. DisentQA: Disentangling parametric and contextual knowledge with counterfactual question answering. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 10056–10070, 2023. doi: 10.18653/v1/2023.acl-long.559. URL https://aclanthology.org/2023.acl-long.559/.

Charles Packer, Sarah Wooders, Kevin Lin, Vivian Fang, Shishir G. Patil, Ion Stoica, and Joseph E. Gonzalez. MemGPT: Towards LLMs as operating systems. *arXiv preprint arXiv:2310.08560*, 2023. URL https://arxiv.org/abs/2310.08560.

Deng Pan, Keerthiram Murugesan, Ting Hua, Nuno Moniz, and Nitesh V. Chawla. Context attribution with multi-armed bandit optimization. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 11651–11662, San Diego, California, United States, July 2026a. Association for Computational Linguistics. ISBN 979-8-89176-395-1. doi: 10.18653/v1/2026.findings-acl.565. URL https://aclanthology.org/2026.findings-acl.565/.

Leyi Pan, Shuchang Tao, Yunpeng Zhai, Lingzhe Zhang, Zhaoyang Liu, Bolin Ding, Aiwei Liu, and Lijie Wen. RLCSD: Reinforcement Learning with Contrastive On-Policy Self-Distillation. *arXiv preprint arXiv:2606.11709*, 2026b. URL https://arxiv.org/abs/2606.11709.

Zhuoshi Pan, Qianhui Wu, Huiqiang Jiang, Xufang Luo, Hao Cheng, Dongsheng Li, Yuqing Yang, Chin-Yew Lin, H. Vicky Zhao, Lili Qiu, and Jianfeng Gao. SeCom: On memory construction and retrieval for personalized conversational agents. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/e56f394bbd4f0ec81393d767caa5a31b-Abstract-Conference.html.

Bhargavi Paranjape, Matthew Lamm, and Ian Tenney. Retrieval-guided counterfactual generation for QA. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1670–1686, 2022. doi: 10.18653/v1/2022.acl-long.117. URL https://aclanthology.org/2022.acl-long.117/.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, pp. 1–22, 2023. doi: 10.1145/3586183.3606763. URL https://doi.org/10.1145/3586183.3606763.

Chenhao Qiu, Dawei Li, Yechao Zhang, Lei Gong, and Zhen Tan. OPD-Aha: From linguistic momentum to visual reflection in multimodal on-policy distillation. *arXiv preprint arXiv:2609.16459*, 2026. URL https://arxiv.org/abs/2609.16459.

Zikun Qu, Min Zhang, Mingze Kong, Zhiwei Shang, Zhengyu Chen, Yikun Ban, Shuang Qiu, and Zhongxiang Dai. SPOT: Sparse probing and outcome calibration for on-policy distillation. *arXiv preprint arXiv:2608.04419v2*, 2026. URL https://arxiv.org/abs/2608.04419v2.

Haruto Sato, Yuki Tanaka, Ren Nakamura, Aoi Kobayashi, and Mei Ito. Learning Evidence Sufficiency Boundaries for Selective Answering in Grounded Multi-Hop QA. *arXiv preprint arXiv:2609.01687*, 2026. URL https://arxiv.org/abs/2609.01687.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024. URL https://arxiv.org/abs/2402.03300.

Weijia Shi, Xiaochuang Han, Mike Lewis, Yulia Tsvetkov, Luke Zettlemoyer, and Wen-tau Yih. Trusting your evidence: Hallucinate less with context-aware decoding. In *Proceedings of the 2024*

Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers), pp. 783–791, 2024. doi: 10.18653/v1/2024.naacl-short.69. URL https://aclanthology.org/2024.naacl-short.69/.

Yiheng Shu, Saisri Padmaja Jonnalagedda, Xiang Gao, Bernal Jiménez Gutiérrez, Weijian Qi, Kamalika Das, Huan Sun, and Yu Su. REMem: Reasoning with Episodic Memory in Language Agent. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=fugnQxbvMm.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. MuSiQue: Multihop questions via single-hop question composition. *Transactions of the Association for Computational Linguistics*, 10:539–554, 2022. doi: 10.1162/tacl_a_00475. URL https://aclanthology.org/2022.tacl-1.31/.

Songjun Tu, Chengdong Xu, Qichao Zhang, Yiwen Ma, Yaocheng Zhang, Linjing Li, Dong Li, Xiangyuan Lan, and Dongbin Zhao. UCOB: Learning to Utilize and Evolve Agentic Skills via Credit-Aware On-Policy Bidirectional Self-Distillation. *arXiv preprint arXiv:2606.29502*, 2026. URL https://arxiv.org/abs/2606.29502.

George Tucker, Surya Bhupatiraju, Shixiang Gu, Richard Turner, Zoubin Ghahramani, and Sergey Levine. The mirage of action-dependent baselines in reinforcement learning. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 5015–5024, 2018. URL https://proceedings.mlr.press/v80/tucker18a.html.

Yu Wang, Ryuichi Takanobu, Zhiqi Liang, Yuzhen Mao, Yuanzhe Hu, Julian McAuley, and Xiaojian Wu. Mem-α: Learning Memory Construction via Reinforcement Learning. *arXiv preprint arXiv:2509.25911*, 2025. URL https://arxiv.org/abs/2509.25911.

Yuran Wang, Zekun Wang, Bohan Zeng, Ruixu Zhang, Wenxuan Liu, Liu Yang, Yifan Dai, Yang Shi, Bozhou Li, Chengzhuo Tong, Daili Hua, Yuanxing Zhang, and Wentao Zhang. Flux-OPD: On-Policy Distillation with Evolving Contexts. *arXiv preprint arXiv:2607.28022*, 2026. URL https://arxiv.org/abs/2607.28022.

Di Wu, Hongwei Wang, Wenhao Yu, Yuwei Zhang, Kai-Wei Chang, and Dong Yu. LongMemEval: Benchmarking chat assistants on long-term interactive memory. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/d813d324dbf0598bbdc9c8e79740ed01-Abstract-Conference.html.

Wujiang Xu, Zujie Liang, Kai Mei, Hang Gao, Juntao Tan, and Yongfeng Zhang. A-Mem: Agentic memory for LLM agents. In *Advances in Neural Information Processing Systems*, volume 38, pp. 17577–17604, 2025. doi: 10.52202/085713-0593. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/19909c36f51abc4856b4560aff3d36d6-Abstract-Conference.html.

Leyan Xue, Feng Xiong, Mingjun Ma, and Changqing Zhang. Distill What the Student Can See: Fisher-Projected On-Policy Distillation for Vision-Language Models. *arXiv preprint arXiv:2608.01263*, 2026. URL https://arxiv.org/abs/2608.01263.

Sikuan Yan, Xiufeng Yang, Zuchao Huang, Ercong Nie, Zifeng Ding, Zonggen Li, Xiaowen Ma, Jinhe Bi, Kristian Kersting, Jeff Z. Pan, Hinrich Schuetze, Volker Tresp, and Yunpu Ma. Memory-R1: Enhancing large language model agents to manage and utilize memories via reinforcement learning. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12805–12825, 2026. doi: 10.18653/v1/2026.acl-long.583. URL https://aclanthology.org/2026.acl-long.583/.

Wenkai Yang, Weijie Liu, Ruobing Xie, Kai Yang, Saiyong Yang, and Yankai Lin. Learning beyond teacher: Generalized on-policy distillation with reward extrapolation. *arXiv preprint arXiv:2602.12125*, 2026a. URL https://arxiv.org/abs/2602.12125.

Yongkang Yang, Zhezheng Hao, Hong Zhang, Yi Liu, Xiankun Lin, Wence Ji, Fanjunduo Wei, Jiarui Yu, Qiang Lin, Xiaoyun Liang, and Hande Dong. Matching Supervision to the Student's Learning Capacity: A Unified Framework for On-Policy Self-Distillation. *arXiv preprint arXiv:2608.08176*, 2026b. URL https://arxiv.org/abs/2608.08176.

Yuxiao Yang, Xiaoyun Wang, and Weitong Zhang. OGLS-SD: On-policy self-distillation with outcome-guided logit steering for LLM reasoning. *arXiv preprint arXiv:2605.12400v2*, 2026c. URL https://arxiv.org/abs/2605.12400v2.

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

Jianing Yin and Tan Tang. DeferMem: Query-time evidence distillation via reinforcement learning for long-term memory QA. *arXiv preprint arXiv:2605.22411*, 2026. URL https://arxiv.org/abs/2605.22411.

Guibin Zhang, Xun Xu, Yanwei Yue, Zikun Su, Wangchunshu Zhou, Xiaobin Hu, and Shuicheng Yan. OPD-Evolver: Cultivating Holistic Agent Evolver via On-Policy Distillation. *arXiv preprint arXiv:2606.17628*, 2026a. URL https://arxiv.org/abs/2606.17628.

Kangning Zhang, Yixing Li, Shuai Shao, Qingyao Li, Zhengxi Lu, Zhiyuan Yao, Shijian Wang, Jianghao Lin, Wenxiang Jiao, Yuan Lu, Weiwen Liu, Weinan Zhang, and Yong Yu. VAD: Attributing visual evidence for target reconstruction in multimodal on-policy distillation. *arXiv preprint arXiv:2607.28590*, 2026b. URL https://arxiv.org/abs/2607.28590.

Tianjun Zhang, Shishir G. Patil, Naman Jain, Sheng Shen, Matei Zaharia, Ion Stoica, and Joseph E. Gonzalez. RAFT: Adapting language model to domain specific RAG. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=rzQGHXNReU.

Wenhao Zhang. Beyond Absolute Imitation: Anchored Residual Guidance for Privileged On-Policy Distillation. *arXiv preprint arXiv:2606.10385*, 2026. URL https://arxiv.org/abs/2606.10385.

Xinliang Frederick Zhang and Lu Wang. TSUBASA: Improving Long-Horizon Personalization via Evolving Memory and Self-Learning with Context Distillation. *arXiv preprint arXiv:2604.07894*, 2026. URL https://arxiv.org/abs/2604.07894.

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*, 2026c. URL https://arxiv.org/abs/2604.17535.

Zhu Zhang, Jixun Wang, Xiaoang Xu, Xiaorong Wang, Zihan Zhou, Zhiyuan Wang, Shuo Wang, Chaojun Xiao, and Yuezhi Zhou. Beyond Teacher Likelihood: Group-Calibrated On-Policy Distillation for Long-Context Reasoning. *arXiv preprint arXiv:2608.19181*, 2026d. URL https://arxiv.org/abs/2608.19181.

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:2601.18734v3*, 2026. URL https://arxiv.org/abs/2601.18734v3.

Yingjie Zhu, Jiasheng Si, Yibo Zhao, Haiyang Zhu, Deyu Zhou, and Yulan He. EXPLAIN, EDIT, GENERATE: Rationale-sensitive counterfactual data augmentation for multi-hop fact verification. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 13377–13392, 2023. doi: 10.18653/v1/2023.emnlp-main.826. URL https://aclanthology.org/2023.emnlp-main.826/.

Appendix Contents

A Teacher targets and calibration analysis **15**
A.1 Shared token partition . . . 15
A.2 Reward-free mirror budget . . . 15
A.3 Student utility and outcome calibration . . . 16
A.4 Assumptions . . . 18
A.5 Logit-update direction and probability responses . . . 19
A.6 Gradient equivalence of the detached-target objective . . . 19

B Data, reward, and evaluation protocol **20**
B.1 Training and implementation . . . 20
B.2 LoCoMo . . . 20
B.3 REALTALK and MuSiQue . . . 21
B.4 Semantic evaluation and reward variants . . . 21

C Baseline definitions and additional comparisons **21**
C.1 Matched component changes . . . 22
C.2 Reward and selector comparisons . . . 23
C.3 Contrast strength . . . 24
C.4 Number of sampled answers . . . 24
C.5 Calibration and training dynamics . . . 27
C.6 Number of training questions . . . 27
C.7 Teacher and student model sizes . . . 30

D Results by question category **32**
D.1 LoCoMo . . . 32
D.2 REALTALK . . . 33
D.3 MuSiQue . . . 34
D.4 LongMemEval . . . 34

E LongMemEval evaluation protocol **36**

F Diagnostic protocol and interpretation **36**
F.1 Mirror-budget replay on fixed prefixes . . . 38
F.2 Paired answer changes and examples . . . 39
F.3 Common-prefix attention cases . . . 39

G Local directional utility estimates at fixed prefixes **39**

H Limitations **43**

A Teacher targets and calibration analysis

A.1 Shared token partition

At each position, the retained token set is $S_t = \mathrm{Top}_K(p_t) \cup \mathrm{Top}_K(q_{F,t})$, where Top_K returns the K highest-probability tokens. The union removes duplicates; its complement forms one tail cell. The assignment matrix $C_t \in \{0,1\}^{|\mathcal{V}|\times(|S_t|+1)}$ maps every vocabulary token to exactly one cell. The resulting distributions are $P_t = C_t^\top p_t$, $Q_{F,t} = C_t^\top q_{F,t}$, and $Q_{C,t} = C_t^\top q_{C,t}$. Each distribution is normalized over the full vocabulary before aggregation, retaining the full remaining mass in the tail. The distillation loss is a KL on this shared partition. The directional log-probability response instead uses each sampled token's own full-vocabulary probability, even when it lies in the tail.

A.2 Reward-free mirror budget

Proof of Proposition 1.

Symbol	Meaning
p, z, F_p	Student vocabulary probabilities, logits, and softmax Jacobian.
K, S	Top-token count per distribution and retained token union.
C	Assignment matrix for the retained tokens and one tail cell.
P, Q_F, Q_C	Student, full-context teacher, and ablated teacher on that partition.
B, Δ	Uncalibrated target and target log-ratio correction $\log B - \log Q_F$.
$\alpha, c, \widetilde{Q}(c)$	Contrast strength, calibration coefficient, and calibrated target.
$Z(c), \delta$	Target normalizer and centered correction $\Delta - \mathbb{E}_{Q_F}\Delta$.
$\kappa(c), c^\star$	Teacher-target KL and the largest feasible coefficient (also denoted c_{RF} in diagnostics).
d, e, h	Logit-update direction, probability response, and log-probability response.
R, A_i, U	Answer reward, leave-one-out advantage, and local directional utility.
$\lambda, c_{\mathrm{out}}$	Outcome-calibration strength and resulting coefficient.
N, q, D, M	Answers per question, minibatch questions, $D = Nq$, and valid response tokens.
sg	Stop-gradient operator.

Table 3: Notation for the MCC-OPD method and outcome-calibration analysis.

Proof. Set $\delta_j = \Delta_j - \sum_k Q_{F,k}\Delta_k$. Expanding the KL gives

$$\kappa(c) = \log \sum_j Q_{F,j} e^{c\Delta_j} - c \sum_j Q_{F,j}\Delta_j = \log \sum_j Q_{F,j} e^{c\delta_j}. \tag{6}$$

Differentiation yields $\kappa'(c) = \mathbb{E}_{\widetilde{Q}(c)}\delta$ and $\kappa''(c) = \mathrm{Var}_{\widetilde{Q}(c)}(\delta) \geq 0$. Since $\mathbb{E}_{Q_F}\delta = 0$, the derivative vanishes at the origin and is nonnegative thereafter. Continuity and monotonicity make the feasible set a closed interval containing zero. For nonconstant δ, the variance is strictly positive at every finite c, so κ is strictly increasing for $c > 0$. An active budget satisfies $\kappa(1) > \kappa(-1) > 0$; the intermediate value theorem gives a unique boundary in $(0, 1)$. □

The implementation returns 1 if feasible and otherwise uses 32 bisection iterations on $[0, 1]$, keeping the feasible endpoint. It evaluates the centered log-sum-exp in double precision. For a constant contrast, $B = Q_F$ and κ is identically zero. Finally, substituting Eq. (2) gives $\widetilde{Q}(-1) \propto Q_F^{1-\alpha} Q_C^{\alpha}$. Thus the mirror is exactly Q_C only at $\alpha = 1$ (or in degenerate cases), rather than for all contrast strengths. The bound constrains the teacher target, not the student's KL to that target, its parameter update norm, or its expected answer reward.

Geometric examples. Figure 5 illustrates the budget at $\alpha = 1.5$ using two exact three-category distributions. With $Q_F = (0.20, 0.60, 0.20)$ and $Q_C = (0.45, 0.40, 0.15)$, the full correction is feasible and $c_{\mathrm{RF}} = 1$. With $Q_F = (0.60, 0.35, 0.05)$ and $Q_C = (0.45, 0.545, 0.005)$, the budget contracts the correction to $c_{\mathrm{RF}} \approx 0.64636$. These constructed distributions illustrate the target rule independently of the empirical model distributions.

A.3 Student utility and outcome calibration

This section defines the local directional utility signal and the outcome-calibration coefficient studied in Section 4.3. Outcome calibration uses the same target family as MCC-OPD and additionally consumes reference-based scores of complete student answers.

To identify the increment's effect, let z be independent vocabulary logits, $p = \mathrm{softmax}(z)$, and $F_p = \mathrm{diag}(p) - pp^\top$. Holding C and the targets fixed,

$$L_{\mathrm{FC}} - L_{\mathrm{full}} = -P^\top \Delta, \qquad d = -\nabla_z (L_{\mathrm{FC}} - L_{\mathrm{full}}) = F_p C \Delta. \tag{7}$$

Here $L_{\mathrm{full}} = D_{\mathrm{KL}}(P \| Q_F)$ and $L_{\mathrm{FC}} = D_{\mathrm{KL}}(P \| B)$. Thus d is the incremental negative logit gradient from uncalibrated FC. The calibrated target induces the additional direction cd. This separates proposing an evidence-sensitive direction from choosing its strength.

An infinitesimal logit move along d changes probabilities along $e = F_p d$. The second softmax Jacobian is necessary: a logit direction and its probability response are different quantities. Define

$$h(v) = d_v - p^\top d, \qquad e_v = p_v h(v), \qquad \mathbb{E}_{v\sim p}[h(v)] = 0. \tag{8}$$

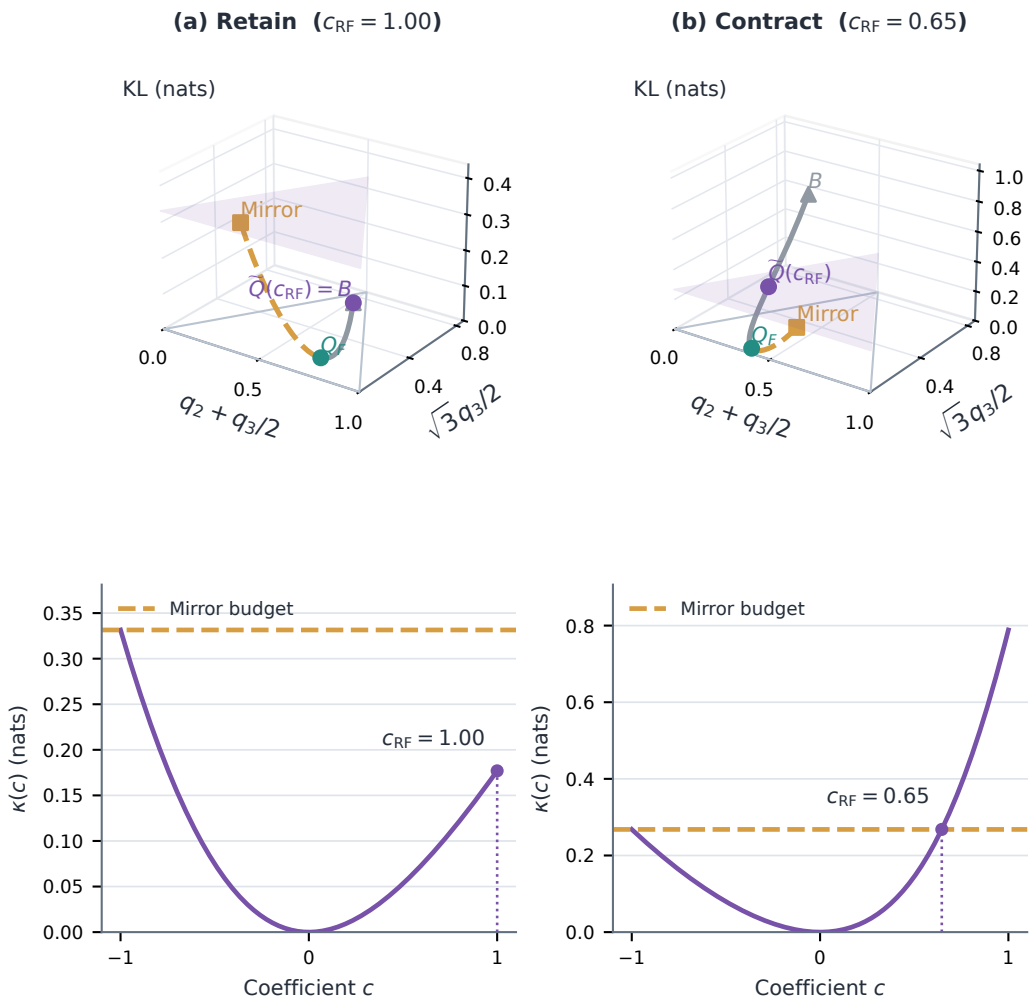

Figure 5: Exact three-category mathematical illustrations, not empirical model distributions. Q_F is the full-context teacher distribution, B the uncalibrated target and $\widetilde{Q}(c_{\mathrm{RF}})$ the MCC-OPD target. The simplex coordinates are $q_2 + q_3/2$ and $\sqrt{3}q_3/2$; height is $D_{\mathrm{KL}}(Q_F \| \widetilde{Q}(c))$. The horizontal plane is the mirror budget at $c = -1$. The lower panels show the same paths as $\kappa(c)$. MCC-OPD chooses the largest feasible coefficient c_{RF} in $[0, 1]$, retaining or contracting the forward correction. Here $\alpha = 1.5$, so the mirror is generally distinct from Q_C. Distributions and exact coordinates are supplied with the figure source.

Here e_v is the directional probability response, while $h(v) = e_v/p_v$ is the directional log-probability response: $h(v) = \partial_\epsilon \log \mathrm{softmax}(z + \epsilon d)_v|_{\epsilon=0}$. Positive $h(v)$ indicates a local increase in the token's probability. It uses the token's full-vocabulary probability, even when it belongs to the aggregated tail. Consequently, a large teacher contrast need not produce a large student response.

Let R be the completed-answer reward and $V(s, v) = \mathbb{E}[R \mid s, y_t = v]$ under a fixed continuation policy. This value is analytical; we do not fit a critic.

Proposition 2 (Local directional utility identity)**.** *Fix the state, continuation policy, scorer, and stopping rule, and assume an integrable reward. If $y_t \sim p$, then for a state baseline $b(s)$,*

$$U(s) := \frac{\partial}{\partial \epsilon} \mathbb{E}_{v \sim \mathrm{softmax}(z+\epsilon d)}[V(s, v)]\Big|_{\epsilon=0} = e^\top V = \mathbb{E}[(R - b(s))h(y_t) \mid s]. \tag{9}$$

The identity links an outcome-weighted response to the local utility of this specific direction. It requires the actual sampling distribution to match the actor probabilities used for h. The analysis below gives the assumptions, proof, and a counterexample to extending the result to arbitrary shared parameters.

We sample $N > 1$ independent complete answers to each question and compute leave-one-out advantages (Ahmadian et al., 2024):

$$A_i = R_i - \frac{1}{N-1} \sum_{j \neq i} R_j, \qquad c_{\mathrm{out},it} = 1 + \lambda \frac{M}{D} \mathrm{sg}[A_i h_{it}(y_{it})]. \tag{10}$$

Here D counts complete answers in the minibatch, and each baseline uses only answers to the same question. Conditional on the input, policy, and scorer, the other answers supply an action-independent baseline; they need not branch from or visit the current prefix. Under Proposition 2's assumptions, $\mathbb{E}[A_i h_{it}(y_{it}) \mid s_{it}] = U(s_{it})$. The correction strengthens when it raises a token's probability in an above-baseline answer or lowers it in a below-baseline answer; the opposite combinations weaken it. The coefficient is not clipped or sign-gated. Equal rewards recover uncalibrated FC.

The factor M/D converts the added token-mean loss contribution into an average of within-answer token sums. Lemma 2 derives this gradient equivalence for the realized batch; the weighting is not length-invariant. Individual products are noisy directional signals, not causal token credit or guarantees of finite-step gains.

A.4 Assumptions

A1: Fixed partition. At a fixed student-generated prefix, let $C \in \{0, 1\}^{|\mathcal{V}| \times (|S|+1)}$ assign each vocabulary token to one explicit token or the tail cell. Every row has one unit entry, so $P = C^\top p$ and $C\Delta$ lifts a cell correction to tokens. The vocabulary is finite; the student and teacher cell probabilities in logarithms are positive. We hold C and both teacher targets fixed during differentiation, without differentiating through top-K selection.

A2: Local intervention. Only the next-token independent logits at this state change along a fixed d. The input, continuation policy, scorer, decoding and stopping rules remain fixed, and the reward is integrable. The analytical $V(s, v)$ is not a fitted value model.

A3: Matching sampling. The sampled next-token distribution equals the actor softmax p used to compute h. Synchronized, temperature-one, untruncated rollouts target this condition.

A4: Independent sampled answers. Conditional on question, policy, reference, and fixed scorer, each sampled answer is an independent trajectory scored on its own. The leave-one-out baseline excludes the current answer; other answers need not visit its intermediate state.

A5: Fixed batch and detachment. For gradient equivalence, the realized batch, valid-token mask, $M > 0$, and $D = Nq > 0$ are fixed after excluding distributed padding. The token partition, teacher targets, advantages, log-probability responses, and coefficients are detached. After data-parallel averaging, the effective reduction is a global valid-token mean.

A.5 LOGIT-UPDATE DIRECTION AND PROBABILITY RESPONSES

Lemma 1 (Exact counterfactual logit direction)**.** *Let B be the normalized target in Eq.* (2) *and $\Delta = \log B - \log Q_F$. Then*

$$D_{\mathrm{KL}}(P\|B) - D_{\mathrm{KL}}(P\|Q_F) = -P^\top \Delta, \qquad -\nabla_z \big[D_{\mathrm{KL}}(P\|B) - D_{\mathrm{KL}}(P\|Q_F)\big] = F_p C \Delta. \tag{11}$$

For any finite detached scalar c, the additional negative logit gradient from targeting $\widetilde{Q}(c) \propto Q_F \exp(c\Delta)$ instead of Q_F is $cF_pC\Delta$.

Proof. Subtracting the KL expressions cancels $P_j \log P_j$ and leaves $-\sum_j P_j \Delta_j = -p^\top C\Delta$. The softmax Jacobian is $\partial p/\partial z = F_p$, giving the second equality. Finally, $\log \widetilde{Q}(c) = \log Q_F + c\Delta - \log Z(c)$. Its final term is constant across cells and $\sum_j P_j = 1$, so it has zero logit gradient. Detachment fixes c and Δ during differentiation. □

Tail-token response. Write $a = C\Delta$ and $\bar{a} = p^\top a = P^\top \Delta$. Then

$$d_v = p_v(a_v - \bar{a}), \qquad h(v) = p_v(a_v - \bar{a}) - \sum_u p_u^2 (a_u - \bar{a}). \tag{12}$$

Tail tokens share $a_v = \Delta_\perp$ but retain different p_v; their d_v and $h(v)$ therefore need not agree. The implementation uses the actual sampled token probability, not the tail-cell mass.

Proof of Proposition 2. Holding d fixed at the current state, the directional derivative of the next-token distribution under $z \mapsto z + \epsilon d$ is $F_p d$. Its vth component is

$$(F_p d)_v = p_v(d_v - p^\top d) = p_v h(v), \qquad \sum_v p_v h(v) = 0. \tag{13}$$

Multiplying by $V(s,v)$ and summing gives $e^\top V = \sum_v p_v h(v) \mathbb{E}[R \mid s, v] = \mathbb{E}[h(y_t) R \mid s]$ by the law of total expectation. The expectation of $b(s)h(y_t)$ is zero. For the implemented leave-one-out baseline, condition on the current answer's visited state and the question. Each other answer is an independent trajectory from the same fixed policy, so its reward is independent of the current sampled action; the conditional product with $h(y_t)$ again has zero expectation. Other answers need not visit the same intermediate state. This last argument requires the generating distribution to equal the policy used to compute h. □

The leave-one-out construction supplies an action-independent baseline under these assumptions; it does not establish a general variance-reduction guarantee.

A.6 GRADIENT EQUIVALENCE OF THE DETACHED-TARGET OBJECTIVE

Lemma 2 (Detached-target gradient equivalence)**.** *For a fixed global batch with M valid response positions and $D = Nq$ sampled answers, let c_{it} be defined by Eq.* (10)*. The gradient of the implemented token-mean KL to detached $\widetilde{Q}(c_{it})$ equals the gradient of*

$$\frac{1}{M} \sum_{i,t} D_{\mathrm{KL}}(P_{\theta,it}\|B_{it}) - \frac{\lambda}{D} \sum_{i,t} \mathrm{sg}[A_i h_{it}(y_{it})] P_{\theta,it}^\top \mathrm{sg}[\Delta_{it}]. \tag{14}$$

The extra negative logit gradient at (i,t) after global reduction is $\lambda D^{-1} \mathrm{sg}[A_i h_{it}(y_{it})] d_{it}$. The gradient equivalence also holds through a differentiable shared-parameter student evaluated on these fixed prefixes, by the chain rule; it is not a guarantee of improved return.

Proof. For each position, $D_{\mathrm{KL}}(P\|\widetilde{Q}(c)) = D_{\mathrm{KL}}(P\|Q_F) - cP^\top\Delta + \log Z(c)$, whereas $D_{\mathrm{KL}}(P\|B) = D_{\mathrm{KL}}(P\|Q_F) - P^\top\Delta$ because $B = \widetilde{Q}(1)$ and $Z(1) = 1$. Their difference is $-(c-1)P^\top\Delta + \log Z(c)$. The last term has zero student gradient because the target is detached. Substituting $c - 1 = \lambda(M/D)\,\mathrm{sg}[Ah]$ and dividing by M gives Eq. (14); Lemma 1 supplies its logit gradient. This is equality of gradients, not numerical losses. After global token-mean reduction, the extra term is D^{-1} times the sum of valid token contributions within each answer. It is an answer-average of token sums, not a length-invariant mean. This algebra is for the realized batch and does not assert an unbiased stochastic gradient under variable-length sampling. □

Interpretation and scope. For a local directional utility U and $\eta_{\text{step}} > 0$, the quadratic design $\max_\gamma\{\gamma U - \gamma^2/(2\eta_{\text{step}})\}$ yields $\gamma^* = \eta_{\text{step}} U$. This motivates a signed correction but does not prescribe the training coefficient λ. For a fixed state with independently updated logits, replacing the stochastic Ah by its conditional mean $U(s)$ makes an added infinitesimal update $\eta_{\text{step}} U(s) d$ contribute $\eta_{\text{step}} U(s)^2 \geq 0$ to the local first-order expected return relative to the uncalibrated FC update. This is a one-state calculation, not a nonnegative-benefit claim for shared parameters, simultaneous states, Adam, finite steps, or held-out answers. The actual M/D factor is a batch-reduction device, not a claim that every state uses a fixed deterministic η_{step}. For a concrete shared-parameter counterexample, let the reward logit gradient be $g = (1/6, 0, -1/6)$, the proposed direction be $d = (0.1, -0.1, 0)$, and the logit Jacobian for a scalar parameter be $J = (1, 2, 0)^\top$. Although $d^\top g = 1/60 > 0$, the inner product of the parameter reward gradient with the mapped rank-one correction has sign $(J^\top g)(J^\top d)(d^\top g) < 0$. Thus the local positive alignment does not survive arbitrary parameter sharing. Similarly, retaining Q_F as a factor in the calibrated target is an algebraic decomposition: it imposes neither a proximity bound to Q_F nor alignment of the total gradient with full-context distillation.

B Data, reward, and evaluation protocol

B.1 Training and implementation

The common training source is LoCoMo, with 140 training and 77 validation questions. The main training configuration uses $N = 4$ answers per question and a learning rate of 2×10^{-6} for three epochs (27 updates). We evaluate the final MCC-OPD checkpoint. Training uses the official verl framework (version 0.9.0), scaled dot-product attention (SDPA), up to 16 questions per minibatch, a prompt cap of 28,672 tokens, and a response cap of 4,096. Fresh answers are sampled at temperature one without top-k or top-p truncation; a complete sampled batch is used for one update. The local utility identity assumes exact agreement with the actor softmax, which is stronger than matching these nominal sampling settings.

Partition methods use the union of student top-64 and full-context teacher top-64 tokens, plus the true complementary probability mass. The log-probability response uses the actual sampled token probability, including when that token lies in the tail. Outcome calibration uses $N = 4$, $\alpha = 1$, $\lambda = 0.02$; main MCC-OPD uses $N = 4$, $\alpha = 1.5$. The uncalibrated controls share the corresponding N and α. Teacher-guided baselines use the same frozen 8B teacher. OPSD uses a frozen initial 1.7B self-teacher that additionally reads training references. ExOPD also scores a frozen initial 1.7B reference; GRPO uses this model as its KL reference. Teacher-SFT uses the offline supervision and optimization settings described in Appendix C.

B.2 LoCoMo

Our QA split contains 140 training, 77 validation, and 1,192 test questions from categories 1–4. The split is by question. The test set began with a caption-preserving pool of 1,213 questions. We excluded 21 whose answer or evidence could not be recovered reliably from the source conversations. For retained questions, we clarified wording, corrected references, and adjusted evidence pointers where the source supported the original intended question. These included plans described as completed events, uncertain dates or counts, and mismatched people or events. The evaluation set includes corrected date anchors for three temporal questions, with the corresponding model outputs regenerated using the corrected questions. Source conversation text and captions were not rewritten. The table label "LoCoMo" therefore refers to this specified QA protocol, not scores on the unmodified release.

Counterfactual teacher scoring removes annotated supporting turns and their captions from the teacher's ablated view, while the student always sees full memory. The modified answer-F1 training reward is computed on the extracted final answer after normalization. For category 1, it is the minimum of comma-separated reference-item coverage F1 and whole-answer F1; other categories use whole-answer F1, with category 3 references truncated at the first semicolon. Count words are normalized for count questions, and source-supported equivalent references are accepted. This training reward is distinct from evaluation F1.

Evaluation uses full conversations and greedy decoding, with at most 4,096 response tokens. For an initially empty extracted answer, the completion procedure retains at most 3,832 generated tokens, inserts the answer boundary, and allows up to 256 additional tokens within the response budget. EM uses normalized full-answer equality. BLEU-1 is the mean of unsmoothed sentence-level unigram BLEU with a 13a tokenizer; category 3 uses the reference prefix before its first semicolon.

B.3 REALTALK AND MUSIQUE

REALTALK (Lee et al., 2025) contains 728 evaluation questions from ten real conversations. Every method receives the full conversation, including timestamps and captions; no question is removed for length. The context window is 40,960 tokens with a 2,560-token response cap. Two temporal references were corrected against the source conversation, without rewriting the model input. These are supplied-history answering results, not retrieval scores.

For MuSiQue (Trivedi et al., 2022), we use 2,172 questions retained from the 2,417-question development release after source-grounded review. Recoverable question, reference, or evidence defects were repaired to preserve the intended question; unrecoverable cases were excluded. All supplied paragraphs, including distractors, remain in their original order. Neither gold answers nor annotated reasoning chains are given to the student. The context window is 12,288 tokens with a 4,096-token response cap. This reviewed evaluation set differs from the unmodified public release. Neither benchmark supplies training data for the main-table models.

B.4 SEMANTIC EVALUATION AND REWARD VARIANTS

Semantic accuracy uses one fixed rubric per benchmark and common question guidance across methods. Normalized exact matches to accepted references are accepted directly; empty answers are incorrect. Other distinct scoring inputs are judged with DeepSeek Flash at temperature zero. Two judgments are collected, with a third vote to resolve disagreement; identical inputs reuse their verdict. This score is separate from lexical metrics and from the original benchmark's reported scores.

The default outcome-calibration reward uses extracted-answer F1 and does not call a training judge. The semantic reward ablation instead uses binary correctness during training for both outcome calibration and GRPO. Its training scores and the held-out evaluation verdicts have separate inputs and roles. The semantic and F1 versions of GRPO are distinct trained models, reported separately in Table 6.

C BASELINE DEFINITIONS AND ADDITIONAL COMPARISONS

Full-context OPD objectives. The partition baseline targets Q_F with the reverse KL used by our methods. The sampled baseline freezes the old student and teacher log probabilities of each generated token and uses their difference as the advantage in a clipped policy surrogate. These objectives are not interchangeable implementations of the same finite-sample loss. Both use the full prompt and no answer reward.

Teacher-SFT. Starting from the same Qwen3-1.7B student, we supervise one fixed Qwen3-8B answer for each of the 140 LoCoMo training questions. The teacher uses greedy decoding with the full supplied memory and does not receive the reference answer. All 140 outputs are retained. We use LlamaFactory for full-parameter supervised fine-tuning with a learning rate of 10^{-5}, batch size 16, a 32,768-token window, and three epochs (27 updates), reporting the third-epoch model. These 140 fixed trajectories are reused across epochs; online $N = 4$ methods instead generate 1,680 fresh student answers.

ExOPD. We implement ExOPD's reward extrapolation (Yang et al., 2026a) with scale $\lambda_{\text{ex}} = 1.25$, the frozen 8B teacher T, and the initial 1.7B student as a frozen reference R. At each sampled token, the detached advantage is

$$a_t^{\text{ex}} = \lambda_{\text{ex}}(\log T_t - \log R_t) - (\log S_{\text{old},t} - \log R_t). \tag{15}$$

All three models receive the same full prompt and student prefix. The update uses the sampled-token policy surrogate, lower/upper clipping thresholds 0.2/0.28, and global token-mean reduction. It uses no evidence deletion or answer-level reward. The setting follows our common $N = 4$ training budget; scoring the additional reference means that equal rollout counts do not imply equal computation.

CROP and its controls. We implement CROP's paper-defined selector (Li et al., 2026) with original, meaning-preserving paraphrase, and condition-changing counterfactual prompts. Its selection score subtracts paraphrase JSD from counterfactual JSD, retaining the highest-scoring 20% of positions within each answer. Teacher top-16 returned tokens and one residual bucket define each selection JSD; an unreturned token's mass is not recovered with an additional query. The update target remains the full-context teacher distribution, unlike our FC target extrapolation. The clipped sampled update uses lower/upper clip thresholds 0.2/0.28. The CROP control without paraphrase correction removes the paraphrase subtraction; random selection retains 20% of positions with the same rounding. All training questions retain their original full prompt and reference. The view construction was independently reviewed; one retained source-ambiguous item did not establish a strict answer flip. We therefore do not claim that every constructed triplet satisfies a fully verified answer-changing intervention.

SA-OPD selectors. SA-OPD (Jiang et al., 2026) compares the teacher–student sampled log-probability gap with and without the task prompt. The *paper selector* setting applies its batch-level absolute-gap rule with selection fractions 0.2 and 0.3. The *code selector* uses per-answer dual-tail filtering and an absolute-gap removal budget in $[0.004, 0.012]$, following the published selector code. Both use symmetric policy clipping at 0.2, dual clipping at 3, and update-gap clipping at 10. A separate all-token SA-OPD control preserves these update settings. We align the no-prompt teacher and student with the same initial boundary token, freeze selection over the complete batch, and reduce the loss by the number of retained tokens. These are documented adaptations to the common training backend, not claims to reproduce the original paper's benchmark scores or entire training stack.

C.1 Matched component changes

Table 4 reports all four evaluation metrics for adding the evidence contrast and for calibrating it. Each difference holds the contrast strength and number of sampled answers fixed. The final block supplies the four-benchmark matched Judge and F1 differences shown in Figure 3(c).

Table 4: **Matched component changes**, in percentage points. Each difference holds the sampling count and contrast strength fixed. All blocks use $N = 4$. The first two use $\alpha = 1$; the last uses $\alpha = 1.5$. OPD denotes the partition-KL baseline; Uncalibrated uses the FC target with $c = 1$.

Comparison	Benchmark	Judge	F1	EM	BLEU-1
Uncalibrated − OPD	LoCoMo	+1.43	+3.01	+2.52	+2.00
	REALTALK	+2.61	+2.16	+1.10	+1.04
	MuSiQue	+1.66	+1.92	+1.52	+2.25
	LongMemEval	+1.60	+3.45	+2.40	+2.83
Outcome calibration − Uncalibrated	LoCoMo	−0.42	+0.08	+0.84	+0.93
	REALTALK	−0.27	−1.42	−0.27	−0.92
	MuSiQue	−0.05	+0.08	+0.60	+0.35
	LongMemEval	−1.00	+0.26	+1.00	+0.89
MCC-OPD − Uncalibrated	LoCoMo	+0.76	+0.70	+0.08	+1.10
	REALTALK	+1.10	−0.66	−0.27	+0.03
	MuSiQue	+2.95	+2.95	+2.76	+2.63
	LongMemEval	+1.00	+3.00	+3.81	+2.28

Student utility as a calibration signal. Outcome calibration estimates how a local counterfactual update would change the student's expected answer reward. We call this first-order change *student utility*. The estimator combines token log-probability responses with leave-one-out reward advantages and can strengthen, weaken, or reverse the correction. MCC-OPD controls the target shift

Table 5: **Calibration strategies at $N = 4$.** Outcome calibration uses F1 rewards, $\alpha = 1$, and $\lambda = 0.02$; MCC-OPD uses no outcome rewards and $\alpha = 1.5$. Matched component comparisons at each contrast strength appear in Table 4. Values are percentages.

		LoCoMo				REALTALK			
Method	N	**Judge**	**F1**	**EM**	**B1**	**Judge**	**F1**	**EM**	**B1**
Outcome calibration	4	62.67	48.27	20.89	32.53	26.24	23.49	5.08	9.99
MCC-OPD (ours)	4	63.17	47.97	20.89	32.87	26.37	23.66	4.95	10.76
		MuSiQue				**LongMemEval**			
Method	N	**Judge**	**F1**	**EM**	**B1**	**Judge**	**F1**	**EM**	**B1**
Outcome calibration	4	37.98	33.44	22.38	29.97	46.49	30.27	14.03	22.59
MCC-OPD (ours)	4	39.46	35.31	24.03	31.63	48.10	31.64	15.63	23.30

using only the paired teacher distributions. Appendix A.3 gives the utility derivation and its assumptions; Appendix G examines the estimator's implementation through fixed-prefix diagnostics.

Calibration-strategy comparison. MCC-OPD achieves higher LLM-judge accuracy than outcome calibration on all four benchmarks without using answer-level rewards (Table 5). On MuSiQue and LongMemEval, the advantage spans all four metrics. Outcome calibration has 0.30-point higher LoCoMo F1 and 0.14-point higher REALTALK EM. Both strategies use $N = 4$ at their chosen contrast strengths; Table 4 isolates the effect of calibration by holding the contrast strength fixed for each strategy.

C.2 Reward and selector comparisons

Reward and update rule. With the same binary correctness reward and $N = 8$, outcome calibration reaches Judge accuracies of 62.42%, 25.14%, and 39.46% on LoCoMo, REALTALK, and MuSiQue. These exceed GRPO by 2.94, 2.61, and 2.21 points, respectively. The reward choice also matters: F1-reward GRPO improves lexical scores over its semantic-reward version and reaches 39.04% Judge accuracy on MuSiQue. Table 6 reports both reward settings for each method.

Outcome diversity limits the utility signal. When all answers in a question group receive the same reward, their leave-one-out advantages vanish and outcome calibration reduces to uncalibrated FC. In separate $N = 8$ runs with binary semantic rewards, fewer than a third of question groups contain different rewards (Figure 10). Distillation still supplies supervision for the remaining groups. MCC-OPD computes its coefficient directly from the teacher distributions, so its calibration does not depend on variation in answer rewards.

Table 6: **Reward and update-rule ablation.** All four models use LoCoMo training and $N = 8$; results are percentages. Outcome calibration uses the same $\alpha = 1$, $\lambda = 0.02$ across reward settings.

Method	Benchmark	Judge	F1	EM	BLEU-1
Outcome calibration, F1 reward	LoCoMo	63.51	48.32	20.39	32.33
	REALTALK	24.73	23.78	5.08	9.96
	MuSiQue	38.12	33.83	22.33	30.10
Outcome calibration, semantic reward	LoCoMo	62.42	48.16	21.14	32.81
	REALTALK	25.14	23.77	5.22	9.92
	MuSiQue	39.46	34.52	22.97	30.71
GRPO, F1 reward	LoCoMo	59.82	45.41	19.63	30.57
	REALTALK	24.04	23.20	5.63	9.77
	MuSiQue	39.04	34.58	23.20	31.07
GRPO, semantic reward	LoCoMo	59.48	38.84	13.51	25.13
	REALTALK	22.53	22.09	3.02	9.02
	MuSiQue	37.25	31.93	20.58	28.14

Table 7: **Complete selector controls.** All models use the same training source and $N = 8$. The all-token control retains SA-OPD update settings; sampled OPD is the control for the CROP family. Values are percentages.

Method	Benchmark	Judge	F1	EM	BLEU-1
OPD (sampled tokens)	LoCoMo	61.33	45.69	19.30	30.22
	REALTALK	23.49	22.23	5.22	10.04
	MuSiQue	36.51	32.14	21.82	28.61
Random 20%	LoCoMo	56.54	42.59	18.79	29.54
	REALTALK	21.43	19.84	4.81	8.19
	MuSiQue	37.20	33.72	23.34	30.45
CROP (no paraphrase correction)	LoCoMo	59.14	44.93	19.38	29.69
	REALTALK	23.08	22.85	5.36	10.27
	MuSiQue	34.90	32.07	22.15	29.04
CROP	LoCoMo	59.90	46.45	20.55	30.56
	REALTALK	23.35	21.91	5.91	9.54
	MuSiQue	37.20	34.03	23.57	30.63
SA-OPD (all tokens)	LoCoMo	61.83	46.70	19.04	31.61
	REALTALK	24.31	23.77	5.49	10.46
	MuSiQue	37.15	32.60	21.55	28.87
SA-OPD (paper selector)	LoCoMo	61.74	45.87	18.04	30.15
	REALTALK	25.14	23.60	4.95	10.93
	MuSiQue	36.14	31.89	20.40	27.86
SA-OPD (code selector)	LoCoMo	61.58	45.61	17.95	29.72
	REALTALK	24.59	23.39	5.49	10.37
	MuSiQue	36.46	32.45	21.87	28.92

C.3 Contrast strength

Figure 6 reports the five evaluated MCC-OPD strengths at $N = 4$: $\alpha \in \{0.75, 1, 1.25, 1.5, 2\}$. The highlighted points use the main setting $\alpha = 1.5$. Each benchmark and metric is shown separately, retaining their different responses to the contrast strength.

C.4 Number of sampled answers

Table 8 compares four and eight answers per question at each of two fixed contrast strengths. All rows use the same LoCoMo source, three epochs, and final checkpoint; the $\alpha = 1.5$, $N = 4$ row is the MCC-OPD model in Table 1. Changing N changes the sampled training trajectories and token budget.

Table 8: **MCC-OPD rollout-count ablation.** Values are percentages after the common empty-answer completion procedure. Within each MCC-OPD pair, only the rollout count N changes in the method configuration. The OPD (partition KL) row provides the common $N = 8$ baseline.

	LoCoMo				REALTALK				MuSiQue			
Method	Judge	F1	EM	B1	Judge	F1	EM	B1	Judge	F1	EM	B1
OPD (partition KL), $N = 8$	61.49	46.12	19.04	30.97	25.55	24.13	5.08	10.70	35.96	31.56	20.49	27.92
MCC-OPD, $\alpha = 1$, $N = 4$	62.25	47.57	20.22	31.88	26.51	25.00	5.63	10.86	38.77	34.56	23.02	31.09
MCC-OPD, $\alpha = 1$, $N = 8$	62.16	46.88	19.71	30.78	25.00	24.08	5.77	10.79	38.54	34.24	22.97	30.67
MCC-OPD, $\alpha = 1.5$, $N = 4$	63.17	47.97	20.89	32.87	26.37	23.66	4.95	10.76	39.46	35.31	24.03	31.63
MCC-OPD, $\alpha = 1.5$, $N = 8$	62.08	47.30	19.88	31.96	25.82	24.54	5.36	11.24	37.48	32.92	22.05	29.37

At $\alpha = 1.5$, moving from $N = 4$ to $N = 8$ changes the equally weighted three-benchmark means by -1.21, -0.73, -0.86, and -0.89 points for Judge, F1, EM, and BLEU-1, respectively. The MuSiQue Judge difference is -1.98 points, with a paired-question 95% bootstrap interval of $[-3.18, -0.78]$. LoCoMo and REALTALK resampling uses their ten source conversations; their

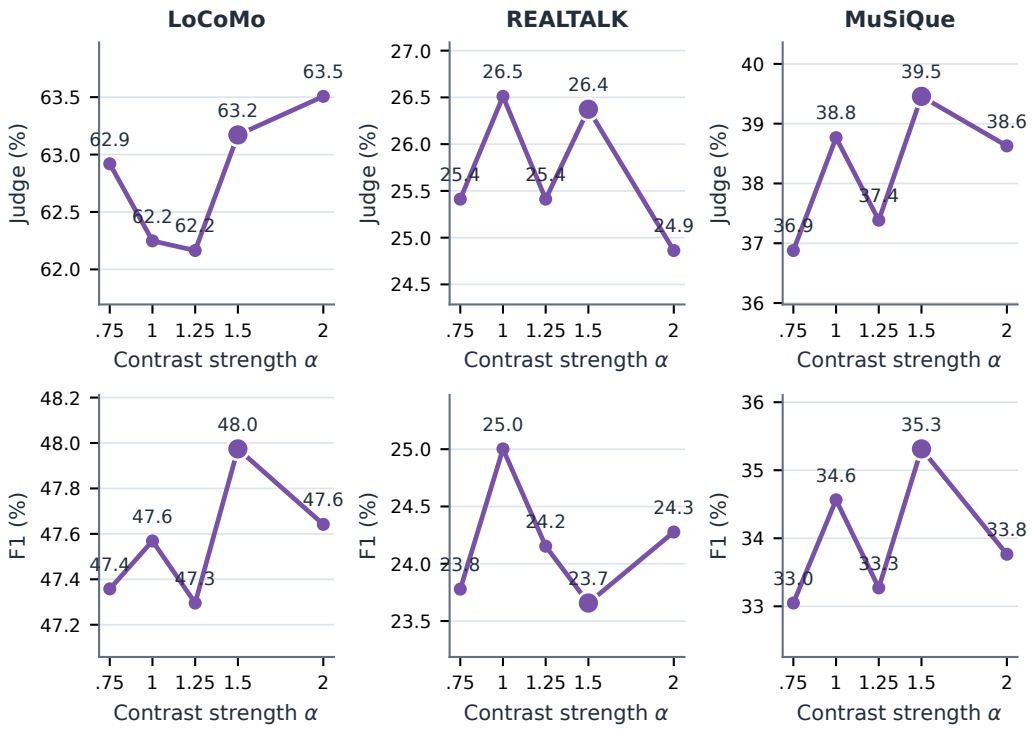

Figure 6: MCC-OPD strength sensitivity with $N = 4$ and the final checkpoint after 27 updates for every configuration. Every marker is an evaluated configuration; lines connect observations. The larger marker denotes the main $\alpha = 1.5$ configuration. Benchmarks and metrics are plotted separately.

Table 9: **Additional results with eight sampled answers.** All rows use $N = 8$ and update 27; the corresponding $N = 4$ results appear in Tables 1 and 5. Judge is LLM-judge accuracy and B1 is BLEU-1; values are percentages.

	LoCoMo				REALTALK			
Method	**Judge**	**F1**	**EM**	**B1**	**Judge**	**F1**	**EM**	**B1**
OPD (partition KL)	61.49	46.12	19.04	30.97	25.55	24.13	5.08	10.70
OPD (sampled tokens)	61.33	45.69	19.30	30.22	23.49	22.23	5.22	10.04
CROP	59.90	46.45	20.55	30.56	23.35	21.91	5.91	9.54
SA-OPD (paper selector)	61.74	45.87	18.04	30.15	25.14	23.60	4.95	10.93
SA-OPD (code selector)	61.58	45.61	17.95	29.72	24.59	23.39	5.49	10.37
Outcome calibration	63.51	48.32	20.39	32.33	24.73	23.78	5.08	9.96
	MuSiQue				**LongMemEval**			
Method	**Judge**	**F1**	**EM**	**B1**	**Judge**	**F1**	**EM**	**B1**
OPD (partition KL)	35.96	31.56	20.49	27.92	47.09	27.41	11.62	19.83
OPD (sampled tokens)	36.51	32.14	21.82	28.61	44.69	27.54	12.22	20.03
CROP	37.20	34.03	23.57	30.63	45.69	28.55	14.03	20.92
SA-OPD (paper selector)	36.14	31.89	20.40	27.86	47.29	27.95	10.82	19.65
SA-OPD (code selector)	36.46	32.45	21.87	28.92	47.09	30.76	14.83	22.87
Outcome calibration	38.12	33.83	22.33	30.10	48.50	31.27	15.03	22.99

Judge intervals cross zero. These 20,000-resample intervals condition on the fitted models. The twelve positive MCC-OPD $N = 8$ versus partition-OPD differences are point estimates, not twelve established significant improvements.

Baseline rollout counts. Table 9 retains the corresponding $N = 8$ baselines. Changing the rollout count affects methods differently: averaged over four benchmarks, moving from $N = 8$ to $N = 4$ changes Judge accuracy by $+0.37$ points for sampled-token OPD and -1.15 points for CROP.

Observed training cost. Table 10 compares the three core methods separately at $N = 4$ and $N = 8$. All runs use 140 LoCoMo training questions, three epochs, 27 updates, and eight A100 40GB GPUs. MCC-OPD uses $\alpha = 1.5$; outcome calibration uses $\alpha = 1$ and $\lambda = 0.02$.

Table 10: **Observed training cost at fixed rollout counts.** Methods are compared within the $N = 4$ and $N = 8$ groups. Each row reports one completed run on eight A100 40GB GPUs. Total time spans worker startup through exit, including initialization, validation, and checkpointing. Step time sums the logged training-step durations, including rollout and scoring.

Method	**Teacher views**	**Answers**	**Response tokens**	**Total min**	**Step min**
$N = 4$ answers per question					
OPD (partition KL)	1	1,680	127,278	53.22	34.63
Outcome calibration	2	1,680	132,560	65.71	45.35
MCC-OPD (ours)	2	1,680	124,943	75.70	47.16
$N = 8$ answers per question					
OPD (partition KL)	1	3,360	259,906	84.63	60.99
Outcome calibration	2	3,360	272,194	106.62	84.08
MCC-OPD (ours)	2	3,360	265,966	96.40	86.78

Within each group, the methods generate the same number of answers; response lengths and teacher computations still differ. Outcome calibration and MCC-OPD score two teacher views, compared with one for OPD (partition KL). CROP scores three, while SA-OPD also evaluates a no-prompt student view. Equal rollout counts therefore do not imply equal FLOPs. Neither calibration adds inference-time teacher calls. The shared empty-answer completion procedure contributes additional evaluation work, as described in Appendix B.

For MCC-OPD at $\alpha = 1.5$, increasing N from four to eight raises the response-token count from 124,943 to 265,966 and total time from 75.70 to 96.40 minutes. The token and wall-time ratios are 2.13 and 1.27, respectively. Figure 7 places these changes alongside the task results. The $\alpha = 1$, $N = 8$ run uses A800 hardware, so its time is not compared with the A100 runs.

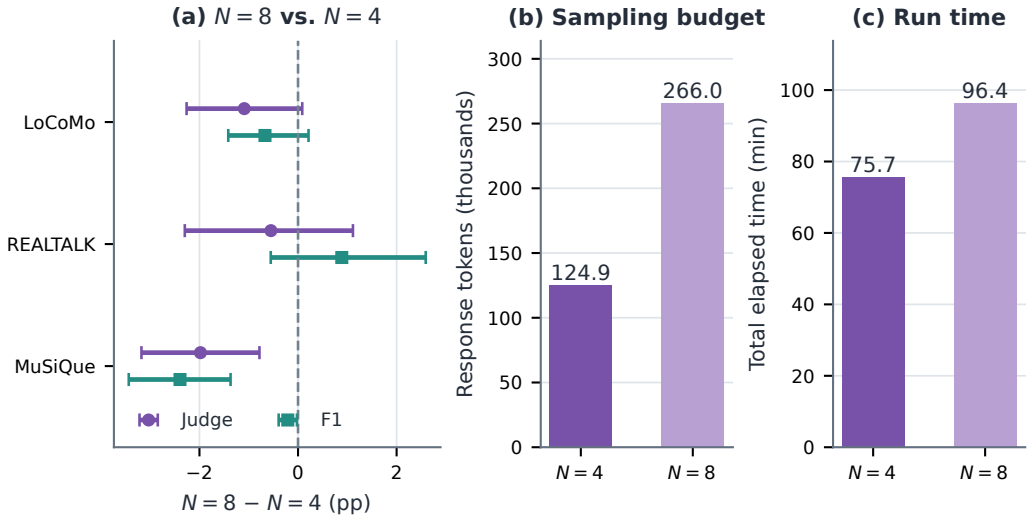

Figure 7: MCC-OPD sampling-number ablation at $\alpha = 1.5$. (a) Paired task differences with 95% intervals using the benchmark-specific units of the main component comparison. (b,c) Actual response tokens and total elapsed time on 8 A100 GPUs. Total elapsed time includes run-specific initialization, validation and saving; these two configurations are not a FLOPs-matched speedup experiment.

C.5 Calibration and training dynamics

Figure 8 shows the calibration coefficient, local direction change, validation F1, and generated response length across the recorded training runs. The validation curve contains only the three observed checkpoints at updates 9, 18, and 27; the reported test results use update 27 throughout. Direction and length statistics follow each run's own sampled trajectories.

C.6 Number of training questions

Figure 9 compares 35, 70, and 140 training questions. This experiment uses $\alpha = 1$ for MCC-OPD, separately from the main $\alpha = 1.5$ model. Holding three epochs fixed changes the update count from 9 to 15 to 27; the figure therefore describes the combined change in data coverage and optimization budget.

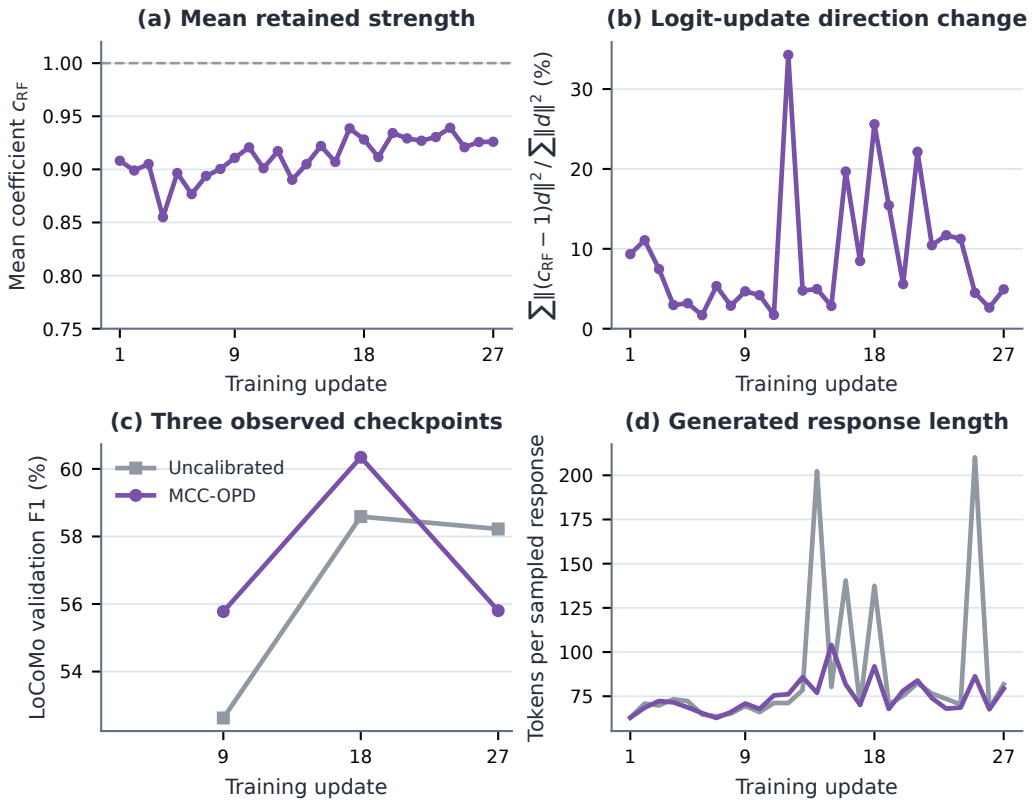

Figure 8: Training dynamics for matched Uncalibrated and MCC-OPD runs with $\alpha = 1.5$, $N = 4$. (a) Mean calibration coefficient c_{RF}. (b) Change in the incremental logit-update direction d, reported as $\sum\|(c_{\mathrm{RF}} - 1)d\|^2 / \sum\|d\|^2$ (percent), with both sums over valid tokens per update. The direction statistic is not a parameter-gradient norm or task utility. (c) All three observed validation checkpoints on 77 questions. (d) Actual mean generated response lengths; curves use each run's own trajectories.

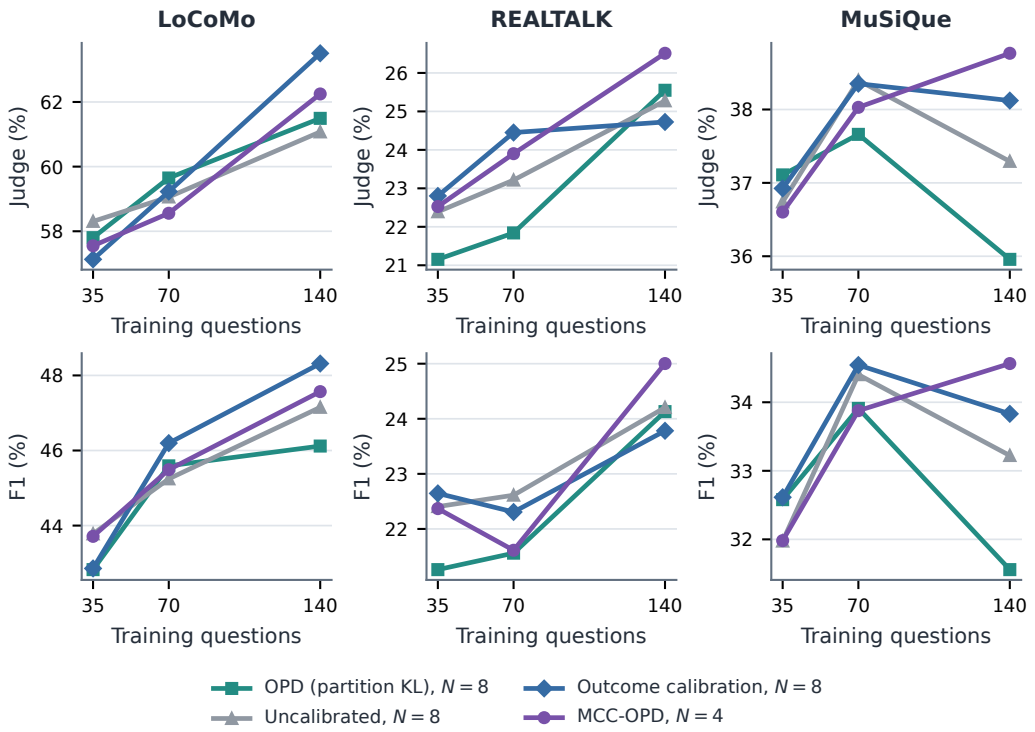

Figure 9: Data-size comparison at $\alpha = 1$ (including MCC-OPD), not the main $\alpha = 1.5$ setting. Three epochs use 9/15/27 updates for 35/70/140 questions. MCC-OPD uses $N = 4$ and the other methods $N = 8$. Thus both training coverage and update count vary, and methods are not compute matched. Metrics are shown separately by benchmark.

C.7 TEACHER AND STUDENT MODEL SIZES

We evaluate all four pairings of Qwen3-4B and Qwen3-8B teachers with Qwen3-0.6B and Qwen3-1.7B students. Every configuration trains on the same 140 LoCoMo questions with $N = 8$, a learning rate of 2×10^{-6}, and three epochs; we evaluate the final checkpoint after 27 updates. MCC-OPD uses $\alpha = 1.5$, while outcome calibration uses $\alpha = 1$, $\lambda = 0.02$, and F1 rewards. These method settings are held fixed across model sizes. Table 11 reports all four metrics on LoCoMo, REALTALK, and MuSiQue. Means below weight the three benchmarks equally.

Student size at a fixed teacher size. Increasing the student from 0.6B to 1.7B improves Judge accuracy on every benchmark for both calibration rules. MCC-OPD's mean Judge accuracy rises by 10.23 points with the 8B teacher and 8.65 points with the 4B teacher; the corresponding gains for outcome calibration are 10.55 and 9.24 points. Mean F1, EM, and BLEU-1 also improve in all four comparisons.

Teacher size at a fixed student size. A larger teacher does not uniformly improve the distilled student. For MCC-OPD, changing the teacher from 8B to 4B raises MuSiQue Judge accuracy from 17.27% to 22.93% for the 0.6B student and from 37.48% to 39.27% for the 1.7B student. Outcome calibration shows the same direction of change. Both methods also achieve higher mean Judge accuracy with the 4B teacher, but lower mean EM and BLEU-1. The preferred teacher size therefore depends on the evaluation metric.

Calibration across model sizes. With the 8B teacher and 0.6B student, MCC-OPD exceeds outcome calibration on 11 of 12 metrics, although their mean Judge accuracies are nearly identical (31.56% versus 31.57%). With the 4B teacher and 1.7B student, outcome calibration leads on 10 of 12 metrics. MCC-OPD's mean Judge differences relative to outcome calibration are -0.32, -0.01, -0.23, and $+0.36$ points for the 8B$\rightarrow$1.7B, 8B$\rightarrow$0.6B, 4B$\rightarrow$1.7B, and 4B$\rightarrow$0.6B pairings, respectively. These results show that the relative performance of the calibration rules depends on the teacher–student pairing.

Table 11: **Teacher and student model-size comparison.** All models are from the Qwen3 family and use $N = 8$ and the final checkpoint after 27 updates. MCC-OPD uses $\alpha = 1.5$; outcome calibration uses $\alpha = 1$ and $\lambda = 0.02$. Judge is LLM-judge accuracy and B1 is BLEU-1; values are percentages. Bold marks the higher score within each teacher–student pairing.

LoCoMo						
Teacher	**Student**	**Method**	**Judge**	**F1**	**EM**	**B1**
8B	1.7B	Outcome calibration	**63.51**	**48.32**	**20.39**	**32.33**
8B	1.7B	**MCC-OPD (ours)**	62.08	47.30	19.88	31.96
8B	0.6B	Outcome calibration	55.87	43.01	17.45	28.84
8B	0.6B	**MCC-OPD (ours)**	**56.12**	**44.55**	**18.88**	**30.27**
4B	1.7B	Outcome calibration	**62.84**	**48.76**	**20.72**	**26.71**
4B	1.7B	**MCC-OPD (ours)**	62.33	48.43	19.88	24.47
4B	0.6B	Outcome calibration	55.45	41.95	**14.60**	**27.73**
4B	0.6B	**MCC-OPD (ours)**	**56.29**	**42.22**	14.35	26.89
REALTALK						
Teacher	**Student**	**Method**	**Judge**	**F1**	**EM**	**B1**
8B	1.7B	Outcome calibration	24.73	23.78	5.08	9.96
8B	1.7B	**MCC-OPD (ours)**	**25.82**	**24.54**	**5.36**	**11.24**
8B	0.6B	Outcome calibration	**21.98**	19.86	3.16	9.20
8B	0.6B	**MCC-OPD (ours)**	21.29	**20.14**	**4.12**	**9.50**
4B	1.7B	Outcome calibration	24.86	**24.20**	**4.40**	**8.96**
4B	1.7B	**MCC-OPD (ours)**	**25.55**	23.73	3.98	8.06
4B	0.6B	Outcome calibration	**22.66**	19.99	2.06	8.64
4B	0.6B	**MCC-OPD (ours)**	21.98	**20.73**	**2.75**	**8.87**
MuSiQue						
Teacher	**Student**	**Method**	**Judge**	**F1**	**EM**	**B1**
8B	1.7B	Outcome calibration	**38.12**	**33.83**	**22.33**	**30.10**
8B	1.7B	**MCC-OPD (ours)**	37.48	32.92	22.05	29.37
8B	0.6B	Outcome calibration	16.85	14.83	6.17	12.74
8B	0.6B	**MCC-OPD (ours)**	**17.27**	**15.44**	**6.77**	**13.63**
4B	1.7B	Outcome calibration	**40.15**	**33.67**	20.58	**28.60**
4B	1.7B	**MCC-OPD (ours)**	39.27	33.22	**21.09**	28.30
4B	0.6B	Outcome calibration	22.01	**17.38**	**7.14**	**13.24**
4B	0.6B	**MCC-OPD (ours)**	**22.93**	17.18	7.00	12.71

D RESULTS BY QUESTION CATEGORY

We report all four metrics for each question category in the four main benchmarks. These use the same checkpoints and answer processing as Tables 1 and 5. The rollout count N is shown for online methods; Teacher-SFT uses 140 fixed offline answers. The categories partition each benchmark; their question counts sum to 1,192, 728, 2,172, and 499, respectively. All scores are percentages.

D.1 LOCOMO

Table 12: LoCoMo results by question category. n is the number of questions; N is the training rollout count.

Method	N	Judge	F1	EM	BLEU-1
Single-hop ($n = 662$)					
Qwen3-1.7B (Student)	–	72.36	40.25	17.07	26.38
Qwen3-8B (Teacher)	–	84.89	63.22	36.86	45.53
Teacher-SFT	–	77.64	53.22	28.10	38.12
OPD (partition KL)	4	78.85	52.03	25.23	36.42
OPD (sampled tokens)	4	76.89	53.63	27.95	37.03
CROP	4	74.32	53.02	29.91	38.62
SA-OPD (paper selector)	4	78.40	53.81	28.10	37.96
SA-OPD (code selector)	4	76.59	52.54	27.49	38.49
ExOPD	4	77.95	54.91	29.76	39.04
OPSD	8	73.56	41.85	20.09	28.40
GRPO (semantic reward)	8	76.28	44.09	19.94	30.40
Outcome calibration	4	81.87	57.80	29.91	41.59
MCC-OPD (ours)	4	81.72	57.27	30.06	41.76
Multi-hop ($n = 206$)					
Qwen3-1.7B (Student)	–	52.91	27.04	3.40	11.58
Qwen3-8B (Teacher)	–	61.65	37.20	5.34	16.55
Teacher-SFT	–	56.31	31.52	4.85	14.77
OPD (partition KL)	4	59.22	36.53	9.22	17.35
OPD (sampled tokens)	4	54.37	34.95	10.68	17.57
CROP	4	50.49	32.18	10.19	15.48
SA-OPD (paper selector)	4	58.74	34.62	9.22	17.87
SA-OPD (code selector)	4	56.31	35.00	9.71	17.86
ExOPD	4	54.85	33.21	8.74	16.86
OPSD	8	53.88	29.57	5.34	13.46
GRPO (semantic reward)	8	56.80	28.16	5.34	13.01
Outcome calibration	4	54.85	37.29	10.19	18.32
MCC-OPD (ours)	4	58.74	36.74	10.68	18.89
Temporal ($n = 255$)					
Qwen3-1.7B (Student)	–	21.96	34.69	3.53	22.22
Qwen3-8B (Teacher)	–	33.33	46.92	6.67	29.12
Teacher-SFT	–	26.27	41.84	7.84	25.04
OPD (partition KL)	4	22.35	40.05	5.10	25.35
OPD (sampled tokens)	4	22.35	38.80	4.71	23.66
CROP	4	18.04	38.68	5.10	23.93
SA-OPD (paper selector)	4	21.57	38.71	5.10	24.43
SA-OPD (code selector)	4	21.18	40.55	3.92	24.17
ExOPD	4	21.57	39.63	7.06	25.64
OPSD	8	21.18	38.31	5.49	25.19
GRPO (semantic reward)	8	21.18	39.30	4.71	25.40
Outcome calibration	4	25.10	39.87	7.45	25.95
MCC-OPD (ours)	4	24.31	39.53	7.06	26.06
Open-domain ($n = 69$)					
Qwen3-1.7B (Student)	–	49.28	15.11	1.45	6.71
Qwen3-8B (Teacher)	–	53.62	23.66	8.70	11.86
Teacher-SFT	–	53.62	19.66	7.25	10.13
OPD (partition KL)	4	49.28	24.26	14.49	16.62

Continued on next page

Table 12 (continued)

Method	N	Judge	F1	EM	BLEU-1
OPD (sampled tokens)	4	40.58	17.71	11.59	11.81
CROP	4	47.83	25.24	18.84	16.65
SA-OPD (paper selector)	4	43.48	17.82	11.59	11.44
SA-OPD (code selector)	4	44.93	20.22	13.04	12.46
ExOPD	4	43.48	18.61	13.04	12.50
OPSD	8	40.58	21.10	11.59	13.07
GRPO (semantic reward)	8	47.83	18.73	8.70	9.70
Outcome calibration	4	40.58	20.74	15.94	12.48
MCC-OPD (ours)	4	42.03	23.49	14.49	14.46

D.2 REALTALK

Table 13: REALTALK results by question category. n is the number of questions; N is the training rollout count.

Method	N	Judge	F1	EM	BLEU-1
Multi-hop ($n = 301$)					
Qwen3-1.7B (Student)	–	20.93	27.16	1.66	10.42
Qwen3-8B (Teacher)	–	27.91	34.39	5.65	14.34
Teacher-SFT	–	20.27	28.34	3.65	12.51
OPD (partition KL)	4	20.60	29.60	3.65	12.34
OPD (sampled tokens)	4	21.93	29.81	6.31	13.75
CROP	4	17.94	27.72	5.32	12.66
SA-OPD (paper selector)	4	22.26	30.48	5.65	14.29
SA-OPD (code selector)	4	22.26	30.08	5.32	13.50
ExOPD	4	18.27	30.26	5.65	13.24
OPSD	8	21.26	27.66	3.32	11.65
GRPO (semantic reward)	8	21.59	29.60	1.66	11.01
Outcome calibration	4	21.26	30.11	4.65	13.40
MCC-OPD (ours)	4	22.59	30.01	5.32	14.81
Temporal ($n = 319$)					
Qwen3-1.7B (Student)	–	19.12	15.11	2.51	6.57
Qwen3-8B (Teacher)	–	38.87	27.68	9.09	13.43
Teacher-SFT	–	25.71	19.10	3.76	9.03
OPD (partition KL)	4	22.88	17.62	5.33	8.57
OPD (sampled tokens)	4	21.94	16.82	4.39	7.02
CROP	4	22.26	16.88	5.64	7.05
SA-OPD (paper selector)	4	21.00	16.79	4.70	6.96
SA-OPD (code selector)	4	22.57	16.38	4.39	6.36
ExOPD	4	22.57	17.51	6.90	7.51
OPSD	8	19.12	16.09	4.70	7.18
GRPO (semantic reward)	8	17.87	16.09	4.08	7.47
Outcome calibration	4	25.71	17.99	5.64	7.55
MCC-OPD (ours)	4	24.76	17.72	3.76	6.94
Commonsense ($n = 108$)					
Qwen3-1.7B (Student)	–	37.04	14.34	0.93	6.09
Qwen3-8B (Teacher)	–	39.81	18.88	3.70	7.82
Teacher-SFT	–	30.56	15.48	0.93	5.62
OPD (partition KL)	4	36.11	18.75	2.78	6.82
OPD (sampled tokens)	4	40.74	22.46	6.48	8.10
CROP	4	35.19	20.00	6.48	7.45
SA-OPD (paper selector)	4	40.74	22.03	6.48	7.49
SA-OPD (code selector)	4	37.96	20.77	4.63	8.69
ExOPD	4	37.04	19.54	5.56	7.92
OPSD	8	39.81	18.10	3.70	8.92
GRPO (semantic reward)	8	38.89	18.86	3.70	8.07
Outcome calibration	4	41.67	21.27	4.63	7.67

Continued on next page

Table 13 (continued)

Method	N	Judge	F1	EM	BLEU-1
MCC-OPD (ours)	4	41.67	23.50	7.41	10.78

D.3 MuSiQue

Table 14: MuSiQue results by question category. n is the number of questions; N is the training rollout count.

Method	N	Judge	F1	EM	BLEU-1
2-hop ($n = 1185$)					
Qwen3-1.7B (Student)	–	35.44	29.49	17.05	25.30
Qwen3-8B (Teacher)	–	55.19	50.15	37.64	46.17
Teacher-SFT	–	31.39	28.48	19.07	25.40
OPD (partition KL)	4	34.43	31.56	20.08	27.74
OPD (sampled tokens)	4	35.70	32.12	21.18	28.90
CROP	4	35.86	33.54	22.95	30.39
SA-OPD (paper selector)	4	35.78	32.26	21.10	28.80
SA-OPD (code selector)	4	35.78	32.62	20.51	29.08
ExOPD	4	36.37	34.30	23.21	30.87
OPSD	8	38.73	33.47	21.52	29.44
GRPO (semantic reward)	8	36.96	32.18	20.42	28.35
Outcome calibration	4	37.22	33.57	22.28	30.32
MCC-OPD (ours)	4	38.40	34.81	22.87	31.36
3-hop ($n = 672$)					
Qwen3-1.7B (Student)	–	39.29	30.24	18.01	25.79
Qwen3-8B (Teacher)	–	52.23	46.85	36.31	43.16
Teacher-SFT	–	33.93	29.76	19.35	25.51
OPD (partition KL)	4	39.43	32.28	21.28	27.97
OPD (sampled tokens)	4	38.39	33.15	23.07	29.55
CROP	4	38.69	33.20	23.51	29.83
SA-OPD (paper selector)	4	40.03	34.61	24.70	31.39
SA-OPD (code selector)	4	40.62	34.27	24.40	30.87
ExOPD	4	38.10	33.46	23.81	30.14
OPSD	8	41.82	32.42	20.39	28.35
GRPO (semantic reward)	8	38.69	32.36	21.13	28.71
Outcome calibration	4	39.73	33.50	23.07	29.92
MCC-OPD (ours)	4	42.11	37.51	27.23	33.85
4-hop ($n = 315$)					
Qwen3-1.7B (Student)	–	36.51	28.80	17.14	24.30
Qwen3-8B (Teacher)	–	59.68	50.09	37.46	45.90
Teacher-SFT	–	27.62	25.79	14.60	21.33
OPD (partition KL)	4	37.14	29.25	18.73	24.73
OPD (sampled tokens)	4	36.51	30.73	20.95	27.09
CROP	4	33.65	30.89	19.37	27.68
SA-OPD (paper selector)	4	34.92	30.62	19.37	26.12
SA-OPD (code selector)	4	34.60	30.01	18.73	25.96
ExOPD	4	34.29	29.55	19.68	25.33
OPSD	8	35.24	30.32	21.27	26.71
GRPO (semantic reward)	8	35.24	30.06	20.00	26.10
Outcome calibration	4	37.14	32.85	21.27	28.71
MCC-OPD (ours)	4	37.78	32.52	21.59	27.88

D.4 LongMemEval

Table 15: LongMemEval results by question category. n is the number of questions; N is the training rollout count.

Method	N	Judge	F1	EM	BLEU-1
Single-session user ($n = 70$)					
Qwen3-1.7B (Student)	–	74.29	26.23	7.14	15.08
Qwen3-8B (Teacher)	–	90.00	66.66	50.00	54.40
Teacher-SFT	–	74.29	59.64	42.86	48.19
OPD (partition KL)	4	82.86	38.95	15.71	26.49
OPD (sampled tokens)	4	80.00	43.80	21.43	31.46
CROP	4	80.00	46.86	24.29	35.44
SA-OPD (paper selector)	4	80.00	42.33	20.00	30.07
SA-OPD (code selector)	4	84.29	49.47	28.57	36.82
ExOPD	4	82.86	53.28	32.86	42.08
OPSD	8	75.71	29.06	10.00	18.76
GRPO (semantic reward)	8	78.57	25.79	4.29	14.84
Outcome calibration	4	82.86	45.92	22.86	34.51
MCC-OPD (ours)	4	78.57	48.96	28.57	37.72
Single-session assistant ($n = 56$)					
Qwen3-1.7B (Student)	–	87.50	28.52	5.36	17.48
Qwen3-8B (Teacher)	–	94.64	66.71	41.07	47.89
Teacher-SFT	–	82.14	61.72	41.07	49.53
OPD (partition KL)	4	82.14	49.34	28.57	35.78
OPD (sampled tokens)	4	85.71	50.57	30.36	38.87
CROP	4	82.14	44.76	23.21	30.78
SA-OPD (paper selector)	4	80.36	46.69	25.00	32.74
SA-OPD (code selector)	4	82.14	52.48	33.93	39.52
ExOPD	4	83.93	62.64	41.07	46.56
OPSD	8	82.14	32.24	7.14	19.31
GRPO (semantic reward)	8	85.71	32.57	3.57	19.57
Outcome calibration	4	83.93	51.23	28.57	37.71
MCC-OPD (ours)	4	89.29	56.69	33.93	42.49
Single-session preference ($n = 30$)					
Qwen3-1.7B (Student)	–	56.67	16.17	0.00	10.69
Qwen3-8B (Teacher)	–	66.67	16.57	0.00	11.04
Teacher-SFT	–	33.33	18.16	0.00	12.93
OPD (partition KL)	4	56.67	16.82	0.00	11.08
OPD (sampled tokens)	4	50.00	16.48	0.00	11.03
CROP	4	50.00	16.22	0.00	11.12
SA-OPD (paper selector)	4	53.33	16.20	0.00	10.32
SA-OPD (code selector)	4	46.67	17.98	0.00	11.58
ExOPD	4	43.33	14.43	0.00	9.37
OPSD	8	50.00	15.82	0.00	10.16
GRPO (semantic reward)	8	56.67	15.74	0.00	10.14
Outcome calibration	4	46.67	15.53	0.00	10.30
MCC-OPD (ours)	4	53.33	15.18	0.00	9.82
Multi-session ($n = 132$)					
Qwen3-1.7B (Student)	–	23.48	13.49	3.79	9.72
Qwen3-8B (Teacher)	–	46.97	28.54	18.94	23.52
Teacher-SFT	–	22.73	14.80	5.30	11.76
OPD (partition KL)	4	24.24	17.21	7.58	14.07
OPD (sampled tokens)	4	25.76	18.72	12.12	16.59
CROP	4	25.00	17.06	9.09	14.18
SA-OPD (paper selector)	4	24.24	17.02	9.09	13.60
SA-OPD (code selector)	4	26.52	20.21	12.88	17.71
ExOPD	4	26.52	20.33	13.64	18.31
OPSD	8	24.24	14.62	6.82	11.13
GRPO (semantic reward)	8	25.00	15.32	6.06	11.56
Outcome calibration	4	26.52	18.56	11.36	16.17
MCC-OPD (ours)	4	25.76	17.47	9.85	14.87

Continued on next page

Table 15 (continued)

Method	N	Judge	F1	EM	BLEU-1
Temporal reasoning ($n = 133$)					
Qwen3-1.7B (Student)	–	30.08	20.04	2.26	11.26
Qwen3-8B (Teacher)	–	57.89	40.46	16.54	28.90
Teacher-SFT	–	27.82	26.00	8.27	16.36
OPD (partition KL)	4	29.32	23.05	5.26	14.22
OPD (sampled tokens)	4	30.83	26.48	7.52	17.25
CROP	4	30.08	26.09	7.52	17.06
SA-OPD (paper selector)	4	34.59	27.12	7.52	17.13
SA-OPD (code selector)	4	28.57	25.16	7.52	15.94
ExOPD	4	29.32	25.21	6.77	16.77
OPSD	8	29.32	18.54	1.50	10.34
GRPO (semantic reward)	8	26.32	19.75	2.26	10.66
Outcome calibration	4	33.83	29.59	8.27	19.01
MCC-OPD (ours)	4	34.59	29.87	9.02	17.86
Knowledge update ($n = 78$)					
Qwen3-1.7B (Student)	–	48.72	12.55	1.28	7.03
Qwen3-8B (Teacher)	–	60.26	42.50	29.49	37.45
Teacher-SFT	–	43.59	31.25	14.10	21.96
OPD (partition KL)	4	47.44	24.65	11.54	18.94
OPD (sampled tokens)	4	51.28	27.63	12.82	22.19
CROP	4	51.28	26.57	15.38	21.50
SA-OPD (paper selector)	4	52.56	29.26	14.10	24.03
SA-OPD (code selector)	4	47.44	26.45	11.54	20.87
ExOPD	4	50.00	34.04	20.51	27.93
OPSD	8	46.15	20.19	7.69	14.54
GRPO (semantic reward)	8	46.15	16.07	3.85	10.09
Outcome calibration	4	42.31	27.83	15.38	22.73
MCC-OPD (ours)	4	50.00	31.44	17.95	25.31

E LongMemEval evaluation protocol

We evaluate 499 LongMemEval questions (Wu et al., 2025) with source-grounded QA corrections. None is used to train these LoCoMo-only models. Inputs contain 60 retrieved history entries per question; five questions receive one additional source user turn to restore missing evidence. Judging uses the benchmark's base rubric with shared question-specific guidance for twelve questions. This is a retrieved-and-supplemented evidence condition rather than the original full-history or oracle setting. All models receive the same inputs and share the applicable judging protocol.

Table 1 reports the four aggregate metrics on all 499 questions. Table 15 separates the six memory tasks; Table 16 additionally reports their unweighted mean and accuracy on the 30 unanswerable questions. Lexical scores remain supplementary diagnostics for LongMemEval.

The aggregate advantage does not extend to every memory task or to abstention. MCC-OPD's abstention accuracy is 46.67%, compared with 56.67% for the $N = 4$ OPD baseline. The category tables retain these differences alongside the overall scores.

F Diagnostic protocol and interpretation

Evidence deletion. The frozen replay uses all 140 training questions and eight existing answers each. Answers were generated during the third training epoch by successive policies; scoring freezes the final outcome-calibrated checkpoint's token partition. This diagnostic uses a separate $N = 8$ outcome-calibration run and does not observe its online training updates. The full and support-deleted teacher distributions are reused on the same token coordinates. The added control deletes non-support, caption-free turns while preserving all annotated support and captions. Re-tokenizing the complete prompts verifies exactly matched deletion lengths for every question (12–1,995 tokens). Only 74 questions also match the number of deleted turns; positions, speakers, and meaning are not matched. Non-support turns are not guaranteed to be irrelevant.

Table 16: **LongMemEval task and abstention summaries (499 questions).** Task avg. is the unweighted mean across six task accuracies; Abstain is accuracy on the 30 unanswerable questions. N is the training rollout count. Values are percentages.

Method	N	Task avg.	Abstain
Qwen3-1.7B (Student)	–	53.46	56.67
Qwen3-8B (Teacher)	–	69.41	60.00
Teacher-SFT	–	47.32	40.00
OPD (partition KL)	4	53.78	56.67
OPD (sampled tokens)	4	53.93	50.00
CROP	4	53.08	56.67
SA-OPD (paper selector)	4	54.18	50.00
SA-OPD (code selector)	4	52.60	56.67
ExOPD	4	52.66	53.33
OPSD	8	51.26	50.00
GRPO (semantic reward)	8	53.07	53.33
Outcome calibration	4	52.68	46.67
MCC-OPD (ours)	4	55.26	46.67

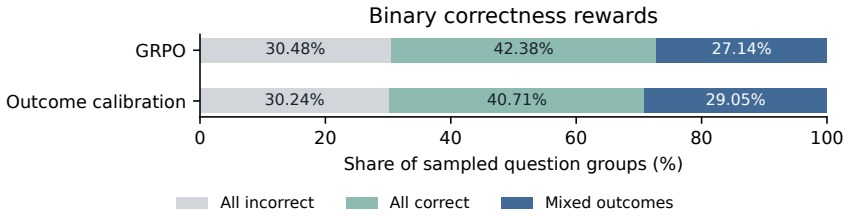

Figure 10: **Availability of semantic outcome contrasts.** Across each method's own 420 sampled question groups, only 27.14% (GRPO) and 29.05% (outcome calibration with semantic rewards) contain both reward values. Homogeneous groups provide no task-reward contrast. Outcome calibration then reduces to uncalibrated FC; this does not imply that GRPO's total gradient, including its regularization, is zero. This is a trajectory diagnostic, not a causal explanation of the accuracy gap.

Aggregation and semantic labels. JSD is computed on the common partition, not the full vocabulary. We first average positions within an answer, then answers within a question, then questions equally. Confidence intervals use 4,000 paired question bootstrap samples, not independent training runs. The existing semantic labels contain 477 correct and 136 wrong answers, 394 without an exact question–answer match in the archived semantic reviews (unknown), and 113 flagged for scoring-rule boundaries or source/reference ambiguity (boundary cases). One wrong answer has no usable answer region, leaving 135 wrong answers for the promotion diagnostic. Unknown and boundary cases are excluded from that diagnostic, rather than treated as wrong. All 9,158 usable answer tokens are explicit partition tokens, permitting exact token log-probability comparisons. Only eight questions contain both clearly correct and clearly wrong answers; this limits within-question correctness contrasts.

Target promotion. For each ablated view, the replay constructs the uncalibrated target $B \propto Q_F^2/Q_C$ at $\alpha = 1$. Promotion is the increase in mean log probability of the existing answer tokens relative to Q_F. At the 0.01-nat threshold, question-weighted promotion of wrong answers is 22.80% for support deletion and 25.27% for its control; the difference's 95% interval is $[-19.79, 15.36]$ percentage points. For correct answers, the corresponding rates are 13.49% and 23.78%, with difference interval $[-18.64, -2.10]$. Answer-weighted rates differ: support deletion promotes 23/135 wrong answers and 65/477 correct answers. We do not mix these weighting schemes in Figure 4. Sensitivity thresholds of near-zero and 0.1 nat likewise do not establish a reduction in wrong-answer promotion relative to the control. Promotion under a frozen target is not a causal change in a trained model's error rate.

Semantic reward availability. Figure 10 counts question groups separately on the complete training trajectories of outcome calibration and GRPO with binary correctness rewards. Each has 420 groups of eight answers. Mixed rewards occur in 29.05% of outcome-calibration groups and 27.14% of GRPO groups. The policies and trajectories differ, making this a descriptive comparison. Reward ties remove the outcome-calibration term, but leave FC distillation and do not imply a zero total GRPO gradient.

F.1 Mirror-budget replay on fixed prefixes

Figure 11 applies the MCC-OPD target rule at $\alpha = 1.5$ to 1,120 saved outcome-calibration trajectories from the 140 training questions. The replay covers 91,755 valid response tokens and retains each saved full/ablated teacher pair and shared partition. It yields $c_{\mathrm{RF}} = 1$ at 72.10% of positions; this is a different prefix distribution from the online MCC-OPD measurements in Figure 3. The panels describe the budget rule and its position-dependent behavior on these fixed inputs.

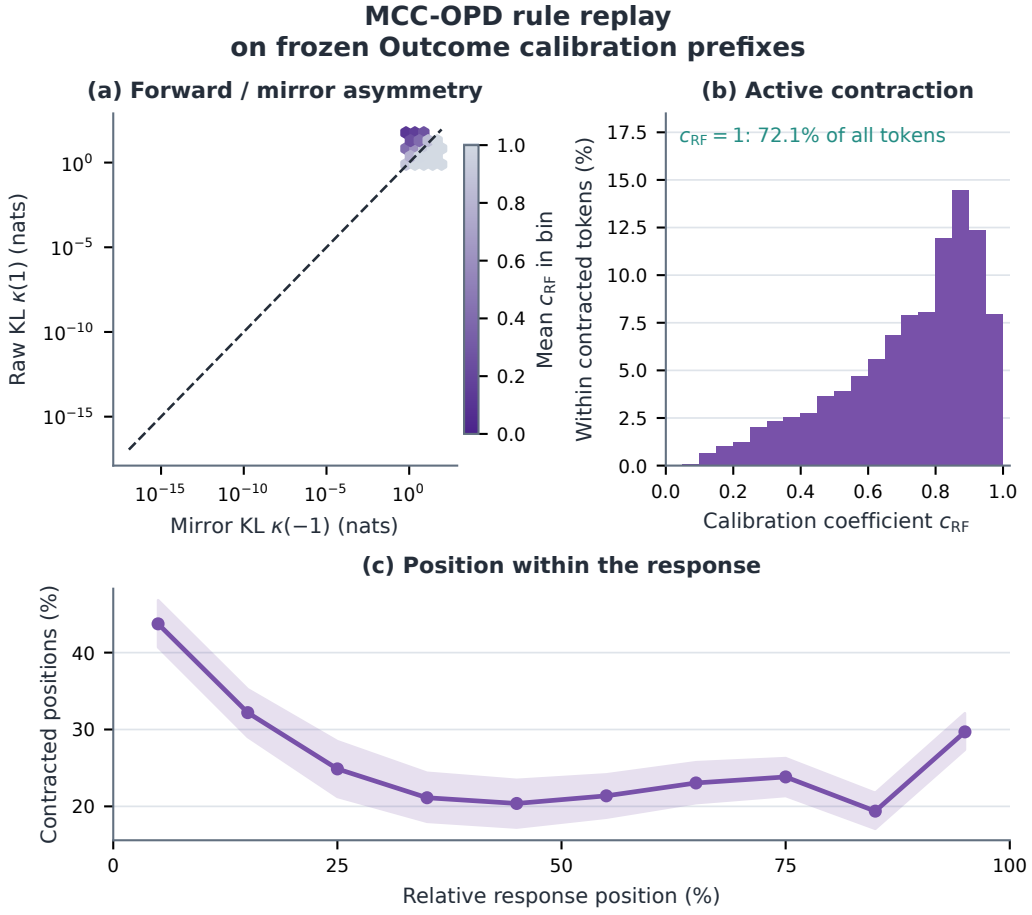

Figure 11: MCC-OPD target-rule replay with $\alpha = 1.5$ on all 1,120 saved trajectories from the Outcome calibration run (140 training questions). These are fixed inputs to the replay, not MCC-OPD online traces. Full-context and evidence-ablated teacher log-probabilities use the saved shared token partition after removing padded cells. (a) Each hexagon averages c_{RF} for its tokens; the diagonal is the raw feasibility boundary. Twenty zero-KL endpoints are omitted from log axes. (b) The histogram is conditioned on $c_{\mathrm{RF}} < 1$; the $c_{\mathrm{RF}} = 1$ point mass among all tokens is reported separately. (c) Means first average positions within response bins, then responses within questions. Empty bins in short responses contribute no observation. Bands are 95% question-bootstrap intervals. Positions include the complete saved response, including protocol tokens; trends are descriptive, not prefix interventions.

F.2 Paired answer changes and examples

Figure 12 decomposes the matched MCC-OPD–Uncalibrated difference into the two directions of correctness change: incorrect to correct and correct to incorrect. Both models use $\alpha = 1.5$, $N = 4$, and the final checkpoint. The paired intervals use the same benchmark-specific resampling units as the main component comparison.

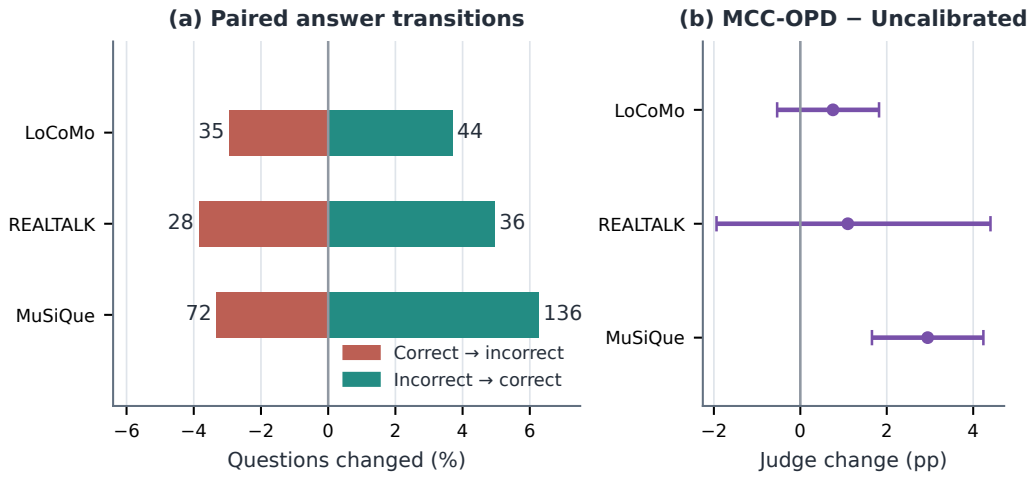

Figure 12: Paired answer changes for matched Uncalibrated and MCC-OPD ($\alpha = 1.5$, $N = 4$). Bars show percentages within each benchmark; labels are question counts. Net changes include paired 95% intervals using conversation clusters for LoCoMo/REALTALK and questions for MuSiQue.

Figures 13–15 give selected examples from each benchmark, including both directions of correctness change. Cases were selected from the fixed Judge groups for clear evidence and short answers; their prevalence is represented by Figure 12, not by the number of examples. The REALTALK panel also includes semantically correct answers with different lexical F1 scores. Each pair uses the same supplied input; the independent generated answers are not a fixed-prefix intervention.

F.3 Common-prefix attention cases

We select four distinct questions by a fixed ID order, using OPD (partition KL) as the baseline. The cases cover a baseline error corrected by outcome calibration, a baseline error corrected by MCC-OPD, a baseline success lost by a calibrated variant, and success for all three. All happen to come from one conversation. We replay the full input and the OPD (partition KL) answer prefix in the three students and the 8B teacher, holding token IDs fixed. Figure 16 shows that evidence-attention increases can coexist with reproducing the same answer tokens, including a wrong answer. All three students agree with every greedy next-token prediction along this fixed trajectory. In the first case, the teacher differs only on terminal punctuation while retaining the incorrect answer content. In two other cases it disagrees at the first answer token; replay alone does not determine its free continuation. These selected cases illustrate prefix conditioning, not a population-level gain in evidence use or a causal role of attention.

G Local directional utility estimates at fixed prefixes

We further examine the local directional utility estimator using 140 fixed reasoning prefixes, one per training question, and 16 fresh continuations per prefix. Prefixes are reasoning midpoints from the third epoch of an independent outcome-calibration run; the scoring and continuation policy is its frozen final checkpoint. A fixed mean reward from historical answers supplies the question baseline, independently of the new continuations. This isolates conditional variation and differs from the online fresh leave-one-out estimator in Eq. (10).

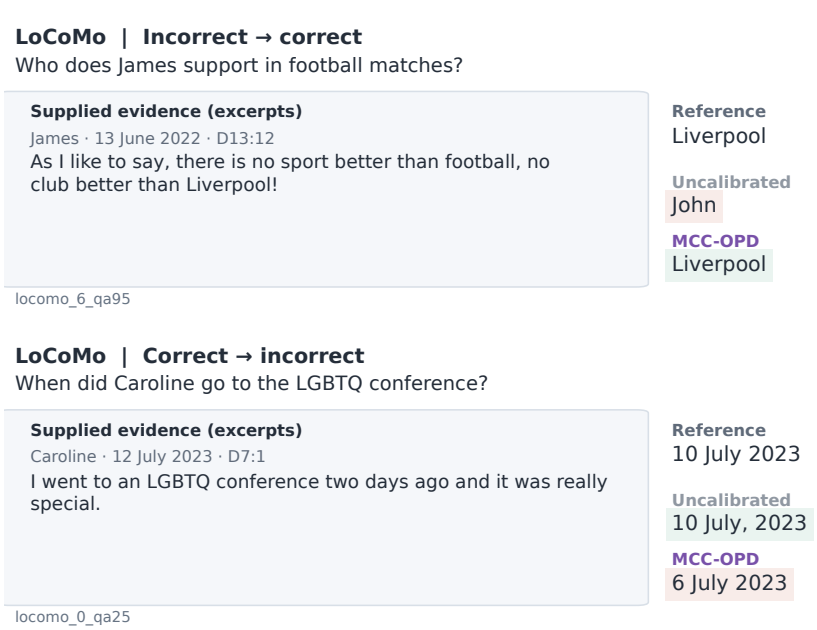

Figure 13: LoCoMo examples from matched Uncalibrated and MCC-OPD test outputs. The question, reference and answers retain their recorded wording. Evidence excerpts come from the supplied prompt or supporting paragraphs. Green/red answer backgrounds indicate semantic correctness under the frozen Judge. Both directions of answer change are shown.

Historical control variate. As a supplementary estimator study, we construct a token-return predictor from 8,400 previous answers to the same training questions. It pools final rewards by question and token occurrence, without matching the intermediate prefix; repeated occurrences and zero rewards are retained. Its predicted advantage $m(v)$ is supported on the explicit partition tokens and is zero in the tail. Following control-variate principles (Gu et al., 2017; Tucker et al., 2018), define

$$\widehat{U}_0 = h(Y)(R-b), \qquad H = h(Y)m(Y), \qquad \widehat{U}_\eta = \widehat{U}_0 - \eta\big(H - \mathbb{E}_p H\big). \tag{16}$$

For fixed m, h, and matched sampling $Y \sim p$, the added term has zero expectation, while

$$\mathrm{Var}(\widehat{U}_\eta) = \mathrm{Var}(\widehat{U}_0) + \eta^2\,\mathrm{Var}(H) - 2\eta\,\mathrm{Cov}(\widehat{U}_0, H). \tag{17}$$

Thus historical prediction can raise or lower variance. A shuffled control preserves the within-question reward multiset and token set while disrupting their correspondence. All estimators reuse the same fresh continuations. This extension is not required by either reward-free or outcome calibration.

Observed variation. At $\eta = 0.25$ and 0.5, mean conditional variance is respectively 3.80% and 4.39% below the uncorrected estimator; at $\eta = 1$ it is 4.08% higher, and the shuffled $\eta = 0.5$ control is 1.30% higher. Only the $\eta = 0.25$ paired variance-difference interval excludes zero. At $\eta = 0.5$, reward-prediction MSE is 0.03662, versus 0.03778 for the baseline and 0.04775 after shuffling. The real-minus-shuffled MSE difference has a 95% question bootstrap interval of $[-0.02067, -0.00200]$, while the corresponding variance difference's interval includes zero. All paired mean-difference intervals include zero. These measurements provide limited evidence about this fixed-prefix sample, not an online training benefit.

Sampling and coverage limits. No continuation reaches its length cap. Historical token coverage is 97.94% by actor probability mass and 97.81% by observed first actions, but 100 of 140 prefixes yield only one distinct first action among the 16 samples. Coverage therefore does not establish accurate conditional values. Although the sampling settings are temperature one without truncation, Transformers actor scoring with SDPA and vLLM rollout generation in bfloat16



Figure 14: REALTALK examples from matched Uncalibrated and MCC-OPD test outputs. The question, reference and answers retain their recorded wording. Evidence excerpts come from the supplied prompt or supporting paragraphs. Green/red answer backgrounds indicate semantic correctness under the frozen Judge. Both directions of answer change are shown. The final pair has identical semantic correctness but different lexical F1.

MuSiQue | Incorrect → correct

When did the owner of Bucephalus die?

Supplied evidence (excerpts)

Alexandria Bucephalous · paragraph 7

was a city founded by Alexander the Great in memory of his beloved horse Bucephalus.

Alexandria Bucephalous · paragraph 7

Bucephalus had died after the Battle of the Hydaspes in 326 BC.

Hellenistic period · paragraph 13

between the death of Alexander the Great in 323 BC and the emergence of the Roman Empire

Reference

323 BC

Uncalibrated

326 BC

MCC-OPD

323 BC

2hop__826864_17335

MuSiQue | Correct → incorrect

What year did the Leone brothers start their band?

Supplied evidence (excerpts)

Madina Lake · paragraph 3

Madina Lake is an American alternative rock band formed in Chicago in 2005.

Matthew Leone · paragraph 19

Matthew Jon Leone ( ) is the bassist of rock band Madina Lake. He was born on May 31, 1975.

Matthew Leone · paragraph 19

Matthew is three minutes older than his identical twin Nathan Leone, who is the lead vocalist in Madina Lake.

Reference

2005

Uncalibrated

2005

MCC-OPD

1975

2hop__756620_127399

Figure 15: MuSiQue examples from matched Uncalibrated and MCC-OPD test outputs. The question, reference and answers retain their recorded wording. Evidence excerpts come from the supplied prompt or supporting paragraphs. Green/red answer backgrounds indicate semantic correctness under the frozen Judge. Both directions of answer change are shown.

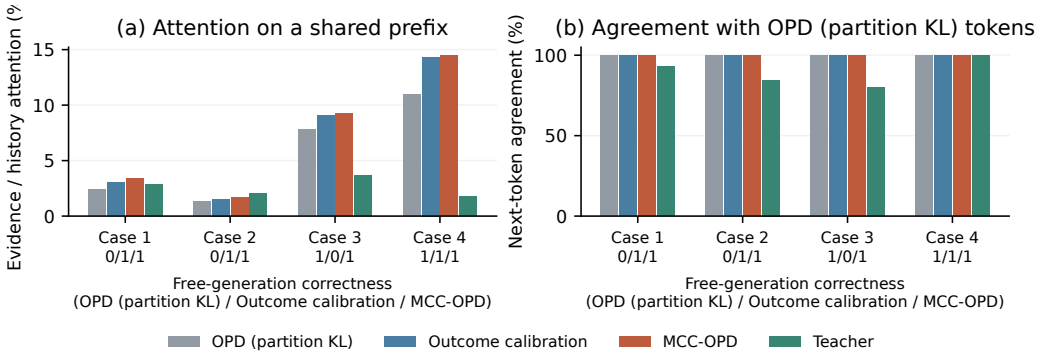

Figure 16: **Attention and continuation at a common prefix.** All models replay the same reasoning and answer tokens generated by OPD (partition KL). Left: answer-to-evidence attention as a fraction of answer-to-history attention. Right: greedy agreement with the next fixed OPD (partition KL) answer token. Labels under each case give the correctness of each student's separate, original free-generation answer. Those verdicts are not replay outcomes.

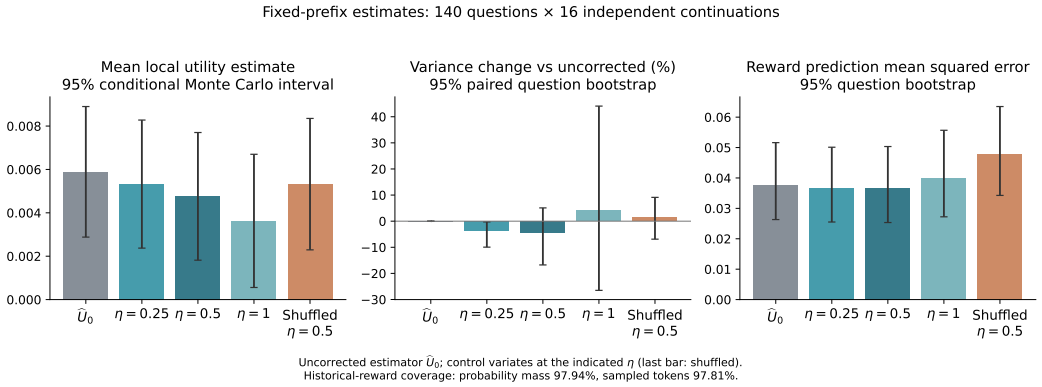

Figure 17: **Fixed-prefix local-utility estimator diagnostics.** The uncorrected estimator $\widehat{U}_0$ uses the independent historical question baseline described here. Other bars apply a control variate at the indicated η, with shuffled rewards in the last bar. Left: estimated means with normal-approximation conditional Monte Carlo intervals. Center: variance differences relative to $\widehat{U}_0$, divided by its mean observed conditional variance. Right: reward-prediction mean squared error (MSE) for $b + \eta m(Y)$. The latter two panels use question bootstrap intervals. This evaluates conditional scalar estimates, not parameter-gradient variance.

(BF16) differ numerically: absolute first-token log-probability differences have mean 0.0302, 99th percentile 0.6134, and maximum 1.3738. These measured differences prevent treating the analytical matched-distribution assumption as empirically verified by this diagnostic.

H LIMITATIONS

The mirror budget bounds divergence between teacher targets on a shared token partition. It does not bound the student's parameter update or guarantee correctness, and its reference scale is a design choice rather than an optimal task-dependent budget. MCC-OPD inherits the chosen evidence intervention and the student's prefix: missing contrasts or shared prefix errors cannot be repaired by scaling alone.

The observed gains depend on the task and evaluation metric. Our reported comparisons focus on LoCoMo, REALTALK, MuSiQue, and LongMemEval; they do not establish a uniform advantage across transfer tasks or answer styles.

Our experiments study answering from supplied evidence. The counterfactual teacher view relies on annotated supporting evidence, and we do not jointly train retrieval or memory editing. Automatically identifying which evidence to remove and integrating calibration with retrieval and memory updates are directions for future work. Evaluation in open-ended, multi-turn interactions would further assess how the trained answer policy handles evolving memories.