ON THE EDGE OF UNDERSTANDING: CHURN IN LLM POST-TRAINING

Anonymous authors
Paper under double-blind review

ABSTRACT

Beneath steady accuracy improvements in LLM post-training, prior work finds that many queries flip between solved and unsolved across training epochs. However, the underlying mechanisms driving these flips remain poorly understood. Through a large-scale analysis, we find that a single gradient update flips 21% of learnable queries between solved and unsolved throughout training, across both in-distribution and out-of-distribution datasets — a phenomenon we term LLM churn. Accuracy gain is the surplus of positive over negative churn, and both persist even after accuracy plateaus. Churning queries reside on the LLM's edge of understanding, where a razor-thin confidence margin between correct and incorrect answers makes them $12.5\times$ more vulnerable to perturbation. Persistent churn incurs substantial costs: solved queries that flip to unsolved leave a 25.8 pp forgetting gap, and 28.5% of final accuracy comes from unreliable answers on queries that flip repeatedly across training checkpoints. Finally, we find that churn is not benign: applying sharpness-aware minimization after accuracy plateaus reduces churn while lifting final accuracy by 5 pp for models up to 4B parameters, with no benefits on larger models. Our work highlights the importance of analyzing the evolution of query-level performance during post-training and points to a promising path for improving learning efficiency via reduced churn.

1 INTRODUCTION

Progress in LLM post-training is conventionally measured by accuracy on held-out evaluation sets. However, this aggregate metric only reflects how many examples the model answers correctly, concealing which individual examples change their outcomes: a single gradient update can turn some incorrect answers into correct ones while simultaneously flipping others from correct to incorrect. Although correctness flips have been observed in post-training [16, 22], existing work has not systematically characterized these bidirectional flips at the single-update level. Furthermore, there is a lack of understanding as to how these flips relate to final accuracy and prediction stability.

To this end, across 8 models, 10 in-distribution and out-of-distribution datasets, and 4 post-training methods, we track sample-level correctness trajectories throughout training and quantify the positive

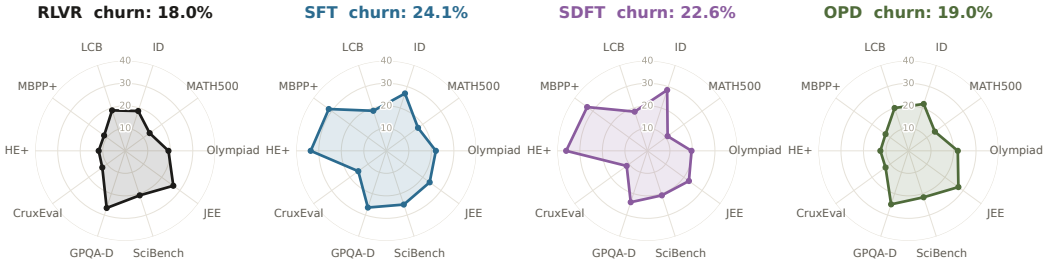

Figure 1: Churn rate by post-training method: reinforcement learning with verifiable rewards (RLVR, 7 models), supervised fine-tuning (SFT, 8 models), self-distillation fine-tuning (SDFT, 8 models), and on-policy distillation (OPD, 3 models), averaged over models spanning three model families up to 9B.

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.

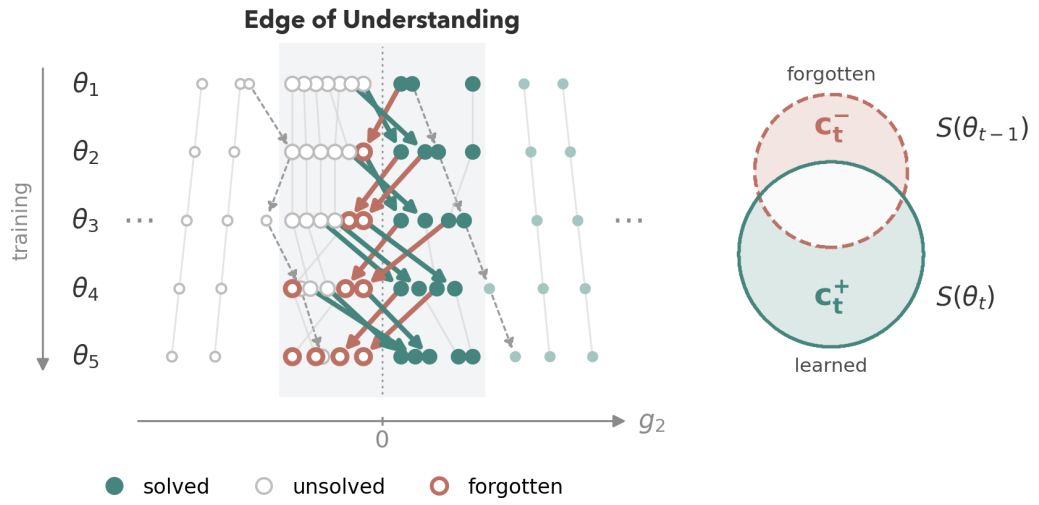

Figure 2: **Churn on the edge leaves a forgetting gap** (schematic). Updates push queries with a margin near g_2=0 across the line (teal: positive churn c_t^+; red: negative churn c_t^-), and the forgotten queries pile up.

and negative flips induced by individual gradient updates. We find that among evaluation queries answered correctly at least once during training, a single update flips the correctness of 20.9% of samples on average (Figure 1), a phenomenon we term **LLM churn**. The term draws an analogy to policy churn in deep reinforcement learning, which concerns changes in greedy actions [36]. From this perspective, accuracy gains reflect only the net difference between positive and negative flips. Even after aggregate accuracy reaches a plateau, churn in both directions persists.

Churning queries reside on the model's *edge of understanding*, where the confidence margin between correct and incorrect answers is razor-thin (Figure 6). In other words, correctly answered churning queries almost "happen to be correct" by near-chance: minor parameter perturbations flip them to incorrect 12.5× more often than correctly answered non-churning queries (Figure 7), highlighting their inherent unreliability.

LLM churn leaves lasting traces to the final model checkpoint. Churning queries that were answered correctly at some point but end up incorrect form a forgetting gap [22], which averages to 25.8 percentage points in our experiments (Figure 3). Churning queries that finish correct but flips repeatedly during training are indistinguishable in final accuracy from queries that stayed correct throughout (Table 22). Predictions counted equally toward final accuracy can therefore differ substantially in stability.

Since churning queries flip under small perturbations, we test whether seeking flatter regions of the loss landscape can stabilize them. In SFT experiments on closed-book QA, applying sharpness-aware minimization (SAM) [10] after accuracy plateaus substantially reduces churn and improves final accuracy for models up to 4B parameters, but we do not observe an accuracy gain for the evaluated 9B model. This suggests that plateau churn can be curbed under certain conditions, though achieving consistent gains across larger models and broader tasks remains an open challenge.

Our contributions are:

- **Phenomenon: LLM churn.** We identify LLM churn as the governing dynamic behind post-training, showing that accuracy gains reflect only the net difference between positive and negative flips. Across 8 models, 10 datasets, and 4 post-training methods, we quantify that a single gradient update flips response correctness on an average of 20.9% of learnable queries, a turnover that persists even after accuracy reaches a plateau (Section 3.2, Appendix D).
- **Cause: edge of understanding.** We attribute churn to near-zero decision margins. Churning queries sit on the model's edge of understanding, where correct and incorrect candidates receive

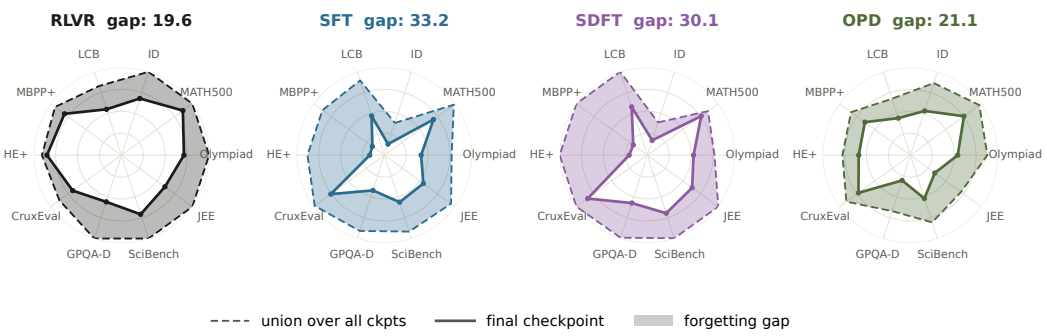

Figure 3: **Forgetting gap by method**, averaged over models. Dashed: fraction of learnable queries ("union" in the legend); solid: accuracy of the final checkpoint; shaded ring: forgetting gap (Equation 7).

nearly equal confidence. Under slight parameter perturbations, these queries flip 12.5 times more frequently than consistently solved ones (Section 3.3).

- **Consequence: forgetting gap and prediction stability.** We reveal two costs of persistent churn: unrecovered negative flips leave behind an average forgetting gap of 25.8 percentage points, while recovered flips are counted toward standard accuracy despite remaining unreliable and unstable. Tracking churn thus separates answers that are stably learned from those that happen to be correct when training stops, which accuracy alone cannot distinguish (Section 3.4).
- **Preliminary intervention.** We show that plateau churn can be mitigated. While standard hyperparameter tuning and replay have little effect, applying sharpness-aware minimization (SAM) after the plateau reduces churn by 31% and raises final accuracy by 5 percentage points for models up to 4B parameters. However, these gains do not transfer to the evaluated 9B model (Section 3.5, Appendix E).

2 Related Work

Negative flips. A negative flip occurs when an individual example transitions from correct to incorrect during training, a query-level change that accuracy hides. Negative flips have been studied in conventional machine learning models [1, 6, 37, 41] and, more recently, in LLM post-training. Harmon et al. [16] count queries that the base model solves but the final checkpoint does not, for two post-training methods. Li et al. [22] extend this to count queries solved at any epoch but not by the final checkpoint, and also observe that a single query can oscillate between solved and unsolved across epochs for two post-training methods; a similar observation is made by Mousavi et al. [31]. However, these studies observe churn only at the epoch level and do not explain why it occurs. We conduct a large-scale analysis to quantify churn induced by a single gradient step across four post-training methods. We identify churn as the dynamics that governs post-training, trace its cause to the model's edge of understanding, and use it to explain the formation of query-level forgetting seen at coarser granularity in Harmon et al. [16], Li et al. [22], Mousavi et al. [31].

Policy churn. A closely related phenomenon in deep reinforcement learning is policy churn [36, 40], where a single gradient step changes the greedy action in roughly 10% of all states. We instead study correctness flips in LLM post-training, which directly affect what the model can solve. Across RLVR, SFT, OPD, and SDFT, a single gradient update flips the correctness of 20.9% of learnable queries. LLM churn is not benign: on average, 25.8% of queries end training as negative flips, solved at some point but unsolved by the final checkpoint. It also allows lucky guesses on churning queries to be counted toward final accuracy.

Catastrophic forgetting. Catastrophic forgetting is typically measured as a drop in dataset-level accuracy on earlier tasks under sequential training [7, 9, 12, 21, 26–29, 35, 48]. We instead measure query-level forgetting, on both in-distribution and out-of-distribution data, induced by individual gradient updates in single-task post-training. Furthermore, standard continual learning methods do not reduce query-level forgetting.

3 EXPERIMENT

3.1 SETUP

Post-training paradigms and data. We study four post-training methods across two regimes: on-policy RLVR (via Dr.GRPO [15, 25, 38]) and OPD [2] trained on an equal mixture of OpenCodeReasoning [3], OpenMathReasoning [30], and OpenScienceReasoning [32], alongside off-policy SFT [43] and SDFT [39] trained on OpenThoughts-114k [14]. Because the two regimes use different training data, we analyze churn within each method and do not compare across methods.

Evaluation benchmarks. From each training mixture, we hold out an in-distribution (ID) evaluation set. We also assess out-of-distribution (OOD) generalization across nine reasoning benchmarks spanning mathematics (MATH500 [19, 23], OlympiadBench [17]), multidisciplinary science (JEEBench [4], SciBench [42], GPQA-Diamond [34]), and code (CruxEval [13], HumanEval+ and MBPP+ [5, 8, 24], LiveCodeBench [20]).

Models and protocol. We evaluate eight open models up to 9B parameters: Qwen3-1.7B/4B/8B [46], Qwen3.5-2B/4B/9B [33], and Gemma4-E2B/E4B [11]. SFT and SDFT cover all eight, RLVR covers seven (excluding Qwen3.5-9B), and OPD uses three teacher-student pairs: Qwen3-8B to 1.7B, Qwen3.5-9B to 2B, and Gemma4-E4B to E2B. We keep 20 uniformly spaced checkpoints per run and train each configuration with three random seeds. Unless stated otherwise, reported numbers are means over the three seeds, with $\pm$ denoting the standard deviation across seeds. Greedy decoding uses batch-invariant kernels [18] to ensure deterministic evaluation. Appendix A provides full training details.

3.2 PHENOMENON: CHURN

Let $r(a \mid q) \in \{0, 1\}$ indicate whether response a correctly answers query q, and let $\pi_\theta^{\mathrm{g}}(q)$ denote the greedy response of an LLM π_θ. For a validation set $\mathcal{Q}$, the queries solved by θ form the set

$$S(\theta) \;=\; \big\{\, q \in \mathcal{Q} \;:\; r\big(\pi_\theta^{\mathrm{g}}(q) \mid q\big) = 1 \,\big\}. \tag{1}$$

Across checkpoints $\theta_1, \ldots, \theta_T$ of a post-training run, the *learnable queries* are those solved at least once:

$$L \;=\; \bigcup_{t=1}^{T} S(\theta_t). \tag{2}$$

Churn captures query-level learning and forgetting between consecutive checkpoints θ_{t-1} and θ_t. The *positive* and *negative churn rates* are the fractions of learnable queries that become solved or become unsolved, respectively; the *churn rate* is their sum:

$$c_t^{+} = \frac{\big|S(\theta_t) \setminus S(\theta_{t-1})\big|}{|L|}, \qquad c_t^{-} = \frac{\big|S(\theta_{t-1}) \setminus S(\theta_t)\big|}{|L|}, \qquad c_t = c_t^{+} + c_t^{-}. \tag{3}$$

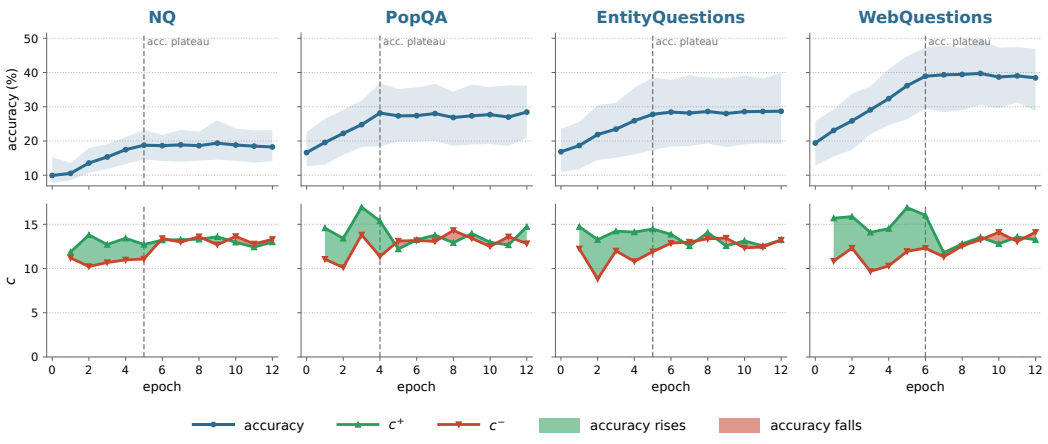

Figure 4: Accuracy (top) and c^{+}, c^{-} (bottom) over SFT epochs on closed-book QA.

Accuracy change on step t is proportional to the difference between positive and negative churn rate:

$$\underbrace{\frac{S(\theta_t)}{|\mathcal{Q}|}}_{\text{Accuracy}_t} - \underbrace{\frac{S(\theta_{t-1})}{|\mathcal{Q}|}}_{\text{Accuracy}_{t-1}} = \frac{|L|}{|\mathcal{Q}|} \cdot (c_t^+ - c_t^-)$$

We report rates averaged over $t = 2, \ldots, T$.[1] Figure 2 illustrates both directions of churn.

Method		ID	MATH500	Olympiad	JEE	SciBench	GPQA-D	CruxEval	HE+	MBPP+	LCB
RLVR	c^+	9.4 ±1.6	6.8 ±5.5	9.7 ±5.0	13.2 ±4.6	10.4 ±3.9	13.4 ±4.3	6.4 ±4.1	5.8 ±3.7	5.8 ±1.8	9.5 ±2.7
	c^-	9.2 ±2.3	6.6 ±5.6	9.5 ±5.2	13.1 ±5.3	10.4 ±4.2	13.3 ±5.0	6.2 ±4.2	6.0 ±4.0	5.8 ±2.3	9.5 ±3.6
	$\boldsymbol{c}$	**18.6** ±3.6	**13.4** ±10.9	**19.2** ±9.9	**26.3** ±9.2	**20.8** ±7.8	**26.7** ±7.7	**12.6** ±8.1	**11.8** ±6.9	**11.6** ±3.3	**19.0** ±5.4
SFT	c^+	13.0 ±8.1	9.0 ±5.7	11.0 ±5.6	11.7 ±5.1	12.5 ±5.1	13.0 ±3.6	7.5 ±4.9	16.1 ±8.3	15.1 ±6.7	8.6 ±5.0
	c^-	13.8 ±11.6	8.4 ±4.3	11.0 ±4.9	12.1 ±6.0	12.5 ±5.5	13.5 ±3.9	7.9 ±5.1	17.6 ±10.4	16.5 ±8.8	10.1 ±5.6
	$\boldsymbol{c}$	**26.8** ±16.8	**17.4** ±8.6	**22.0** ±8.3	**23.8** ±8.7	**25.0** ±8.0	**26.5** ±6.0	**15.4** ±9.0	**33.6** ±11.6	**31.6** ±10.2	**18.7** ±8.3
SDFT	c^+	14.3 ±6.4	5.6 ±4.0	9.8 ±5.2	11.3 ±5.8	10.6 ±4.3	12.0 ±3.3	5.7 ±3.7	17.6 ±8.3	15.7 ±6.6	8.7 ±4.6
	c^-	14.1 ±5.7	5.5 ±4.0	9.8 ±6.2	11.4 ±5.9	10.2 ±4.6	12.0 ±3.3	5.7 ±3.7	18.5 ±8.4	17.5 ±9.2	9.6 ±5.6
	$\boldsymbol{c}$	**28.4** ±8.4	**11.1** ±5.9	**19.6** ±7.9	**22.8** ±9.2	**20.8** ±6.7	**24.0** ±5.1	**11.3** ±6.5	**36.1** ±8.2	**33.1** ±8.7	**18.3** ±6.6
OPD	c^+	10.9 ±1.2	7.2 ±1.9	10.8 ±2.3	13.6 ±2.2	10.8 ±1.6	12.4 ±3.8	6.3 ±1.4	6.2 ±3.7	6.3 ±2.7	10.0 ±1.5
	c^-	11.0 ±1.3	7.2 ±1.9	11.1 ±2.5	13.8 ±2.3	10.9 ±2.0	12.6 ±3.8	6.2 ±1.4	6.3 ±3.8	6.3 ±2.4	10.0 ±1.5
	$\boldsymbol{c}$	**21.9** ±1.7	**14.4** ±3.1	**21.9** ±3.5	**27.4** ±2.7	**21.7** ±2.5	**25.0** ±4.8	**12.5** ±2.4	**12.5** ±6.7	**12.6** ±4.5	**20.0** ±1.7

Table 1: **Churn rate by method** (% of learnable queries, Equation 3): positive c^+, negative c^-, and total c, averaged over consecutive checkpoints; $\pm$ is the standard deviation over all checkpoint intervals of all models.

Table 1 shows consistent *LLM churn* across seven RLVR runs, eight SFT runs, eight SDFT runs, and three OPD runs, with churn rates averaged over the ten datasets ranging from 18.0% to 24.1% across methods (Figure 1). In other words, a single gradient update changes whether the model answers roughly one in five learnable queries correctly. Churn occurs on both ID and OOD datasets, with average rates of 23.9% and 20.6%, respectively, showing that it extends beyond the training distribution (per model and dataset in Tables 5–8).

We do not observe substantial accuracy gains on these benchmarks, which could be attributed to the limited range of batch size and training steps we could consider within our computational constraints. To study churn under substantial accuracy improvement and prolonged training, we run SFT on four closed-book QA datasets with four models (Qwen3.5-0.8B/4B/9B and Gemma4-E2B; configurations in Appendix D). We observe that the net churn $c_t^+ - c_t^-$ tracks the accuracy change, and that $c_t^+ \approx c_t^-$ once accuracy plateaus, with both remaining at 13.1% on average (Figure 4; per model in Figure 18, Appendix D).

3.3 Cause: the edge of understanding

What separates queries that churn from those that stay stably correct or wrong? To answer this, we must first decide which queries churn: a query may churn early in training and settle later. For θ_t, we partition $\mathcal{Q}$ into three groups:

$$\begin{aligned} &\textit{unforgettable} && \mathcal{U}_t = \textstyle\bigcap_{i=t}^{T} S(\theta_i) \\ &\textit{unsolvable} && \mathcal{N}_t = \mathcal{Q} \setminus \textstyle\bigcup_{i=t}^{T} S(\theta_i) \\ &\textit{churning} && \mathcal{C}_t = \mathcal{Q} \setminus (\mathcal{U}_t \cup \mathcal{N}_t) \end{aligned} \tag{4}$$

To understand what distinguishes these three groups, we measure the model's confidence margin between correct and wrong responses. Let $\ell_\theta(a \mid q)$ denote the average per-token log-probability of

[1] In practice, we take one gradient step from each of the 20 intermediate checkpoints and measure the churn rate.



Figure 5: Averaged churn rate dynamics for each method; bands are min–max across models.

response a to query q. We define the confidence margin g_2 as

$$\ell_\theta(a \mid q) = \frac{1}{|a|} \log \pi_\theta(a \mid q), \tag{5}$$

$$g_2(q;\theta) = \max_{a\,:\,r(a|q)=1} \ell_\theta(a \mid q) - \max_{a\,:\,r(a|q)=0} \ell_\theta(a \mid q), \tag{6}$$

where $r(a \mid q) \in \{0, 1\}$ indicates response correctness. A positive margin means the model favors a correct response; a negative margin means it favors an incorrect response; a margin near zero means the two receive similar confidence.

For each query, we draw eight samples at temperature 1 and select the sample of opposite correctness to the greedy response with the highest per-token log-probability. The confidence margin is estimated by the difference in their per-token log-probabilities, computed at each checkpoint and averaged across checkpoints (per-group values for every model and dataset in Appendix C).

Across all steps, we compare $g_2(q;\theta_t)$ among $\mathcal{N}_t$, $\mathcal{U}_t$, and $\mathcal{C}_t$. The confidence margin separates the three groups: unforgettable queries have large positive margins (confidently right), unsolvable queries have large negative margins (confidently wrong), and churning queries have near-zero margins, where the model barely prefers either answer (Table 2, Figure 6, bottom). The absolute margin of churning queries is $5.6\times$ smaller than that of unforgettable queries and $11.2\times$ smaller than that of unsolvable queries (ratio of means over the four methods). Churning queries are also less confident in whichever answer they give: their correct answers are less likely than those of unforgettable queries, and their incorrect answers less likely than those of unsolvable queries (Figure 6, top). This suggests churning queries lie near the **edge of understanding**, where the model is weakly confident in both correct and wrong answers and barely prefers either, so a small update can flip its preference.

Group		RLVR	SFT	SDFT	OPD
Unforgettable	$\mathcal{U}_t$	$+0.048$	$+0.054$	$+0.059$	$+0.067$
Churning	$\boldsymbol{\mathcal{C}_t}$	$\mathbf{+0.015}$	$\mathbf{+0.014}$	$\mathbf{+0.011}$	$\mathbf{+0.001}$
Unsolvable	$\mathcal{N}_t$	-0.079	-0.121	-0.143	-0.117
# of Runs	n	77	88	90	22

Table 2: Mean g_2 by group and method.

Churning queries are also fragile under random weight perturbation. We add isotropic Gaussian noise with standard deviation $\rho \in \{0.001, 0.003, 0.01, 0.03\}$ to every weight of the final checkpoint and measure each group's *flip rate*: the fraction of its queries whose greedy correctness changes (Figure 7, bottom). At ρ=0.001, the flip rate of $\mathcal{C}_t$ is 8–18$\times$ that of $\mathcal{U}_t$, $12.5\times$ on average over runs (Figure 7, top; Table 17). The ratio falls as ρ grows (Table 17), suggesting churning queries are more sensitive than unforgettable queries under small rather than arbitrarily large perturbations, similar to controlled parameter updates during continual learning.

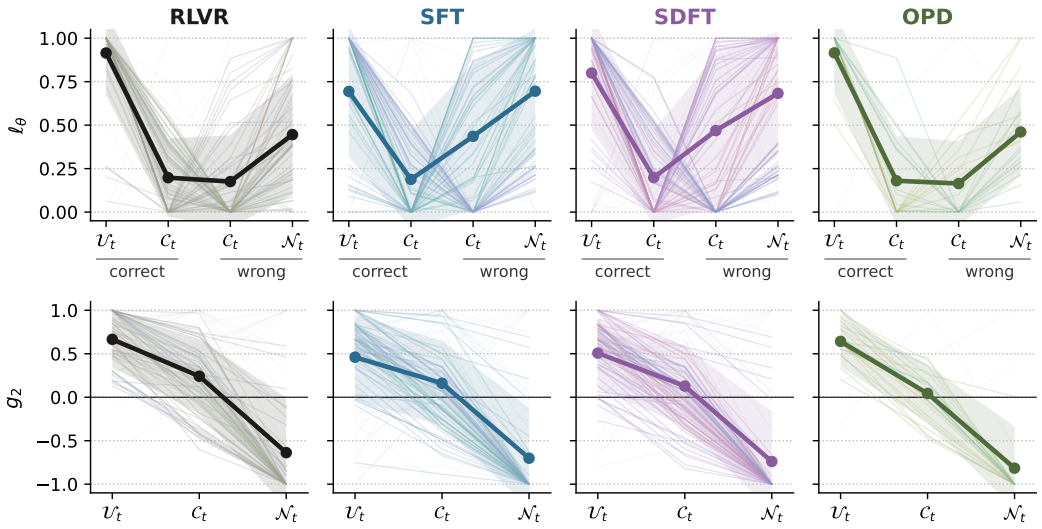

Figure 6: **Edge of understanding.** Top: ℓ_θ normalized to $[0, 1]$. Bottom: g_2 normalized to $[-1, 1]$. Each thin line is a (model, dataset) pair.

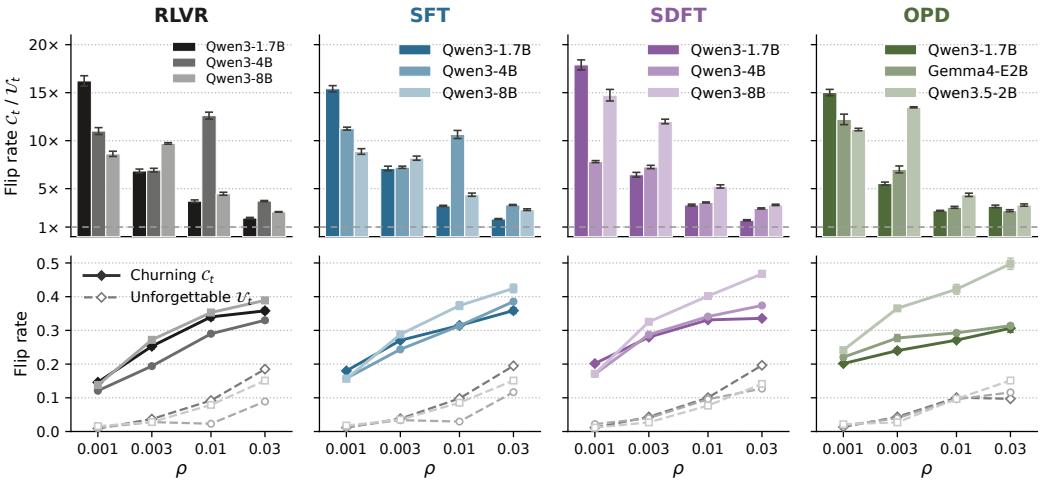

Figure 7: **Sensitivity to perturbation.** Top: ratio of flip rates at ρ=0.001. Bottom: flip rate vs. ρ.

Together, the confidence margin and the perturbation results give a mechanistic picture of churn: churning queries sit near g_2=0, where any small parameter change, whether a gradient step or random noise, can move them to the other side of the line.

3.4 Consequence: The Forgetting Gap

If queries on the edge keep crossing the line with every update, and churn does not decay after accuracy plateaus (Section 3.2), then which side each of them is on when training stops is largely dependent on the time at which post training is set to conclude. The final checkpoint θ_T therefore solves only part of the learnable set L, even though each query in L was solved at some point. Following Li et al. [22], we quantify this loss with the *forgetting gap*, the fraction of all queries that are learnable but unsolved by θ_T:

$$\text{Forgetting Gap} = \frac{|L \setminus S(\theta_T)|}{|\mathcal{Q}|} = \underbrace{\frac{|L|}{|\mathcal{Q}|}}_{\text{learnable}} - \underbrace{\frac{|S(\theta_T)|}{|\mathcal{Q}|}}_{\text{final accuracy}}. \tag{7}$$

Method		ID	MATH500	Olympiad	JEE	SciBench	GPQA-D	CruxEval	HE+	MBPP+	LCB
	Final	42.9 ±10.8	79.2 ±8.4	47.5 ±10.7	51.6 ±17.8	54.4 ±12.0	47.3 ±10.3	65.0 ±23.2	72.1 ±11.5	61.0 ±11.8	39.9 ±16.0
	Learnable	63.1 ±11.1	91.4 ±3.1	66.6 ±6.3	84.1 ±7.6	76.7 ±8.3	84.5 ±4.7	82.8 ±18.6	77.0 ±11.3	71.1 ±8.7	60.1 ±22.5
RLVR	**Gap**	**20.1** ±2.2	**12.2** ±5.5	**19.2** ±4.7	**32.4** ±10.4	**22.3** ±6.0	**37.2** ±6.3	**17.8** ±10.0	**4.9** ±2.1	**10.1** ±7.9	**20.2** ±8.3
	Best	44.4 ±11.5	80.9 ±8.2	49.6 ±10.8	54.1 ±16.8	56.3 ±12.1	51.4 ±9.9	66.2 ±23.0	73.4 ±10.8	63.2 ±10.3	42.4 ±17.2
	Majority	42.1 ±10.2	80.3 ±8.3	47.0 ±11.2	50.9 ±19.6	54.7 ±13.5	46.0 ±11.7	65.7 ±24.4	71.6 ±10.8	60.8 ±11.1	39.2 ±17.4
	Final	9.0 ±6.9	69.7 ±18.7	32.7 ±11.1	53.9 ±23.0	47.2 ±11.1	40.0 ±10.5	73.6 ±9.9	15.4 ±8.2	15.4 ±6.6	35.2 ±18.4
	Learnable	25.6 ±10.3	91.5 ±7.9	55.4 ±11.3	83.0 ±11.4	73.5 ±9.7	77.6 ±5.8	93.1 ±5.2	77.0 ±15.2	70.4 ±13.4	66.2 ±17.0
SFT	**Gap**	**16.7** ±5.6	**21.8** ±11.1	**22.7** ±7.8	**29.2** ±15.2	**26.3** ±7.9	**37.7** ±11.4	**19.5** ±7.2	**61.6** ±19.2	**55.0** ±9.1	**31.0** ±11.8
	Best	14.8 ±5.2	73.5 ±17.2	35.8 ±9.6	61.4 ±19.6	52.1 ±10.8	46.3 ±12.1	81.1 ±8.9	38.9 ±13.5	33.0 ±10.4	51.5 ±15.2
	Majority	8.2 ±6.3	65.8 ±17.7	29.8 ±9.7	47.5 ±21.8	43.6 ±10.3	35.9 ±12.1	68.8 ±9.3	14.3 ±7.6	14.5 ±6.2	32.8 ±17.1
	Final	11.2 ±4.7	69.4 ±36.2	35.1 ±21.1	53.1 ±23.0	53.4 ±21.5	48.5 ±14.8	79.6 ±11.9	17.6 ±5.3	15.1 ±5.7	42.0 ±16.6
	Learnable	24.8 ±7.8	78.2 ±35.4	51.0 ±24.4	84.1 ±13.1	76.4 ±16.0	83.6 ±4.6	94.6 ±4.3	84.9 ±11.9	76.0 ±9.8	72.5 ±15.8
SDFT	**Gap**	**13.5** ±5.8	**8.8** ±4.7	**15.9** ±4.6	**31.0** ±13.0	**23.0** ±7.8	**35.0** ±11.5	**15.0** ±8.0	**67.2** ±12.1	**60.9** ±7.3	**30.5** ±11.2
	Best	14.3 ±4.8	71.6 ±35.6	39.0 ±20.5	60.9 ±20.5	59.2 ±20.7	53.6 ±12.4	84.4 ±10.2	45.5 ±13.3	43.9 ±9.1	56.8 ±15.8
	Majority	10.4 ±4.3	67.1 ±35.1	33.5 ±20.4	49.2 ±20.5	51.4 ±21.9	43.6 ±12.3	75.3 ±13.4	16.5 ±5.0	14.3 ±5.3	39.0 ±16.1
	Final	33.5 ±10.9	69.1 ±6.7	36.0 ±4.5	29.0 ±7.9	39.8 ±17.4	25.6 ±14.0	69.2 ±9.8	50.0 ±10.1	48.7 ±10.6	32.2 ±15.6
	Learnable	54.6 ±15.4	88.9 ±2.4	58.5 ±12.0	60.1 ±15.2	61.8 ±22.4	56.7 ±31.2	85.3 ±6.9	65.8 ±8.0	63.5 ±8.6	49.1 ±20.2
OPD	**Gap**	**21.1** ±4.6	**19.9** ±7.3	**22.5** ±9.7	**31.1** ±7.3	**22.0** ±5.5	**31.1** ±17.2	**16.1** ±3.1	**15.9** ±5.6	**14.8** ±3.5	**16.8** ±5.3
	Best	34.6 ±11.1	72.0 ±5.1	39.7 ±4.8	33.4 ±9.8	42.6 ±17.0	30.1 ±16.6	70.1 ±10.1	52.4 ±10.1	51.0 ±9.6	33.7 ±16.6
	Majority	32.8 ±10.3	71.3 ±6.5	36.1 ±3.9	29.2 ±8.6	40.0 ±17.5	23.5 ±13.0	69.5 ±15.1	49.5 ±10.0	49.0 ±11.2	28.4 ±12.4

Table 3: **Forgetting gap by method.** Best: best single checkpoint. Majority: majority vote over all checkpoints.

Table 3 reports a substantial forgetting gap across all training paradigms: seven RLVR runs, eight SFT runs, eight SDFT runs, and three OPD runs. Averaged over the ten datasets, the forgetting gap is 19.6 percentage points under RLVR, 32.2 under SFT, 30.1 under SDFT, and 21.1 under OPD (Figure 3). The forgetting gap for each model and dataset is reported in Tables 18–21 (Appendix G.1).

The forgetting gap, like the churn rate, exists on both in-distribution and out-of-distribution datasets, which makes it intuitively related to catastrophic forgetting, but at the query level, and it is observable in general training and does not require sequential training. The forgetting gap increases during early training and plateaus mid to late training (Figures 24–27, Appendix G.4). The gap does not shrink with model size (Figure 23, Appendix G.3).

Selecting the best checkpoint per model (early stopping) reduces the average forgetting gap across the ten datasets by only 1–5 percentage points. Majority voting over all intermediate checkpoints [22] fares no better: it changes accuracy by -3.1 to -0.3 percentage points, i.e., it slightly hurts (Table 3).

The forgetting gap measures loss relative to what the model learns *during* training. A complementary question is how much of what the model knew *before* training survives. Let θ_0 be the base model; we define *retention* as the fraction of queries solved by the base model that remain solved by the final checkpoint:

$$\text{Retention} = \frac{|S(\theta_0) \cap S(\theta_T)|}{|S(\theta_0)|}. \tag{8}$$

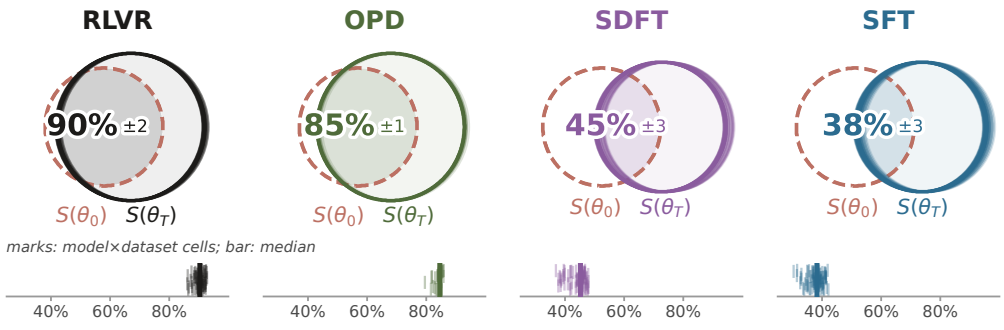

Figure 8: Retention by method.

Figure 8 reports retention by method: RLVR and OPD retain 85–90% of base-solved queries, SFT and SDFT 38–45%. Because RLVR/OPD and SFT/SDFT are trained on different data (Section 3), we do not attribute this difference to the training method.

Churn also leaves its mark on the queries that the final checkpoint does solve. Averaged over the four methods, 28.5% of final accuracy comes from queries that flipped from correct to incorrect at least three times across the 20 checkpoints of training before ending correct (18.4% under RLVR, 36.7% under SFT, 34.7% under SDFT, and 24.2% under OPD; Table 22, Appendix H). These answers count toward final accuracy like any other, but they sit on the edge of understanding (Section 3.3) and largely "happen to be correct" at θ_T, with a near-equal probability of being wrong.

3.5 Can churn be reduced?

We test whether churn can be reduced on the closed-book QA setting of Figure 4, changing one ingredient of SFT at a time, with three training seeds per configuration (Appendix E). Two observations motivate the candidates. Since the forgetting gap is negative churn accumulated into query-level forgetting, we test continual-learning methods designed against forgetting: weight decay, replay [35], SDFT [39], and PCGrad [47], together with a different learning rate (5×10^{-6} or 5×10^{-5}) and batch size (128 or 256). Since churning queries flip under small weight perturbations (Figure 7), we also test sharpness-aware minimization (SAM [10]), which seeks parameters whose loss is flat under such perturbations. None of the continual-learning methods or hyperparameter changes moves the churn rate by more than 1% or final accuracy by more than 0.6 points relative to plain SFT (Table 11). SAM does: on Qwen3.5-0.8B, Qwen3.5-4B, and Gemma4-E2B it lowers churn by 29–32% under either schedule, but the effect on accuracy depends on when it is enabled. From the first epoch, it costs 1–2 points of final accuracy on every model; enabled only after accuracy plateaus, it raises final accuracy by 5–6 points on all four datasets for these three models (Table 12). The contrast between the two schedules suggests that churn is benign, or even useful, while accuracy is rising, in line with the view of policy churn as a form of exploration [36], but harmful once accuracy has plateaued, when the model's understanding of a query should be stabilized rather than perturbed. On Qwen3.5-9B, however, SAM reduces churn by only 6–8% under either schedule and leaves accuracy unchanged after the plateau (-0.2 points), so SAM is not a consistent remedy across model sizes. What these results establish is that **churn after the plateau is not benign, and curbing it can improve accuracy**; a remedy that does so reliably at scale remains open.

4 Discussion and Future Work

Imagine learning one new fact and having 20% of your memories wiped out. This is essentially what happens during post-training. True continual learning requires steady accumulation, which in turn demands curbing negative churn. A trustworthy language model should form lasting competence rather than landing on correct answers by chance while failing unpredictably on others. Addressing churn offers a direct path to harvest the 25.8 percentage point forgetting gap and lift overall performance. Because methods like SAM show that stabilizing weights under perturbation can cut plateau churn on smaller models, future post-training methods should focus on boundary stability—ensuring that models retain what they know instead of silently trading old capabilities for new ones.

References

[1] George Adam, Benjamin Haibe-Kains, and Anna Goldenberg. Maintaining stability and plasticity for predictive churn reduction. *arXiv preprint arXiv:2305.04135*, 2023.

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

[3] Wasi Uddin Ahmad, Sean Narenthiran, Somshubra Majumdar, Aleksander Ficek, Siddhartha Jain, Jocelyn Huang, Vahid Noroozi, and Boris Ginsburg. OpenCodeReasoning: Advancing data distillation for competitive coding. In *Second Conference on Language Modeling (COLM)*, 2025.

[4] Daman Arora, Himanshu Gaurav Singh, and Mausam. Have LLMs advanced enough? a challenging problem solving benchmark for large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2023.

[5] Jacob Austin, Augustus Odena, Maxwell Nye, Maarten Bosma, Henryk Michalewski, David Dohan, Ellen Jiang, Carrie Cai, Michael Terry, Quoc Le, and Charles Sutton. Program synthesis with large language models. *arXiv preprint arXiv:2108.07732*, 2021.

[6] Dara Bahri and Heinrich Jiang. Locally adaptive label smoothing improves predictive churn. In *International Conference on Machine Learning (ICML)*, pages 532–542, 2021.

[7] Arslan Chaudhry, Marcus Rohrbach, Mohamed Elhoseiny, Thalaiyasingam Ajanthan, Puneet K. Dokania, Philip H. S. Torr, and Marc'Aurelio Ranzato. On tiny episodic memories in continual learning. *arXiv preprint arXiv:1902.10486*, 2019.

[8] Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021.

[9] Matthias De Lange, Gido van de Ven, and Tinne Tuytelaars. Continual evaluation for lifelong learning: Identifying the stability gap. In *International Conference on Learning Representations (ICLR)*, 2023.

[10] Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. Sharpness-aware minimization for efficiently improving generalization. In *International Conference on Learning Representations (ICLR)*, 2021.

[11] Gemma Team. Gemma 4 technical report. *arXiv preprint arXiv:2607.02770*, 2026.

[12] Ian J Goodfellow, Mehdi Mirza, Da Xiao, Aaron Courville, and Yoshua Bengio. An empirical investigation of catastrophic forgetting in gradient-based neural networks. *arXiv preprint arXiv:1312.6211*, 2013.

[13] Alex Gu, Baptiste Rożière, Hugh Leather, Armando Solar-Lezama, Gabriel Synnaeve, and Sida I. Wang. CRUXEval: A benchmark for code reasoning, understanding and execution. In *International Conference on Machine Learning (ICML)*, 2024.

[14] Etash Guha, Ryan Marten, Sedrick Keh, Negin Raoof, Georgios Smyrnis, Hritik Bansal, Marianna Nezhurina, Jean Mercat, Trung Vu, Zayne Sprague, et al. OpenThoughts: Data recipes for reasoning models. In *International Conference on Learning Representations (ICLR)*, 2026.

[15] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645:633–638, 2025. doi: 10.1038/s41586-025-09422-z.

[16] Jackson Harmon, Andreas Hochlehnert, Matthias Bethge, and Ameya Prabhu. Mapping post-training forgetting in language models at scale. In *NeurIPS 2025 Workshop on Continual and Compatible Foundation Model Updates*, 2025.

[17] Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding Hu, Zhen Leng Thai, Junhao Shen, Jinyi Hu, Xu Han, Yujie Huang, Yuxiang Zhang, et al. OlympiadBench: A challenging benchmark for promoting AGI with olympiad-level bilingual multimodal scientific problems. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*, 2024.

[18] Horace He and Thinking Machines Lab. Defeating nondeterminism in LLM inference. Thinking Machines Lab: Connectionism, `https://thinkingmachines.ai/blog/defeating-nondeterminism-in-llm-inference/`, 2025.

[19] Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the MATH dataset. In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, 2021.

[20] Naman Jain, King Han, Alex Gu, Wen-Ding Li, Fanjia Yan, Tianjun Zhang, Sida Wang, Armando Solar-Lezama, Koushik Sen, and Ion Stoica. LiveCodeBench: Holistic and contamination free evaluation of large language models for code. In *International Conference on Learning Representations (ICLR)*, 2025.

[21] James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114 (13):3521–3526, 2017.

[22] Yuetai Li, Zhangchen Xu, Fengqing Jiang, Bhaskar Ramasubramanian, Luyao Niu, Bill Yuchen Lin, Xiang Yue, and Radha Poovendran. Temporal sampling for forgotten reasoning in llms. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2026. URL https://aclanthology.org/2026.acl-long.1305.pdf. arXiv:2505.20196 (Temporal Sampling for Forgotten Reasoning in LLMs).

[23] Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *International Conference on Learning Representations (ICLR)*, 2024.

[24] Jiawei Liu, Chunqiu Steven Xia, Yuyao Wang, and Lingming Zhang. Is your code generated by ChatGPT really correct? rigorous evaluation of large language models for code generation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

[25] Zichen Liu, Changyu Chen, Wenjun Li, Penghui Qi, Tianyu Pang, Chao Du, Wee Sun Lee, and Min Lin. Understanding R1-Zero-like training: A critical perspective. In *Second Conference on Language Modeling (COLM)*, 2025.

[26] David Lopez-Paz and Marc'Aurelio Ranzato. Gradient episodic memory for continual learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

[27] Yun Luo, Zhen Yang, Fandong Meng, Yafu Li, Jie Zhou, and Yue Zhang. An empirical study of catastrophic forgetting in large language models during continual fine-tuning. *IEEE Transactions on Audio, Speech and Language Processing*, 33:3776–3786, 2025. doi: 10.1109/TASLPRO.2025.3606231.

[28] Sanket Vaibhav Mehta, Darshan Patil, Sarath Chandar, and Emma Strubell. An empirical investigation of the role of pre-training in lifelong learning. *Journal of Machine Learning Research*, 24(214):1–50, 2023.

[29] Seyed Iman Mirzadeh, Mehrdad Farajtabar, Razvan Pascanu, and Hassan Ghasemzadeh. Understanding the role of training regimes in continual learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[30] Ivan Moshkov, Darragh Hanley, Ivan Sorokin, Shubham Toshniwal, Christof Henkel, Benedikt Schifferer, Wei Du, and Igor Gitman. AIMO-2 winning solution: Building state-of-the-art mathematical reasoning models with OpenMathReasoning dataset. *arXiv preprint arXiv:2504.16891*, 2025.

[31] Seyed Mahed Mousavi, Simone Alghisi, and Giuseppe Riccardi. What does loss optimization actually teach, if anything? knowledge dynamics in continual pre-training of LLMs. *arXiv preprint arXiv:2601.03858*, 2026.

[32] NVIDIA. OpenScienceReasoning-2. https://huggingface.co/datasets/nvidia/OpenScienceReasoning-2, 2025.

[33] Qwen Team. Qwen3.5: Towards native multimodal agents. https://qwen.ai/blog?id=qwen3.5, February 2026.

[34] David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. GPQA: A graduate-level google-proof Q&A benchmark. In *First Conference on Language Modeling (COLM)*, 2024.

[35] David Rolnick, Arun Ahuja, Jonathan Schwarz, Timothy P. Lillicrap, and Greg Wayne. Experience replay for continual learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

[36] Tom Schaul, André Barreto, John Quan, and Georg Ostrovski. The phenomenon of policy churn. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

[37] Raphael Schumann, Elman Mansimov, Yi-An Lai, Nikolaos Pappas, Xibin Gao, and Yi Zhang. Backward compatibility during data updates by weight interpolation. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2846–2861, 2024.

[38] 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.

[39] Idan Shenfeld, Mehul Damani, Jonas Hübotter, and Pulkit Agrawal. Self-distillation enables continual learning. In *International Conference on Machine Learning*, 2026.

[40] Hongyao Tang and Glen Berseth. Improving deep reinforcement learning by reducing the chain effect of value and policy churn. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

[41] Mariya Toneva, Alessandro Sordoni, Remi Tachet des Combes, Adam Trischler, Yoshua Bengio, and Geoffrey J. Gordon. An empirical study of example forgetting during deep neural network learning. In *International Conference on Learning Representations (ICLR)*, 2019.

[42] Xiaoxuan Wang, Ziniu Hu, Pan Lu, Yanqiao Zhu, Jieyu Zhang, Satyen Subramaniam, Arjun R. Loomba, Shichang Zhang, Yizhou Sun, and Wei Wang. SciBench: Evaluating college-level scientific problem-solving abilities of large language models. In *International Conference on Machine Learning (ICML)*, 2024.

[43] Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*, 2022.

[44] Mitchell Wortsman, Gabriel Ilharco, Samir Yitzhak Gadre, Rebecca Roelofs, Raphael Gontijo-Lopes, Ari S. Morcos, Hongseok Namkoong, Ali Farhadi, Yair Carmon, Simon Kornblith, and Ludwig Schmidt. Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In *International Conference on Machine Learning (ICML)*, 2022.

[45] Mitchell Wortsman, Gabriel Ilharco, Jong Wook Kim, Mike Li, Simon Kornblith, Rebecca Roelofs, Raphael Gontijo-Lopes, Hannaneh Hajishirzi, Ali Farhadi, Hongseok Namkoong, and Ludwig Schmidt. Robust fine-tuning of zero-shot models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

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

[47] Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. Gradient surgery for multi-task learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[48] Junhao Zheng, Xidi Cai, Shengjie Qiu, and Qianli Ma. Spurious forgetting in continual learning of language models. In *International Conference on Learning Representations*, 2025.

APPENDIX CONTENTS

A Training configuration **13**

B Churn details **14**

B.1 Persistence across training 14

B.2 Churn changes little with model size 14

B.3 The groups are stable across halves of training 14

B.4 Per-dataset churn tables 14

B.5 Churn rate at every checkpoint 14

C Edge-of-understanding measurements **20**

D Churn dynamics under factual knowledge learning **23**

E Does reducing churn benefit accuracy? SFT ablation **24**

F Sensitivity to perturbation **25**

G Forgetting gap details **25**

G.1 Per-model tables 25

G.2 Per-model radar plots 33

G.3 Forgetting gap changes little with model size 37

G.4 Forgetting gap over training 37

H Correct final answers are recovered negative flips **40**

I AI usage **40**

A TRAINING CONFIGURATION

Table 4 lists the hyperparameters shared by all runs and the method-specific ones. The comparison across SFT, SDFT, RLVR, and OPD is *gradient-step matched*, not accuracy-gain matched: every run takes the same 500 updates with the same learning rate, is repeated over three random seeds, batch size of 128 responses per update, and generation budget, and RLVR uses no KL penalty to a reference policy ($\beta = 0$). The methods are, however, not data matched: RLVR and OPD train on the reasoning mixture and SFT and SDFT on OpenThoughts-114k (Section 3), and the responses come from the model itself for RLVR/OPD and from the reference for SFT/SDFT. Churn and forgetting-gap numbers are therefore reported per method at an equal number and size of updates, and we do not interpret differences between methods as effects of the training algorithm.

Compute resources. All training and evaluation runs were conducted on 2 NVIDIA H100 GPUs.

Table 4: Training configuration. Shared settings apply to every run of every method.

	Hyperparameter	Value
Shared	Learning rate	1×10^{-5}
	Batch size (responses per update)	128
	Updates	500 (checkpoint every 25)
	Max new tokens (training)	16,384
	Max new tokens (evaluation)	8,192
RLVR (Dr.GRPO)	Group size G	8 (16 distinct queries per update)
	Reward	binary verifier, no reward scaling
	KL penalty	none ($\beta = 0$)
OPD	Distillation loss	per-token reverse KL to the teacher, weight $\alpha = 1.0$
	Logits distilled	top-K of the student's distribution, $K = 64$
SFT	Loss	token-level cross-entropy on the reference response
SDFT	Teacher	the frozen base model conditioned on the reference response

B Churn details

B.1 Persistence across training

Under RLVR and OPD the churn rate is almost constant over the run. For Qwen3-1.7B, the negative churn rate measured over the churning group alone (rather than over L) is 16.0% between the first two checkpoints and 14.1% between the last two; for Qwen3-4B it stays between 12.2% and 16.5%. Under SFT and SDFT the first few checkpoints churn more than the rest, mostly through negative churn, after which the churn rate is flat (Figure 5).

B.2 Churn changes little with model size

Figure 9 plots the churn rate against model size, one panel per OOD dataset and one line per method. Models are grouped into three size classes, $\sim$2B (Qwen3-1.7B, Qwen3.5-2B, Gemma4-E2B), 4B (Qwen3-4B, Qwen3.5-4B, Gemma4-E4B), and $\sim$8B (Qwen3-8B, Qwen3.5-9B), and every 25-update checkpoint interval of every model in a class contributes one measurement. Going from $\sim$2B to $\sim$8B, the churn rate averaged over the nine datasets drops by 2.6 points under RLVR ($17.7\% \to 15.1\%$), 2.2 under SFT ($24.4\% \to 22.2\%$), and 4.4 under SDFT ($24.4\% \to 20.0\%$); the OPD students are all $\sim$2B. Fitting churn against $\log_2$ size for each of the 27 method–dataset pairs gives a negative slope in 21 cases, but only one is significant at $p < 0.05$ (SDFT on MATH500), and on the code datasets the curves are flat or rise. A $4\times$ increase in parameters therefore removes at most a fifth of the churn, while the choice of method moves it by up to 25 points on the same dataset (HumanEval+).

B.3 The groups are stable across halves of training

The churning group is defined by changes in correctness, so churn concentrates in it by construction. To rule out this circularity, we define the groups on the first half of training only and measure churn on the second half. Queries marked churning in the first half still churn at 28%–31% in the second half, while unforgettable and unsolvable queries churn at 2%–7%, a 4–12$\times$ separation.

B.4 Per-dataset churn tables

Tables 5–8 give, for every model and dataset, the mean positive (W$\to$C), negative (C$\to$W), and total churn rate over consecutive checkpoints, one table per method; Figure 1 averages these over models.

B.5 Churn rate at every checkpoint

Figures 10–13 plot the churn rate between every two consecutive checkpoints, for every dataset and every model, one figure per method. No curve trends down over training. The noisiest panels (GPQA-Diamond, HumanEval+) have the smallest learnable sets, so a single flipped query moves

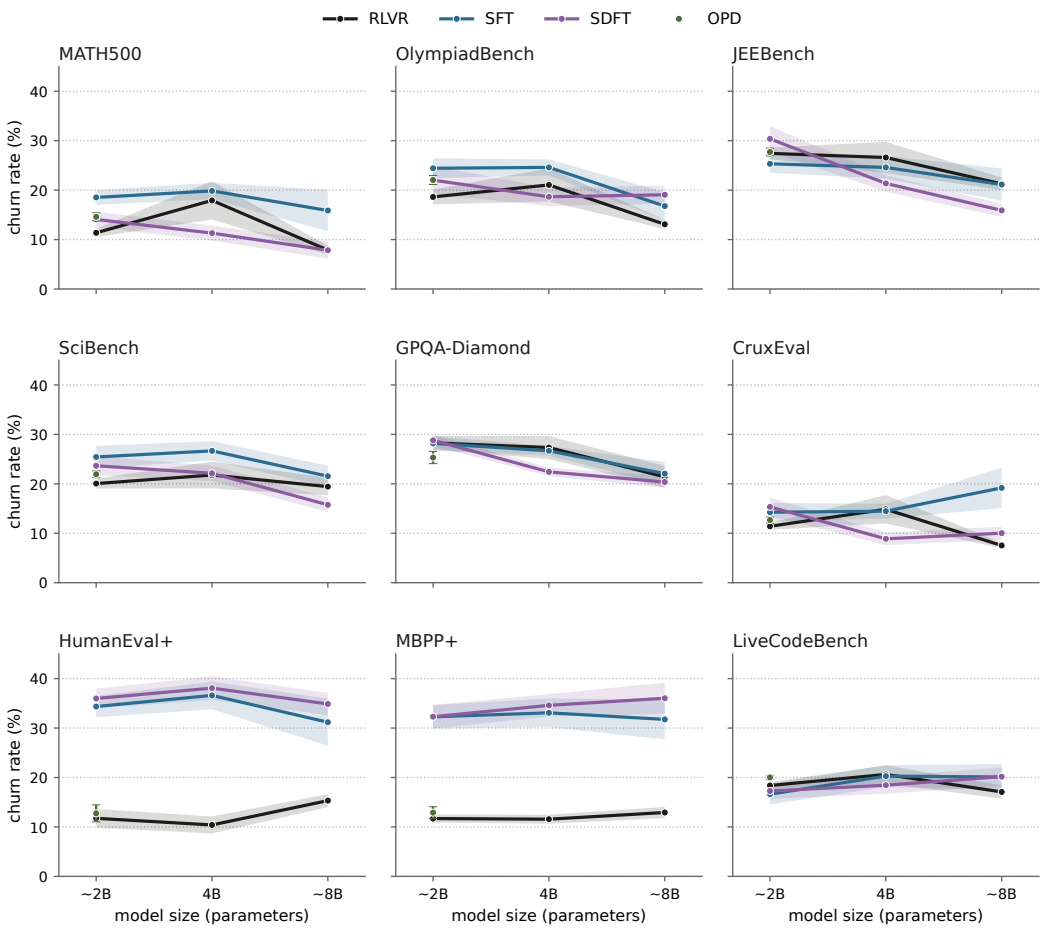

Figure 9: **Churn rate versus model size, per OOD dataset.** Each point is the mean churn rate (Equation 3) over all consecutive 25-update checkpoint intervals of all models in a size class (∼2B: 1.7–2B models; 4B; ∼8B: 8–9B models); bands and whiskers are 95% confidence intervals over those intervals. OPD is a single point because every student is ∼2B.

the rate by up to 0.7 points. Under SFT and SDFT the OOD datasets were evaluated on subsets (100 queries each; 150 for MATH500, 75 for LiveCodeBench, all of GPQA-Diamond), which makes those curves noisier than the RLVR and OPD ones.

Model		ID	MATH500	Olympiad	JEE	SciBench	GPQA-D	CruxEval	HE+	MBPP+	LCB	OOD mean
Qwen3-1.7B	W→C	11.4 ±1.0	5.6 ±1.0	9.2 ±1.5	14.2 ±1.3	11.9 ±1.3	12.8 ±3.1	6.2 ±1.1	10.7 ±3.5	6.9 ±1.4	9.6 ±1.3	9.7
	C→W	11.5 ±1.9	5.5 ±1.4	9.2 ±1.0	14.2 ±2.6	11.9 ±1.2	12.8 ±5.0	6.2 ±0.8	10.6 ±3.8	6.9 ±1.1	9.3 ±1.4	9.6
	Churn	**22.9** ±2.4	**11.2** ±1.0	**18.4** ±1.9	**28.4** ±3.0	**23.8** ±2.5	**25.7** ±4.1	**12.3** ±1.6	**21.3** ±3.4	**13.7** ±2.1	**18.9** ±1.6	**19.3**
Qwen3-4B	W→C	8.5 ±1.0	4.6 ±0.9	6.6 ±1.3	12.5 ±2.1	10.3 ±1.1	11.6 ±3.0	4.6 ±1.0	5.7 ±3.1	6.7 ±0.8	9.6 ±0.9	8.0
	C→W	8.5 ±1.4	4.5 ±1.6	6.5 ±1.6	12.2 ±4.6	10.3 ±1.1	11.9 ±3.8	4.6 ±1.2	5.8 ±3.6	6.4 ±2.6	9.6 ±1.7	8.0
	Churn	**17.0** ±2.0	**9.1** ±2.0	**13.1** ±2.6	**24.7** ±6.0	**20.6** ±2.1	**23.5** ±5.7	**9.2** ±2.1	**11.4** ±5.4	**13.1** ±2.6	**19.1** ±1.9	**16.0**
Qwen3-8B	W→C	9.6 ±0.7	4.0 ±0.9	6.5 ±1.1	10.7 ±1.6	9.5 ±2.7	10.5 ±3.2	3.8 ±0.7	7.7 ±1.6	6.5 ±2.0	8.6 ±1.8	7.5
	C→W	9.6 ±1.5	4.0 ±1.4	6.5 ±1.1	10.6 ±1.7	9.8 ±2.7	10.7 ±3.8	3.8 ±0.7	7.7 ±2.2	6.5 ±3.2	8.4 ±3.6	7.6
	Churn	**19.2** ±2.0	**8.0** ±2.0	**13.0** ±2.1	**21.2** ±3.0	**19.3** ±4.1	**21.2** ±5.1	**7.5** ±1.3	**15.4** ±3.0	**13.0** ±2.6	**17.1** ±3.2	**15.1**
Qwen3.5-2B	W→C	10.7 ±0.8	6.4 ±1.4	11.3 ±1.0	14.7 ±3.2	9.3 ±1.5	15.9 ±3.0	6.2 ±0.9	3.6 ±2.4	6.5 ±1.2	9.3 ±1.5	9.2
	C→W	10.6 ±1.9	6.4 ±1.9	11.2 ±2.5	14.6 ±3.4	9.4 ±2.4	16.0 ±3.4	6.0 ±2.8	3.7 ±2.7	6.5 ±2.0	9.2 ±4.8	9.2
	Churn	**21.3** ±2.3	**12.8** ±3.0	**22.5** ±2.5	**29.2** ±5.0	**18.7** ±3.5	**32.0** ±5.5	**12.3** ±3.1	**7.3** ±4.2	**13.0** ±2.5	**18.5** ±4.3	**18.5**
Qwen3.5-4B	W→C	8.2 ±1.0	19.5 ±2.1	19.7 ±1.4	19.7 ±6.8	17.5 ±3.4	17.8 ±6.6	15.3 ±1.6	8.1 ±2.5	6.9 ±1.8	14.7 ±2.2	15.5
	C→W	8.1 ±2.0	19.3 ±2.9	19.5 ±1.8	19.4 ±7.6	17.0 ±4.8	17.5 ±7.3	14.9 ±2.8	7.9 ±3.8	6.7 ±1.9	14.4 ±3.9	15.2
	Churn	**16.3** ±2.7	**38.7** ±4.0	**39.1** ±2.7	**39.1** ±13.5	**34.5** ±7.3	**35.3** ±11.1	**30.2** ±3.7	**16.1** ±5.5	**13.6** ±2.8	**29.1** ±5.3	**30.6**
Gemma4-E2B	W→C	9.8 ±1.1	5.2 ±2.8	7.5 ±4.5	12.5 ±1.8	8.8 ±1.3	13.8 ±2.3	4.9 ±1.1	3.3 ±1.1	4.2 ±1.2	8.9 ±1.0	7.7
	C→W	9.8 ±2.0	4.7 ±2.5	6.8 ±5.3	12.1 ±4.0	8.7 ±2.4	13.6 ±4.9	4.5 ±1.3	3.3 ±1.3	4.2 ±1.0	9.0 ±1.5	7.4
	Churn	**19.6** ±2.8	**9.9** ±4.8	**14.3** ±8.1	**24.7** ±5.0	**17.5** ±3.0	**27.4** ±5.0	**9.4** ±2.2	**6.6** ±2.0	**8.4** ±1.9	**17.9** ±2.0	**15.1**
Gemma4-E4B	W→C	7.5 ±0.9	3.0 ±1.6	5.3 ±1.1	8.3 ±1.9	5.4 ±1.2	11.4 ±1.9	2.6 ±0.5	1.9 ±1.2	4.0 ±1.5	6.7 ±1.2	5.4
	C→W	7.6 ±2.6	2.7 ±2.0	5.3 ±2.1	8.0 ±2.4	5.3 ±2.3	11.7 ±2.3	2.4 ±0.9	1.9 ±1.1	4.0 ±1.4	7.0 ±2.4	5.4
	Churn	**15.1** ±3.0	**5.7** ±2.9	**10.5** ±2.3	**16.4** ±3.0	**10.6** ±2.7	**23.1** ±3.6	**5.0** ±1.1	**3.7** ±2.3	**8.1** ±2.5	**13.7** ±3.0	**10.8**

Table 5: **Churn rate under RLVR.** Mean over consecutive checkpoints of the fraction of learnable queries flipping incorrect→correct (W→C, positive churn) and correct→incorrect (C→W, negative churn), in percentage points; **Churn** is their sum. ± is the standard deviation over checkpoint intervals. Columns as in Table 18. All runs use 25-step checkpoint intervals; the few pairs spanning 125 steps (where intermediate checkpoints are missing) are excluded, so every cell is a single-interval measurement.

Model		ID	MATH500	Olympiad	JEE	SciBench	GPQA-D	CruxEval	HE+	MBPP+	LCB	OOD mean
Qwen3-1.7B	W→C	14.3 ±4.0	8.8 ±2.2	11.9 ±5.1	12.3 ±5.5	15.0 ±6.1	15.6 ±4.1	10.0 ±3.0	13.4 ±4.3	12.3 ±6.1	9.8 ±4.2	12.1
	C→W	14.1 ±5.6	8.9 ±2.4	12.2 ±4.4	13.9 ±6.1	15.4 ±6.8	15.4 ±3.9	10.5 ±3.2	15.5 ±6.4	13.8 ±8.1	11.1 ±5.9	13.0
	Churn	**28.3** ±6.8	**17.7** ±3.2	**24.1** ±7.8	**26.2** ±6.3	**30.3** ±5.7	**31.1** ±6.1	**20.4** ±4.7	**28.8** ±7.3	**26.1** ±7.3	**20.8** ±6.8	**25.1**
Qwen3-4B	W→C	22.1 ±5.3	8.3 ±2.2	11.1 ±2.9	7.9 ±3.2	14.1 ±4.2	13.4 ±2.0	5.1 ±1.6	18.5 ±10.0	18.8 ±6.7	11.1 ±1.9	12.1
	C→W	20.6 ±7.8	8.6 ±3.6	11.0 ±3.8	7.9 ±4.7	14.0 ±4.4	13.2 ±3.8	5.5 ±2.3	19.6 ±14.1	18.9 ±7.5	10.9 ±5.2	12.2
	Churn	**42.7** ±11.8	**16.9** ±5.2	**22.1** ±5.7	**15.9** ±6.2	**28.2** ±8.0	**26.6** ±4.6	**10.6** ±2.9	**38.1** ±15.1	**37.7** ±12.7	**22.0** ±5.2	**24.2**
Qwen3-8B	W→C	20.8 ±1.3	5.5 ±1.5	8.0 ±3.1	12.3 ±4.6	12.3 ±2.0	11.7 ±2.6	4.6 ±2.0	16.7 ±8.3	16.9 ±5.4	10.9 ±4.0	11.0
	C→W	20.2 ±4.6	5.5 ±2.0	7.5 ±2.9	12.1 ±9.2	11.1 ±6.2	11.5 ±2.3	4.5 ±3.8	17.5 ±15.2	18.1 ±9.7	11.0 ±5.1	11.0
	Churn	**41.0** ±5.0	**11.0** ±3.2	**15.5** ±5.3	**24.4** ±11.1	**23.4** ±6.5	**23.2** ±4.1	**9.1** ±4.4	**34.3** ±16.5	**35.0** ±10.6	**21.9** ±8.1	**22.0**
Qwen3.5-2B	W→C	16.0 ±4.8	9.8 ±2.0	11.2 ±4.7	13.8 ±5.6	13.6 ±3.3	15.0 ±3.9	7.7 ±3.0	19.1 ±10.5	16.0 ±7.3	8.2 ±6.3	12.7
	C→W	15.3 ±5.6	9.8 ±2.9	11.3 ±5.3	13.2 ±4.1	14.0 ±3.4	14.8 ±3.0	8.0 ±3.0	16.9 ±6.3	18.1 ±9.3	8.9 ±4.5	12.8
	Churn	**31.3** ±6.9	**19.6** ±4.1	**22.5** ±6.2	**26.9** ±7.6	**27.6** ±5.4	**29.7** ±4.2	**15.7** ±4.3	**36.0** ±9.7	**34.1** ±8.0	**17.2** ±6.9	**25.5**
Qwen3.5-4B	W→C	13.3 ±4.2	12.1 ±4.7	14.6 ±4.8	14.0 ±3.9	16.3 ±4.6	13.3 ±2.9	9.2 ±3.9	15.8 ±9.2	13.9 ±7.2	9.8 ±6.4	13.2
	C→W	13.4 ±4.0	10.4 ±1.8	13.2 ±4.1	13.5 ±3.8	15.2 ±3.4	12.8 ±2.8	8.8 ±3.1	16.1 ±9.0	13.3 ±6.4	10.2 ±5.5	12.6
	Churn	**26.7** ±4.6	**22.5** ±5.4	**27.8** ±6.6	**27.5** ±7.0	**31.5** ±4.7	**26.1** ±2.8	**18.0** ±5.2	**31.8** ±8.1	**27.2** ±6.1	**20.0** ±10.8	**25.8**
Qwen3.5-9B	W→C	14.4 ±3.4	11.3 ±10.2	9.7 ±7.6	8.5 ±4.4	10.4 ±3.0	10.7 ±4.6	15.2 ±5.9	13.6 ±8.3	13.6 ±7.5	9.1 ±4.4	11.3
	C→W	14.5 ±9.2	9.9 ±8.4	8.4 ±5.4	8.2 ±4.3	9.3 ±6.3	10.1 ±6.0	14.8 ±6.2	13.8 ±13.2	12.9 ±11.4	9.4 ±5.6	10.8
	Churn	**28.9** ±11.8	**21.1** ±18.2	**18.1** ±11.9	**16.6** ±8.3	**19.8** ±6.9	**20.8** ±9.9	**29.9** ±11.2	**27.4** ±14.4	**26.5** ±12.8	**18.4** ±8.9	**22.1**
Gemma4-E2B	W→C	1.8 ±2.3	9.1 ±7.3	13.5 ±7.7	11.0 ±4.3	8.7 ±6.6	11.6 ±2.2	2.8 ±1.6	18.9 ±5.6	16.3 ±6.7	3.6 ±2.9	10.6
	C→W	6.9 ±18.9	7.9 ±4.4	11.8 ±5.8	10.5 ±4.0	7.6 ±2.7	12.0 ±2.0	3.1 ±2.2	18.8 ±4.3	17.6 ±7.9	6.3 ±5.7	10.6
	Churn	**8.7** ±18.7	**17.0** ±8.0	**25.3** ±8.7	**21.5** ±5.3	**16.2** ±6.8	**23.6** ±2.3	**5.9** ±2.9	**37.7** ±5.4	**33.9** ±8.5	**9.9** ±5.2	**21.2**
Gemma4-E4B	W→C	1.2 ±2.0	10.5 ±6.6	12.5 ±5.4	14.5 ±5.4	10.5 ±5.1	13.0 ±2.9	6.2 ±3.2	17.7 ±7.6	14.9 ±4.9	7.0 ±4.2	11.9
	C→W	6.2 ±18.1	8.5 ±3.5	11.3 ±4.7	14.9 ±6.4	9.1 ±3.0	13.7 ±3.5	7.1 ±4.8	20.8 ±9.1	17.0 ±8.5	9.8 ±6.2	12.5
	Churn	**7.5** ±18.7	**19.0** ±5.7	**23.7** ±6.1	**29.4** ±6.7	**19.6** ±4.7	**26.7** ±3.9	**13.4** ±4.6	**38.5** ±7.2	**31.9** ±8.1	**16.8** ±7.0	**24.3**

Table 6: **Churn rate under SFT.** Mean over consecutive checkpoints of the fraction of learnable queries flipping incorrect→correct (W→C, positive churn) and correct→incorrect (C→W, negative churn), in percentage points; **Churn** is their sum. ± is the standard deviation over checkpoint intervals. Columns as in Table 18. The ID column is averaged over the same consecutive checkpoints but pooled across domains, so it is not a mean over per-dataset values like the OOD columns. All eight runs save a checkpoint every 25 updates.

Model		ID	MATH500	Olympiad	JEE	SciBench	GPQA-D	CruxEval	HE+	MBPP+	LCB	OOD mean
	W→C	14.7 ±4.9	4.6 ±2.3	11.0 ±3.6	15.3 ±4.2	12.5 ±4.8	15.2 ±2.6	11.8 ±4.0	17.7 ±7.8	15.4 ±6.4	9.5 ±3.2	12.6
Qwen3-1.7B	C→W	14.5 ±3.8	4.6 ±2.1	11.1 ±4.4	15.9 ±5.9	12.2 ±4.6	15.3 ±2.0	11.7 ±3.7	18.4 ±6.5	16.7 ±6.9	10.1 ±4.4	12.9
	Churn	**29.2** ±6.5	**9.3** ±1.8	**22.0** ±5.2	**31.2** ±7.2	**24.8** ±5.4	**30.5** ±3.2	**23.5** ±4.8	**36.0** ±6.9	**32.0** ±8.8	**19.5** ±5.6	**25.4**
	W→C	15.1 ±5.2	2.8 ±1.5	8.8 ±3.2	10.5 ±4.1	8.7 ±2.6	11.7 ±3.1	5.3 ±1.7	18.7 ±8.3	15.5 ±6.4	10.5 ±3.9	10.3
Qwen3-4B	C→W	14.0 ±4.6	2.9 ±1.1	9.1 ±3.0	10.5 ±3.7	8.1 ±2.8	11.7 ±2.8	5.3 ±2.2	19.7 ±7.3	17.3 ±9.4	10.9 ±4.3	10.6
	Churn	**29.1** ±4.6	**5.7** ±2.1	**17.9** ±4.2	**21.0** ±5.1	**16.8** ±2.9	**23.3** ±3.1	**10.6** ±2.9	**38.5** ±8.6	**32.7** ±9.3	**21.5** ±4.7	**20.9**
	W→C	15.7 ±5.4	2.3 ±1.2	9.6 ±3.3	8.5 ±3.1	7.2 ±3.4	10.9 ±3.0	3.5 ±1.5	15.5 ±7.1	15.1 ±9.3	9.7 ±5.3	9.2
Qwen3-8B	C→W	15.0 ±4.0	2.2 ±0.9	9.2 ±2.6	8.7 ±2.6	6.9 ±3.0	10.9 ±2.9	3.9 ±1.9	16.3 ±7.2	18.1 ±12.6	9.8 ±5.7	9.6
	Churn	**30.7** ±4.9	**4.6** ±1.3	**18.8** ±4.0	**17.3** ±3.8	**14.1** ±4.2	**21.8** ±3.0	**7.4** ±2.1	**31.9** ±7.9	**33.2** ±10.5	**19.5** ±6.3	**18.7**
	W→C	15.9 ±5.9	8.3 ±1.9	11.5 ±3.4	17.0 ±6.3	10.3 ±4.0	13.3 ±2.8	6.9 ±1.7	16.8 ±10.0	13.1 ±7.5	8.7 ±4.2	11.8
Qwen3.5-2B	C→W	15.7 ±5.4	8.4 ±2.8	12.1 ±3.6	18.0 ±4.3	9.9 ±3.5	13.8 ±2.7	7.5 ±2.8	17.2 ±8.4	15.3 ±9.8	9.7 ±6.7	12.4
	Churn	**31.6** ±7.3	**16.7** ±3.3	**23.5** ±6.3	**35.1** ±5.6	**20.2** ±4.3	**27.0** ±3.6	**14.5** ±3.4	**34.0** ±7.9	**28.3** ±8.4	**18.4** ±5.7	**24.2**
	W→C	18.2 ±8.0	5.9 ±2.3	9.4 ±4.1	8.6 ±2.9	11.4 ±5.0	10.0 ±3.0	6.1 ±3.3	17.0 ±9.1	16.1 ±6.8	9.5 ±4.4	10.4
Qwen3.5-4B	C→W	18.1 ±6.8	5.4 ±1.9	8.5 ±3.5	8.6 ±2.5	10.9 ±3.7	9.5 ±2.1	5.4 ±2.1	18.0 ±10.7	17.9 ±10.5	10.3 ±6.1	10.5
	Churn	**36.2** ±6.5	**11.3** ±2.6	**17.9** ±3.8	**17.3** ±3.3	**22.3** ±5.4	**19.5** ±2.5	**11.4** ±4.2	**34.9** ±10.3	**33.9** ±8.2	**19.7** ±7.3	**20.9**
	W→C	15.7 ±5.7	5.7 ±4.9	9.9 ±6.2	7.3 ±3.8	8.9 ±3.8	9.2 ±2.0	6.3 ±2.8	18.7 ±8.6	17.2 ±4.6	10.3 ±5.1	10.4
Qwen3.5-9B	C→W	15.6 ±4.4	4.4 ±1.7	8.0 ±2.6	6.5 ±3.0	7.6 ±3.2	8.9 ±2.2	6.0 ±2.7	19.0 ±7.3	19.3 ±7.5	10.5 ±7.0	10.0
	Churn	**31.3** ±5.1	**10.1** ±4.9	**17.9** ±5.9	**13.9** ±3.7	**16.5** ±4.3	**18.2** ±2.9	**12.2** ±4.2	**37.7** ±5.0	**36.6** ±7.1	**20.9** ±6.0	**20.4**
	W→C	11.6 ±6.9	6.9 ±4.2	7.0 ±8.5	11.3 ±7.0	11.9 ±3.6	13.4 ±3.4	3.8 ±2.3	18.3 ±8.6	15.8 ±4.2	5.6 ±4.1	10.5
Gemma4-E2B	C→W	12.5 ±6.4	8.3 ±6.4	10.5 ±12.7	11.5 ±8.2	13.2 ±6.7	14.1 ±3.5	3.7 ±2.1	18.4 ±7.1	17.7 ±7.7	7.5 ±4.7	11.7
	Churn	**24.1** ±8.3	**15.2** ±7.2	**17.5** ±14.1	**22.7** ±11.2	**25.1** ±8.7	**27.6** ±5.0	**7.5** ±1.7	**36.7** ±8.4	**33.5** ±7.9	**13.1** ±6.1	**22.1**
	W→C	7.5 ±2.9	8.5 ±5.7	11.3 ±6.2	12.3 ±6.6	13.8 ±3.2	11.9 ±2.6	1.8 ±1.1	18.0 ±7.6	17.2 ±6.3	5.7 ±3.8	11.2
Gemma4-E4B	C→W	7.6 ±4.0	7.4 ±5.5	10.0 ±9.2	11.9 ±5.0	12.8 ±3.9	12.0 ±2.7	1.8 ±1.5	21.0 ±12.0	17.6 ±9.0	7.8 ±5.2	11.4
	Churn	**15.1** ±4.2	**15.9** ±5.5	**21.3** ±11.4	**24.2** ±7.2	**26.5** ±4.6	**23.9** ±2.8	**3.6** ±1.4	**39.0** ±8.1	**34.8** ±8.0	**13.6** ±6.2	**22.5**

Table 7: **Churn rate under SDFT.** Mean over consecutive checkpoints of the fraction of learnable queries flipping incorrect→correct (W→C, positive churn) and correct→incorrect (C→W, negative churn), in percentage points; **Churn** is their sum. ± is the standard deviation over checkpoint intervals. Columns as in Table 18. The ID column is averaged over the same consecutive checkpoints but pooled across domains, so it is not a mean over per-dataset values like the OOD columns. All eight runs save a checkpoint every 25 updates.

Model		ID	MATH500	Olympiad	JEE	SciBench	GPQA-D	CruxEval	HE+	MBPP+	LCB	OOD mean
	W→C	11.1 ±0.8	6.6 ±1.0	9.4 ±1.5	13.9 ±2.3	11.3 ±1.7	11.3 ±4.8	7.4 ±0.7	10.5 ±2.4	7.9 ±1.2	10.2 ±1.5	9.8
Qwen3-8B→1.7B	C→W	11.2 ±1.0	6.6 ±1.2	9.8 ±1.9	14.3 ±2.5	11.7 ±2.8	11.6 ±5.7	7.4 ±0.9	10.4 ±2.8	8.1 ±1.5	10.2 ±1.2	10.0
	Churn	**22.3** ±1.0	**13.2** ±1.4	**19.2** ±1.5	**28.2** ±2.9	**23.0** ±2.7	**22.9** ±5.8	**14.8** ±1.2	**21.0** ±3.3	**15.9** ±1.3	**20.5** ±1.3	**19.8**
	W→C	10.5 ±1.5	6.1 ±1.6	10.1 ±1.8	13.2 ±2.1	10.2 ±1.4	12.4 ±3.2	6.7 ±1.3	4.2 ±2.0	7.3 ±3.0	10.2 ±1.6	8.9
Qwen3.5-9B→2B	C→W	10.6 ±1.6	6.0 ±1.4	10.4 ±2.5	13.8 ±2.6	10.3 ±1.5	12.8 ±2.5	6.5 ±1.0	4.5 ±2.5	7.1 ±1.5	10.1 ±2.0	9.1
	Churn	**21.0** ±2.1	**12.0** ±1.7	**20.5** ±2.1	**27.0** ±2.8	**20.5** ±1.6	**25.2** ±3.9	**13.1** ±1.2	**8.8** ±3.7	**14.4** ±3.7	**20.3** ±2.2	**18.0**
	W→C	11.2 ±1.0	9.0 ±1.5	13.0 ±1.7	13.6 ±2.4	10.9 ±1.7	13.6 ±2.9	4.9 ±0.8	3.9 ±1.7	3.7 ±1.3	9.6 ±1.3	9.1
Gemma4-E4B→E2B	C→W	11.2 ±1.1	9.1 ±1.3	13.1 ±1.8	13.3 ±1.9	10.8 ±1.3	13.4 ±2.4	4.8 ±1.0	3.9 ±2.0	3.8 ±1.5	9.6 ±1.2	9.1
	Churn	**22.4** ±1.5	**18.1** ±1.8	**26.1** ±1.8	**26.9** ±2.5	**21.7** ±2.4	**27.0** ±3.5	**9.7** ±1.2	**7.9** ±1.6	**7.5** ±2.1	**19.2** ±1.3	**18.2**

Table 8: **Churn rate under OPD.** Mean over consecutive checkpoints of the fraction of learnable queries flipping incorrect→correct (W→C, positive churn) and correct→incorrect (C→W, negative churn), in percentage points; **Churn** is their sum. ± is the standard deviation over checkpoint intervals. Columns as in Table 18.

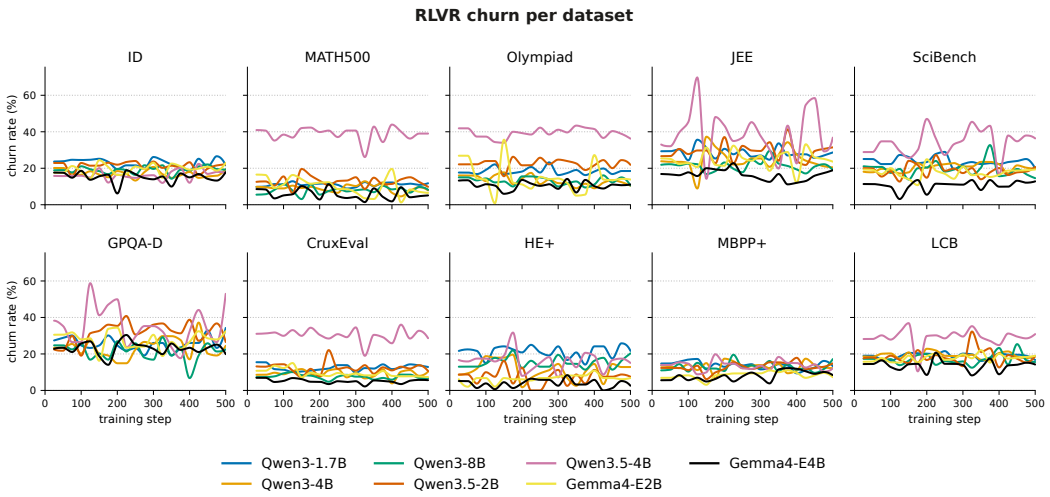

Figure 10: **Churn rate between consecutive checkpoints under RLVR**, per dataset and model (Equation 3).

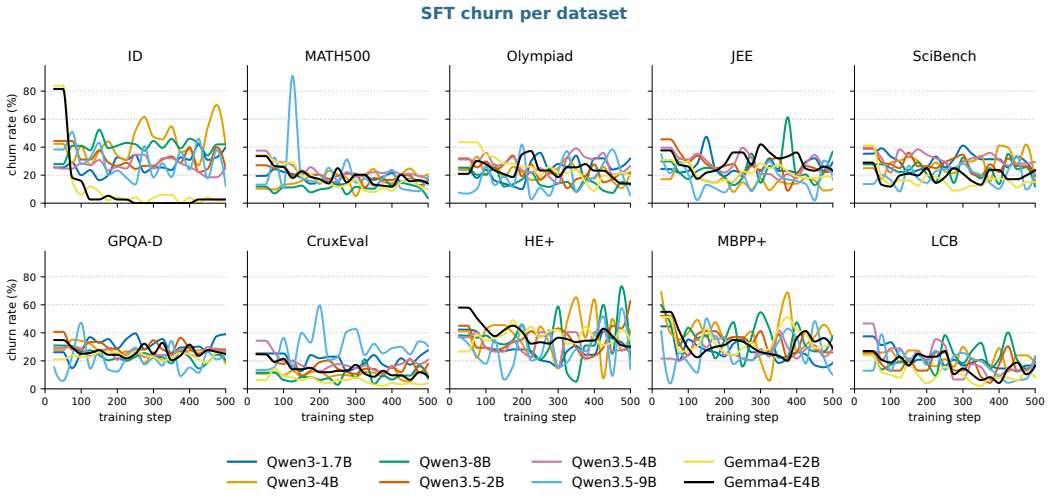

Figure 11: **Churn rate between consecutive checkpoints under SFT.** As Figure 10.

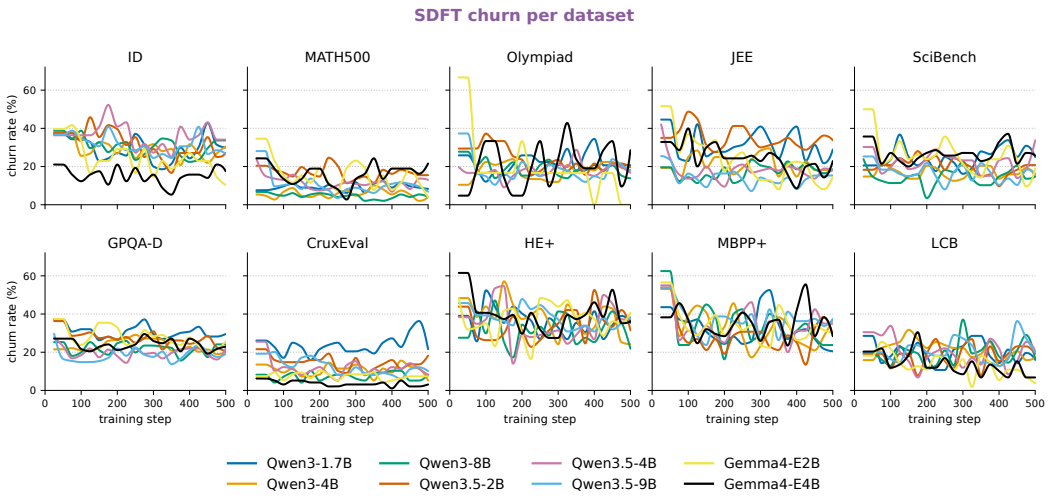

Figure 12: **Churn rate between consecutive checkpoints under SDFT.** As Figure 10.

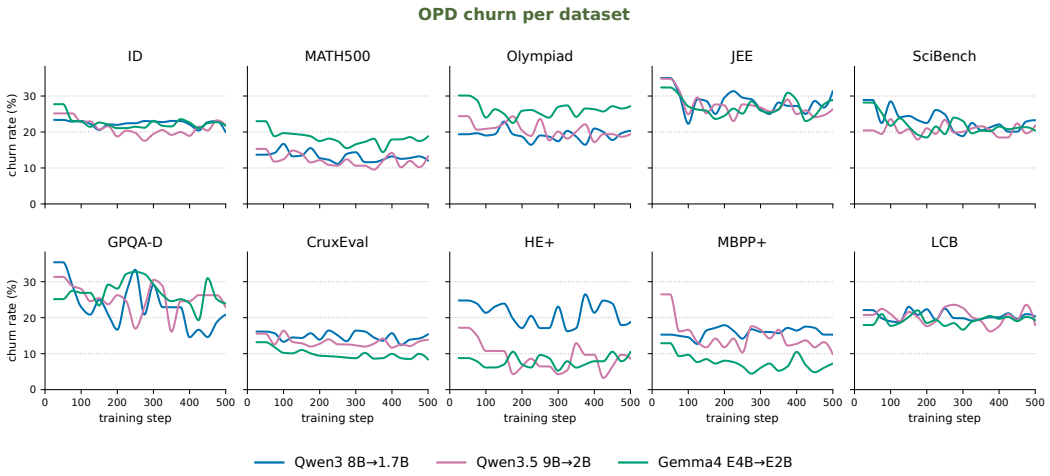

Figure 13: **Churn rate between consecutive checkpoints under OPD.** As Figure 10; one line per teacher→student pair.

C EDGE-OF-UNDERSTANDING MEASUREMENTS

Tables 9 and 10 give the numbers behind Section 3.3 for all four methods on the nine OOD datasets, grouped into three domains (Code: CruxEval, HumanEval+, MBPP+, LiveCodeBench; Math: MATH500, OlympiadBench; Science: JEEBench, SciBench, GPQA-Diamond) and averaged over the datasets of a domain. Groups are decided from each run's own greedy responses across all 20 checkpoints, so the unsolvable/correct cell of Table 9 is empty by definition, and cells in which a group produced no response of the stated correctness are left blank. Log-probabilities are per token because response length differs by up to $9\times$ between groups. Over the 78 method–model–domain cells, the churning group's margin lies between the unforgettable and unsolvable groups' in 67 cells and above the unsolvable group's in 77; unforgettable responses are held more confidently than churning correct responses in 76 cells. In Figures 14–17, cells with a missing group are completed by Gaussian imputation from the fitted group distributions, so those figures are descriptive; Table 2 uses only cells in which all three groups are measured.

		Code				Math				Science			
Method	Model	$\mathcal{U}_t$ (C)	$\mathcal{C}_t$ (C)	$\mathcal{C}_t$ (W)	$\mathcal{N}_t$ (W)	$\mathcal{U}_t$ (C)	$\mathcal{C}_t$ (C)	$\mathcal{C}_t$ (W)	$\mathcal{N}_t$ (W)	$\mathcal{U}_t$ (C)	$\mathcal{C}_t$ (C)	$\mathcal{C}_t$ (W)	$\mathcal{N}_t$ (W)
RLVR	Qwen3-1.7B	-0.040	-0.067	-0.058	-0.062	-0.059	-0.100	-0.117	-0.115	-0.117	-0.152	-0.120	-0.118
	Qwen3-4B	-0.035	-0.076	-0.065	-0.063	-0.115	-0.157	-0.147	-0.138	-0.099	-0.115	-0.108	-0.110
	Qwen3-8B	-0.058	-0.095	-0.096	-0.091	-0.112	-0.151	-0.150	-0.143	-0.108	-0.133	-0.130	-0.126
	Qwen3.5-2B	-0.086	-0.155	-0.162	-0.096	-0.081	-0.138	-0.151	-0.120	-0.110	-0.238	-0.188	-0.147
	Qwen3.5-4B	-0.067	-0.145	-0.140	-0.051	-0.069	-0.127	-0.169	-0.076	-0.089	-0.128	-0.088	-0.153
	Gemma4-E2B	-0.040	-0.083	-0.096	-0.081	-0.039	-0.084	-0.128	-0.098	-0.065	-0.143	-0.162	-0.133
	Gemma4-E4B	-0.065	-0.098	-0.113	-0.093	-0.046	-0.067	-0.093	-0.082	-0.102	-0.192	-0.248	-0.206
SFT	Qwen3-1.7B	-0.106	-0.116	-0.080	-0.075	-0.034	-0.045	–	–	-0.118	-0.122	-0.209	-0.193
	Qwen3-4B	-3.121	-4.032	-0.728	-0.611	-5.658	-5.460	-5.431	-4.867	-4.636	-4.679	-4.635	-4.544
	Qwen3-8B	-0.056	-0.091	-0.028	-0.011	-0.059	-0.068	-0.070	-0.059	-0.020	-0.028	–	–
	Qwen3.5-2B	-0.090	-0.247	-0.172	-0.185	-0.052	-0.128	-0.161	–	-0.113	-0.252	-0.220	-0.211
	Qwen3.5-4B	-0.018	-0.138	-0.195	-0.028	-0.046	-0.141	-0.011	–	-0.020	-0.148	-0.109	-0.062
	Qwen3.5-9B	-0.021	-0.174	-0.149	-0.094	-0.065	-0.148	–	–	-0.044	-0.229	-0.136	-0.074
	Gemma4-E2B	-0.037	-0.101	-0.153	-0.130	-0.018	-0.121	-0.166	-0.143	-0.022	-0.073	-0.119	-0.099
	Gemma4-E4B	-0.035	-0.132	-0.203	-0.149	-0.037	-0.134	-0.184	-0.161	-0.026	-0.107	-0.140	-0.098
SDFT	Qwen3-1.7B	-0.021	-0.045	-0.034	-0.027	-0.063	-0.079	-0.090	-0.081	-0.032	-0.041	-0.049	-0.045
	Qwen3-4B	-0.702	-6.499	-1.191	-1.009	-9.097	-7.796	-7.462	-6.817	-8.610	-8.760	-8.398	-8.568
	Qwen3-8B	-0.044	-0.110	-0.053	-0.045	-0.005	-0.027	–	–	-0.027	-0.046	-0.019	-0.027
	Qwen3.5-2B	-0.071	-0.191	-0.190	-0.168	-0.029	-0.101	-0.164	-0.038	-0.012	-0.117	-0.150	-0.092
	Qwen3.5-4B	-0.021	-0.109	-0.181	-0.050	-0.016	-0.074	-0.225	-0.182	-0.010	-0.106	-0.080	-0.052
	Qwen3.5-9B	-0.025	-0.142	-0.128	-0.095	-0.008	-0.051	-0.048	–	-0.008	-0.152	-0.080	-0.048
	Gemma4-E2B	-0.034	-0.081	-0.120	-0.104	-0.021	-0.160	-0.199	-0.179	-0.064	-0.099	-0.134	-0.119
	Gemma4-E4B	-0.054	-0.129	-0.168	-0.134	-0.045	-0.120	-0.158	-0.141	-0.013	-0.070	-0.092	-0.062
OPD	Qwen3-8B→1.7B	-0.164	-0.207	-0.200	-0.194	-0.142	-0.178	-0.178	-0.168	-0.167	-0.169	-0.167	-0.166
	Qwen3.5-9B→2B	-0.061	-0.092	-0.079	-0.068	-0.048	-0.092	-0.083	-0.074	-0.093	-0.144	-0.139	-0.124
	Gemma4-E4B→E2B	-0.082	-0.136	-0.165	-0.140	-0.044	-0.077	-0.100	-0.088	-0.115	-0.191	-0.230	-0.189

Table 9: **Confidence in the model's own greedy response, by group.** Per-token $\log \pi$ that each checkpoint assigns to its own greedy response, averaged per query, then per group over the checkpoints at which the response was correct (C) or incorrect (W), then over the OOD datasets of each domain (Code: CruxEval, HumanEval+, MBPP+, LiveCodeBench; Math: MATH500, OlympiadBench; Science: JEEBench, SciBench, GPQA-Diamond). Groups as in Equation 4; shaded: the churning group. "–": no response of that correctness in the domain.

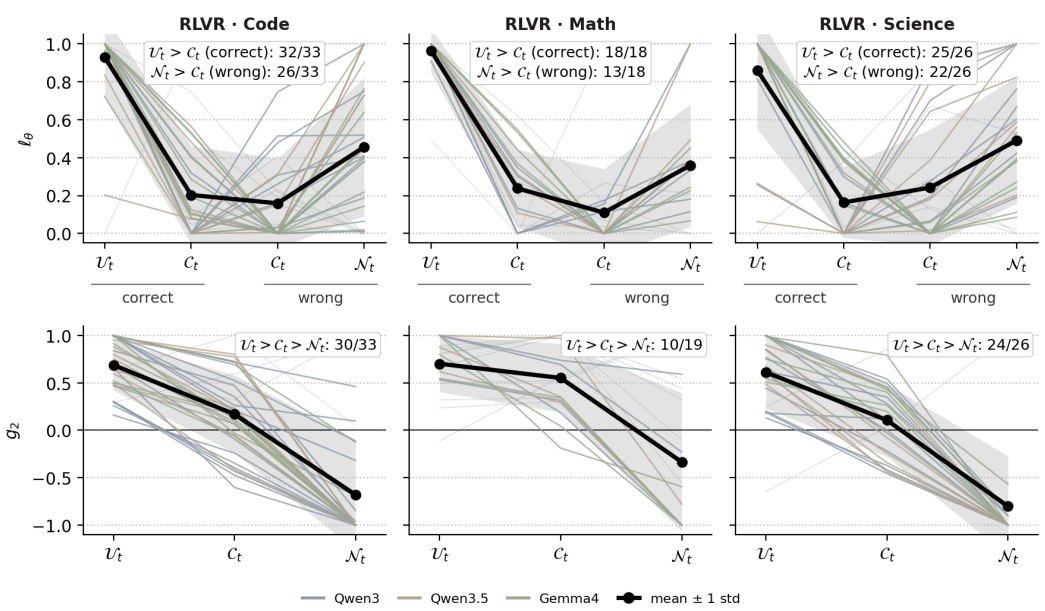

Figure 14: **Edge of understanding under RLVR, per domain.** As Figure 6, shown separately for code, math, and science; each thin line is one cell, scaled so that only the order of the groups remains.

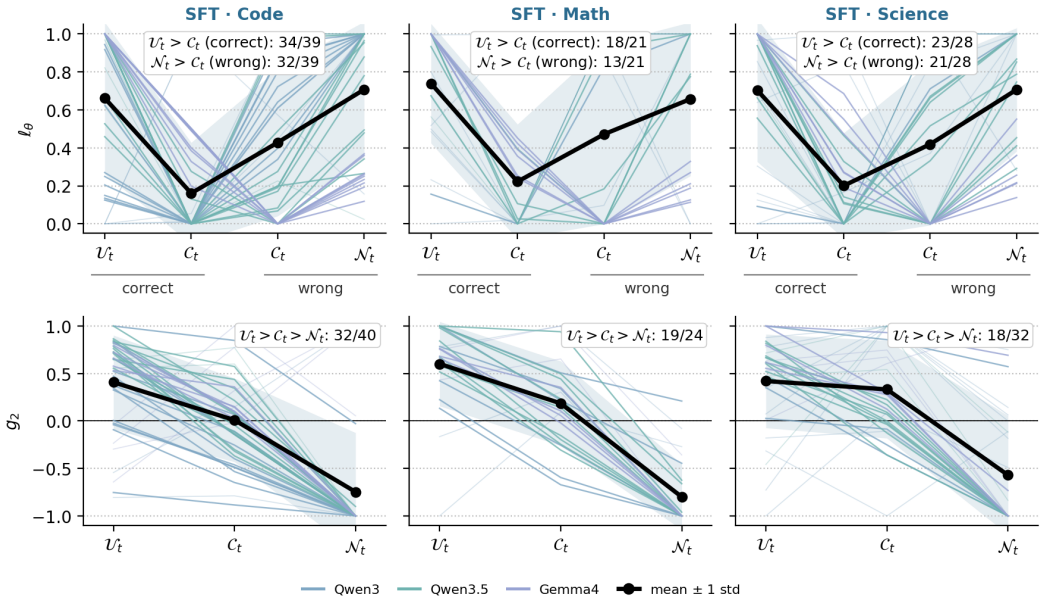

Figure 15: **Edge of understanding under SFT, per domain.** As Figure 14.

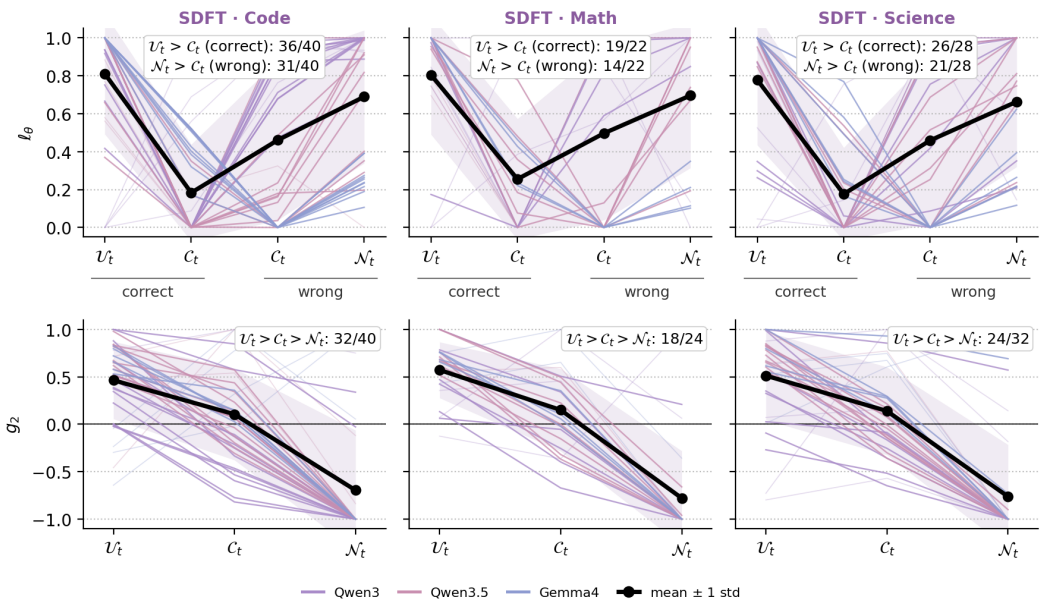

Figure 16: **Edge of understanding under SDFT, per domain.** As Figure 14.

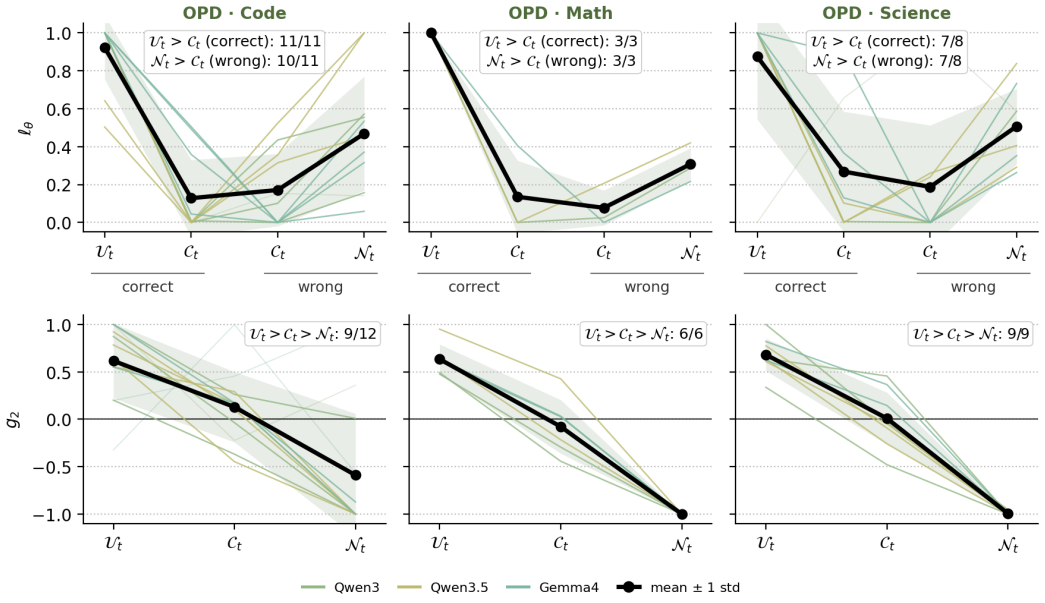

Figure 17: **Edge of understanding under OPD, per domain.** As Figure 14; one line per teacher→student pair and dataset.

Method	Model	Code			Math			Science		
		$\mathcal{U}_t$	$\mathcal{C}_t$	$\mathcal{N}_t$	$\mathcal{U}_t$	$\mathcal{C}_t$	$\mathcal{N}_t$	$\mathcal{U}_t$	$\mathcal{C}_t$	$\mathcal{N}_t$
RLVR	Qwen3-1.7B	+0.021	-0.015	-0.038	+0.058	+0.022	-0.083	+0.036	-0.048	-0.137
	Qwen3-4B	+0.031	-0.024	-0.094	+0.002	+0.047	+0.023	+0.035	+0.008	-0.050
	Qwen3-8B	+0.041	-0.003	-0.053	+0.034	+0.028	+0.032	+0.044	-0.034	-0.122
	Qwen3.5-2B	+0.117	+0.043	-0.170	+0.152	+0.066	-0.208	+0.246	-0.052	-0.335
	Qwen3.5-4B	+0.176	-0.007	-0.267	+0.121	+0.132	-0.090	+0.196	-0.106	-0.327
	Gemma4-E2B	+0.064	+0.004	-0.086	+0.044	+0.037	-0.106	+0.112	+0.023	-0.157
	Gemma4-E4B	+0.080	+0.016	-0.113	+0.091	-0.005	-0.086	-0.000	+0.017	+0.003
SFT	Qwen3-1.7B	+0.039	-0.009	-0.068	+0.048	-0.010	-0.095	+0.004	+0.008	-0.011
	Qwen3-4B	-1.962	-3.413	-5.230	-0.404	+0.124	-0.397	-0.568	-0.104	-0.355
	Qwen3-8B	+0.019	-0.039	-0.093	+0.042	-0.024	-0.058	+0.015	+0.016	-0.021
	Qwen3.5-2B	+0.179	+0.026	-0.240	+0.166	-0.027	-0.214	+0.132	+0.028	-0.206
	Qwen3.5-4B	+0.277	-0.048	-0.355	+0.263	+0.095	-0.259	+0.155	+0.032	-0.230
	Qwen3.5-9B	+0.177	+0.051	-0.231	+0.151	-0.050	-0.256	+0.098	+0.002	-0.195
	Gemma4-E2B	+0.095	+0.024	-0.139	+0.073	+0.014	-0.098	+0.038	+0.031	-0.049
	Gemma4-E4B	+0.016	+0.017	-0.021	+0.075	+0.032	-0.126	+0.038	+0.025	-0.042
SDFT	Qwen3-1.7B	+0.039	-0.009	-0.068	+0.048	-0.010	-0.095	+0.004	+0.008	-0.011
	Qwen3-4B	+1.367	-5.445	-8.659	+0.007	-0.054	-0.217	-0.071	-0.331	-0.825
	Qwen3-8B	+0.026	-0.018	-0.049	+0.029	+0.008	-0.055	+0.011	-0.019	-0.050
	Qwen3.5-2B	+0.213	-0.019	-0.310	+0.206	-0.066	-0.294	+0.157	+0.063	-0.222
	Qwen3.5-4B	+0.144	+0.077	-0.179	+0.239	-0.034	-0.330	+0.278	-0.049	-0.353
	Qwen3.5-9B	+0.163	+0.007	-0.242	+0.134	+0.049	-0.225	+0.211	-0.037	-0.268
	Gemma4-E2B	+0.095	+0.024	-0.139	+0.073	+0.014	-0.098	+0.038	+0.031	-0.049
	Gemma4-E4B	+0.016	+0.017	-0.021	+0.075	+0.032	-0.126	+0.038	+0.025	-0.042
OPD	Qwen3-8B→1.7B	+0.037	-0.016	-0.055	+0.051	-0.038	-0.107	+0.044	-0.014	-0.077
	Qwen3.5-9B→2B	+0.214	-0.028	-0.305	+0.184	-0.030	-0.266	+0.232	-0.053	-0.340
	Gemma4-E4B→E2B	+0.024	+0.022	-0.020	+0.065	+0.003	-0.104	+0.101	+0.019	-0.141

Table 10: **Confidence margin g_2 by group.** Group median of the per-query margin g_2 (Equation 6) between the most confident correct and incorrect responses, averaged over the OOD datasets of each domain (same grouping as Table 9). Groups as in Equation 4; shaded: the churning group, whose margin sits between the confidently-right unforgettable and the confidently-wrong unsolvable groups.

D CHURN DYNAMICS UNDER FACTUAL KNOWLEDGE LEARNING

The mixture of Section 3.2 is one on which accuracy has saturated. To see churn together with learning we fine-tune Qwen3.5-0.8B, 4B, and 9B and Gemma4-E2B for 12 epochs on four closed-book QA datasets (EntityQuestions, NQ, PopQA, WebQuestions; exact-match accuracy), where accuracy rises for the first 4–6 epochs and then plateaus. Figure 18 shows the per-epoch positive and negative churn rates of every run (mean over models in Figure 4). While accuracy rises, c^+ exceeds c^- in 78 of the 79 pre-plateau epochs; after the plateau the two rates coincide (within 0.4 points on average per run), but neither falls: the model keeps learning and forgetting queries at 12–14% per epoch with no net gain.

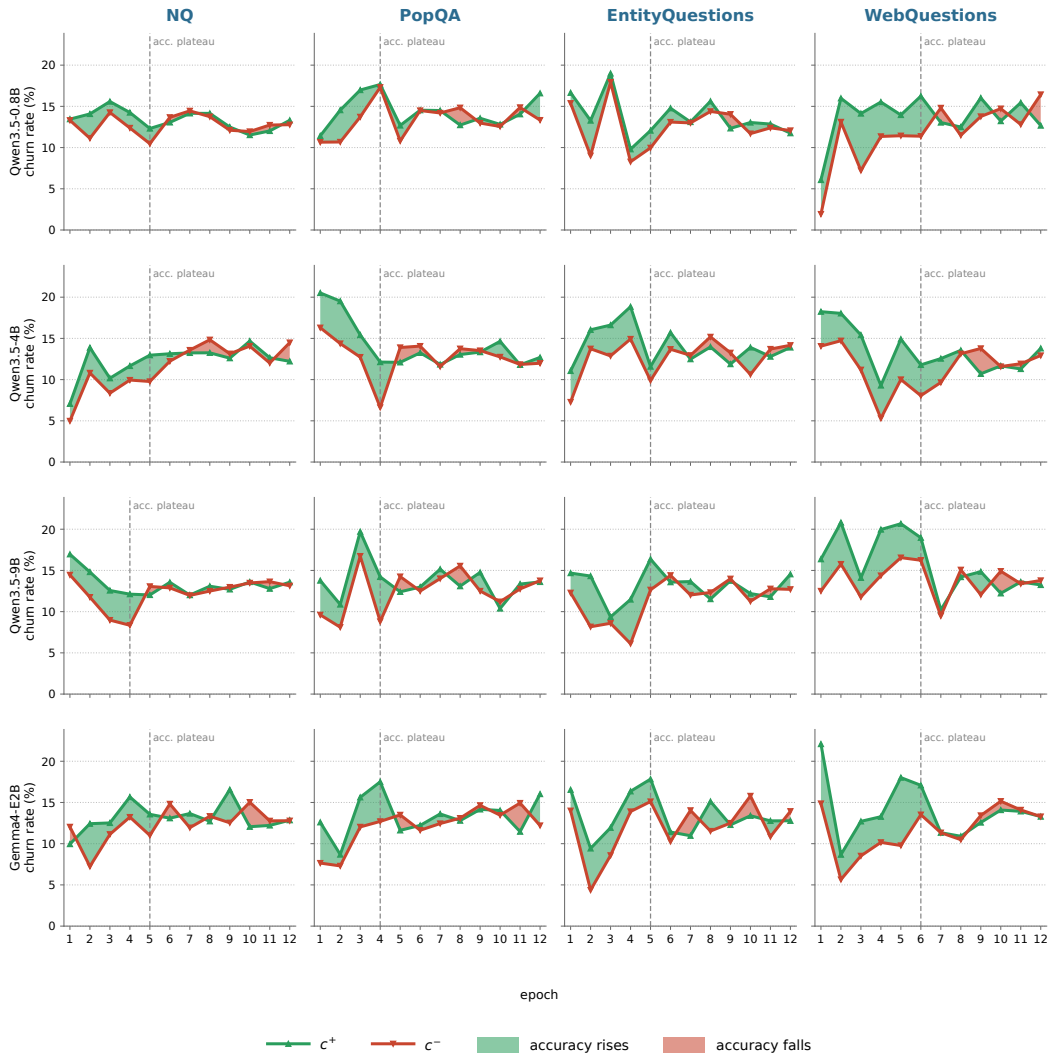

Figure 18: **Positive and negative churn over SFT epochs on closed-book QA.** Rows: Qwen3.5-0.8B/4B/9B and Gemma4-E2B; columns: NQ, PopQA, EntityQuestions, WebQuestions. c^+ (W→C) and c^- (C→W) per epoch; the gap is shaded green where $c^+ > c^-$ (accuracy rises) and red where $c^+ < c^-$ (accuracy falls). The dashed line marks the epoch at which accuracy plateaus. Accuracy rises while $c^+ > c^-$; afterwards the two rates coincide but stay at 12–14%, so the model keeps learning and forgetting queries at the same rate with no net gain.

E Does reducing churn benefit accuracy? SFT ablation

To test whether standard remedies reduce churn, we fine-tune Qwen3.5-0.8B, 4B, and 9B and Gemma4-E2B for 12 epochs on four closed-book QA datasets (EntityQuestions, NQ, PopQA, WebQuestions; exact-match accuracy), with three training seeds per configuration, and vary one ingredient of SFT at a time: sharpness-aware minimization (SAM [10]), which flattens the loss basin and reduces forgetting in sequential fine-tuning [28], enabled either from the first epoch or from epoch 6; weight decay, an L2 pull that EWC [21] generalizes; replay of earlier data [7, 35]; SDFT [39]; PCGrad [47], which projects out conflicting gradient components as GEM [26] does for past tasks; a learning rate of 5×10^{-6} or 5×10^{-5} and a batch size of 128 or 256, training-regime choices known to affect forgetting [29]; and merging the checkpoints of the run [44, 45]. Tables 11 and 12 report the churn rate and the final accuracy per dataset; Tables 13–16 give both per dataset together with the change against plain SFT.

Only SAM changes churn, and only for the three models of at most 4B parameters (Qwen3.5-0.8B, Qwen3.5-4B, Gemma4-E2B): there it lowers the churn rate by 29–32% (13.1% $\rightarrow$ 9.1% on average), whether it is enabled from epoch 0 or from epoch 6; on Qwen3.5-9B it lowers churn by only 6–8% (13.6% $\rightarrow$ 12.6–12.9%). Whether lower churn also means higher accuracy depends on when SAM is enabled. Enabled from epoch 0, SAM lowers final accuracy in all 16 cells (-1.5 points on average). Enabled from epoch 6, once accuracy has plateaued, it raises final accuracy by 5.0–5.7 points for the three smaller models (all 12 cells), but not for Qwen3.5-9B (-0.2 points). Every other intervention leaves the churn rate within 1% (relative) of plain SFT on average and final accuracy within 0.6 points, including both learning rates and both batch sizes. Merging checkpoints raises final accuracy by 0.2 points.

F Sensitivity to perturbation

Table 17 gives the numbers plotted in Figure 7.

G Forgetting gap details

G.1 Per-model tables

Tables 18–21 give, per model and dataset, the accuracy of the final checkpoint, the accuracy of the best single checkpoint, the fraction of learnable queries, and the forgetting gap.

Model	Method	NQ	PopQA	EntityQuestions	WebQuestions
Qwen3.5-0.8B	SFT	12.7 ±0.6	14.0 ±0.5	12.7 ±0.7	13.0 ±0.5
	+ SAM	9.0 ±0.7	11.4 ±0.9	9.0 ±0.8	**7.4** ±0.4
	+ SAM from ep. 6	**8.7** ±0.9	**9.4** ±0.4	**8.5** ±0.5	**9.0** ±0.8
	+ weight decay	12.3 ±0.5	14.0 ±0.5	12.4 ±0.7	12.7 ±0.5
	+ replay	12.9 ±0.7	14.6 ±0.4	12.8 ±0.3	13.0 ±0.8
	+ SDFT	12.5 ±0.5	14.0 ±0.6	12.6 ±0.3	13.2 ±0.7
	+ PCGrad	12.7 ±0.5	13.5 ±0.7	12.8 ±0.6	12.9 ±0.4
	+ lr 5×10^{-6}	12.9 ±0.5	14.4 ±0.5	12.5 ±0.5	12.6 ±0.6
	+ lr 5×10^{-5}	12.8 ±0.8	14.3 ±0.7	12.7 ±0.6	12.7 ±0.9
	+ batch size 128	12.7 ±0.7	13.9 ±0.5	12.7 ±0.9	13.2 ±0.3
	+ batch size 256	12.9 ±1.1	13.9 ±0.7	12.2 ±0.2	12.8 ±0.5
Qwen3.5-4B	SFT	13.1 ±0.3	12.8 ±0.6	14.4 ±0.7	13.0 ±0.7
	+ SAM	**8.1** ±0.4	**9.9** ±0.5	**9.7** ±0.8	**9.5** ±0.7
	+ SAM from ep. 6	8.8 ±0.8	**9.7** ±0.5	10.5 ±0.7	**8.8** ±0.7
	+ weight decay	13.0 ±0.5	12.7 ±0.7	14.2 ±0.4	13.1 ±0.9
	+ replay	13.0 ±0.6	12.7 ±0.3	14.8 ±0.6	12.7 ±0.7
	+ SDFT	13.2 ±0.5	12.4 ±0.5	14.7 ±0.9	12.6 ±0.3
	+ PCGrad	13.0 ±0.4	12.8 ±0.9	14.0 ±0.5	13.2 ±0.3
	+ lr 5×10^{-6}	13.0 ±0.5	12.7 ±1.0	14.7 ±0.6	13.1 ±0.5
	+ lr 5×10^{-5}	13.1 ±0.6	12.8 ±0.7	14.1 ±0.6	12.9 ±0.5
	+ batch size 128	13.1 ±0.6	12.9 ±0.6	14.4 ±0.6	12.9 ±0.5
	+ batch size 256	13.1 ±0.6	13.0 ±0.5	14.4 ±0.3	12.7 ±0.8
Qwen3.5-9B	SFT	15.8 ±0.7	14.1 ±0.7	12.3 ±0.4	12.4 ±0.8
	+ SAM	**14.5** ±0.8	**13.0** ±0.8	**11.4** ±0.4	**11.4** ±0.6
	+ SAM from ep. 6	15.0 ±0.7	13.3 ±0.7	11.6 ±0.6	11.6 ±0.8
	+ weight decay	15.9 ±1.0	13.8 ±0.6	12.5 ±0.8	12.2 ±0.5
	+ replay	15.5 ±0.9	13.7 ±0.8	12.2 ±0.6	12.4 ±0.9
	+ SDFT	16.0 ±0.7	14.0 ±0.6	12.2 ±0.1	12.0 ±0.7
	+ PCGrad	15.5 ±0.4	14.2 ±0.4	12.8 ±1.0	12.9 ±0.4
	+ lr 5×10^{-6}	16.0 ±0.8	14.2 ±0.7	12.0 ±0.9	12.4 ±0.7
	+ lr 5×10^{-5}	15.5 ±0.8	13.8 ±0.4	12.0 ±0.8	12.1 ±0.9
	+ batch size 128	15.7 ±0.6	13.9 ±0.6	12.3 ±0.7	12.3 ±0.7
	+ batch size 256	15.6 ±0.6	13.8 ±0.9	12.2 ±1.0	12.3 ±0.6
Gemma4-E2B	SFT	12.7 ±0.5	12.7 ±0.7	12.7 ±0.6	12.8 ±0.6
	+ SAM	9.4 ±1.1	9.0 ±0.5	9.7 ±0.8	**8.1** ±0.8
	+ SAM from ep. 6	**9.4** ±0.3	**8.5** ±0.6	**7.9** ±0.4	**9.3** ±0.7
	+ weight decay	12.7 ±0.4	12.6 ±0.7	12.8 ±0.5	12.6 ±0.5
	+ replay	12.7 ±0.8	13.0 ±0.5	12.6 ±0.5	13.0 ±0.8
	+ SDFT	12.9 ±0.6	12.7 ±0.7	12.8 ±0.4	12.6 ±0.7
	+ PCGrad	12.7 ±0.9	12.8 ±0.6	12.8 ±0.7	12.7 ±0.3
	+ lr 5×10^{-6}	12.7 ±0.7	12.9 ±0.7	12.5 ±0.8	12.8 ±0.9
	+ lr 5×10^{-5}	13.0 ±0.4	12.6 ±0.5	12.7 ±0.7	13.1 ±0.5
	+ batch size 128	12.8 ±1.0	12.8 ±0.8	12.6 ±0.4	12.8 ±0.4
	+ batch size 256	12.8 ±0.4	12.9 ±0.7	12.7 ±0.5	12.6 ±0.5

Table 11: **SFT ablation: churn rate c (%) per dataset.** Qwen3.5-0.8B/4B/9B and Gemma4-E2B after 12 epochs of SFT on closed-book QA. Grey ±: standard deviation over three training seeds. Bold: lowest churn in each column within a model; shaded: the method with the lowest churn averaged over the four datasets. Checkpoint merging is evaluated only at the end and has no churn rate, so it is omitted.

Model	Method	NQ	PopQA	EntityQuestions	WebQuestions
Qwen3.5-0.8B	SFT	14.1 ±0.5	20.9 ±1.2	19.1 ±1.1	28.8 ±0.9
	+ SAM	12.9 ±1.1	20.3 ±1.4	18.2 ±0.8	27.6 ±1.1
	+ SAM from ep. 6	**18.6** ±1.0	**26.2** ±1.0	**24.6** ±1.2	**34.5** ±1.5
	+ weight decay	14.0 ±1.1	20.2 ±0.9	18.8 ±0.7	27.5 ±0.9
	+ replay	14.2 ±0.9	20.5 ±0.7	18.7 ±0.4	29.3 ±0.4
	+ SDFT	14.1 ±1.2	21.3 ±0.2	18.8 ±1.0	28.7 ±1.0
	+ PCGrad	13.8 ±0.8	21.1 ±1.3	19.2 ±1.3	28.1 ±1.0
	+ lr 5×10^{-6}	14.0 ±0.7	21.2 ±0.9	19.0 ±1.1	29.7 ±0.8
	+ lr 5×10^{-5}	13.9 ±0.6	21.0 ±1.7	18.7 ±0.8	29.1 ±1.3
	+ batch size 128	14.2 ±1.4	21.3 ±1.0	19.1 ±1.4	28.4 ±0.6
	+ batch size 256	13.8 ±1.1	21.0 ±1.1	18.9 ±1.3	29.0 ±0.5
	+ checkpoint merging	15.0 ±1.1	21.4 ±0.7	18.7 ±1.0	28.6 ±0.8
Qwen3.5-4B	SFT	15.5 ±1.2	30.3 ±1.0	30.0 ±0.8	40.1 ±1.0
	+ SAM	15.4 ±0.8	28.8 ±1.6	28.4 ±0.8	37.9 ±0.6
	+ SAM from ep. 6	**20.3** ±0.8	**35.8** ±1.0	**35.7** ±0.9	**47.1** ±0.5
	+ weight decay	14.9 ±1.2	28.8 ±0.7	30.0 ±1.0	39.7 ±0.7
	+ replay	15.2 ±0.8	30.3 ±0.2	30.0 ±0.8	39.9 ±1.0
	+ SDFT	15.4 ±1.0	30.4 ±1.5	30.2 ±0.9	40.0 ±1.3
	+ PCGrad	15.7 ±1.1	30.1 ±1.1	29.6 ±1.2	40.5 ±0.8
	+ lr 5×10^{-6}	15.8 ±0.6	30.3 ±0.8	29.9 ±0.9	40.3 ±0.3
	+ lr 5×10^{-5}	15.6 ±1.3	30.1 ±0.4	29.7 ±1.5	40.6 ±1.1
	+ batch size 128	15.6 ±0.6	30.3 ±0.7	29.8 ±0.8	40.1 ±0.8
	+ batch size 256	15.7 ±1.1	30.1 ±1.1	29.6 ±0.5	39.7 ±1.1
	+ checkpoint merging	16.6 ±1.0	31.8 ±0.8	30.8 ±1.1	38.1 ±1.6
Qwen3.5-9B	SFT	20.2 ±1.4	36.1 ±1.0	39.8 ±0.4	46.9 ±0.9
	+ SAM	19.3 ±0.8	33.2 ±0.8	37.2 ±1.3	44.8 ±1.2
	+ SAM from ep. 6	20.1 ±0.7	36.3 ±0.3	39.6 ±1.1	46.5 ±1.0
	+ weight decay	19.8 ±1.0	35.8 ±1.2	39.3 ±1.0	46.2 ±0.6
	+ replay	19.8 ±1.1	36.1 ±0.9	**40.7** ±1.0	47.4 ±1.2
	+ SDFT	19.4 ±1.1	36.4 ±1.3	39.0 ±1.2	46.6 ±0.9
	+ PCGrad	19.8 ±1.3	36.7 ±0.7	39.1 ±0.4	46.7 ±1.2
	+ lr 5×10^{-6}	20.0 ±1.0	36.2 ±1.3	39.2 ±1.1	45.8 ±1.2
	+ lr 5×10^{-5}	20.2 ±1.3	35.7 ±0.9	40.0 ±0.7	47.1 ±1.0
	+ batch size 128	20.3 ±1.6	36.0 ±0.8	39.5 ±1.6	46.9 ±0.7
	+ batch size 256	**20.3** ±0.6	**37.1** ±1.4	**40.5** ±0.6	**46.3** ±0.6
	+ checkpoint merging	20.2 ±1.0	36.2 ±0.8	38.1 ±0.5	**49.7** ±1.3
Gemma4-E2B	SFT	23.2 ±1.5	26.4 ±0.7	25.9 ±1.1	37.9 ±0.6
	+ SAM	21.6 ±1.3	24.7 ±1.1	25.1 ±0.8	35.4 ±1.4
	+ SAM from ep. 6	**27.3** ±0.8	**31.7** ±0.8	**30.1** ±0.6	**44.3** ±0.9
	+ weight decay	22.6 ±1.0	25.7 ±1.5	25.0 ±0.6	37.8 ±1.9
	+ replay	23.1 ±0.9	25.8 ±0.6	25.5 ±0.7	38.1 ±1.3
	+ SDFT	23.2 ±0.9	26.6 ±0.3	25.8 ±0.7	38.2 ±1.2
	+ PCGrad	22.8 ±1.2	26.7 ±1.2	25.8 ±1.2	38.8 ±0.6
	+ lr 5×10^{-6}	22.8 ±0.4	26.7 ±1.7	26.3 ±1.4	37.9 ±0.6
	+ lr 5×10^{-5}	22.9 ±1.2	26.4 ±1.1	25.8 ±0.9	37.5 ±1.2
	+ batch size 128	23.2 ±0.9	26.9 ±1.0	26.2 ±1.1	38.8 ±1.1
	+ batch size 256	23.2 ±1.1	26.4 ±1.1	25.8 ±1.3	38.2 ±1.2
	+ checkpoint merging	23.3 ±0.6	27.1 ±0.8	26.2 ±1.4	37.7 ±1.5

Table 12: **SFT ablation: final accuracy (%) per dataset.** Exact-match accuracy of Qwen3.5-0.8B/4B/9B and Gemma4-E2B after 12 epochs of SFT on closed-book QA. Grey ±: standard deviation over three training seeds. Bold: highest accuracy in each column within a model; shaded: the method with the highest accuracy averaged over the four datasets.

Method	Churn c (%)				Final accuracy (%)				vs. SFT	
	0.8B	4B	9B	G-E2B	0.8B	4B	9B	G-E2B	Δc (%)	Δacc (pts)
SFT	12.7 ±0.6	13.1 ±0.3	15.8 ±0.7	12.7 ±0.5	14.1 ±0.5	15.5 ±1.2	20.2 ±1.4	23.2 ±1.5	–	–
+ SAM	**9.0** ±0.7	**8.1** ±0.4	**14.5** ±0.8	**9.4** ±1.1	**12.9** ±1.1	**15.4** ±0.8	**19.3** ±0.8	**21.6** ±1.3	**-24.3** ±5.8	**-0.9** ±1.6
+ SAM from ep. 6	**8.7** ±0.9	**8.8** ±0.8	**15.0** ±0.7	**9.4** ±0.3	**18.6** ±1.0	**20.3** ±0.8	**20.1** ±0.7	**27.3** ±0.8	**-22.7** ±5.0	**+3.3** ±1.5
+ weight decay	12.3 ±0.5	13.0 ±0.5	15.9 ±1.0	12.7 ±0.4	14.0 ±1.1	14.9 ±1.2	19.8 ±1.0	22.6 ±1.0	-0.6 ±4.7	-0.4 ±1.6
+ replay	12.9 ±0.7	13.0 ±0.6	15.5 ±0.9	12.7 ±0.8	14.2 ±0.9	15.2 ±0.8	19.8 ±1.1	23.1 ±0.9	-0.3 ±5.6	-0.2 ±1.5
+ SDFT	12.5 ±0.5	13.2 ±0.5	16.0 ±0.7	12.9 ±0.6	14.1 ±1.2	15.4 ±1.0	19.4 ±1.1	23.2 ±0.9	+0.6 ±4.3	-0.2 ±1.6
+ PCGrad	12.7 ±0.5	13.0 ±0.4	15.5 ±0.4	12.7 ±0.9	13.8 ±0.8	15.7 ±1.1	19.8 ±1.3	22.8 ±1.2	-0.7 ±4.2	-0.2 ±1.6
+ lr 5×10^{-6}	12.9 ±0.5	13.0 ±0.5	16.0 ±0.8	12.7 ±0.7	14.0 ±0.7	15.8 ±0.6	20.0 ±1.0	22.8 ±0.4	+0.7 ±4.4	-0.1 ±1.4
+ lr 5×10^{-5}	12.8 ±0.8	13.1 ±0.6	15.5 ±0.8	13.0 ±0.4	13.9 ±0.6	15.6 ±1.3	20.2 ±1.3	22.9 ±1.2	+0.4 ±5.0	-0.1 ±1.7
+ batch size 128	12.7 ±0.7	13.1 ±0.6	15.7 ±0.6	12.8 ±1.0	14.2 ±1.4	15.6 ±0.6	20.3 ±1.6	23.2 ±0.9	-0.1 ±5.4	+0.1 ±1.7
+ batch size 256	12.9 ±1.1	13.1 ±0.6	15.6 ±0.6	12.8 ±0.4	13.8 ±1.1	15.7 ±1.1	**20.3** ±0.6	23.2 ±1.1	+0.3 ±5.3	0.0 ±1.5
+ checkpoint merging	–	–	–	–	15.0 ±1.1	16.6 ±1.0	20.2 ±1.0	23.3 ±0.6	–	+0.5 ±1.5

Table 13: **SFT ablation on NQ.** Churn rate and final accuracy of Qwen3.5-0.8B/4B/9B and Gemma4-E2B (G-E2B) after 12 epochs of SFT. Δc: relative change in churn against plain SFT; Δacc: change in final accuracy in points; both averaged over the four models. Grey $\pm$: standard deviation over three training seeds; in the Δ columns it is pooled over models as the square root of the mean variance. Bold: best value in each column; shaded: the method with the lowest churn and the method with the highest accuracy, averaged over models. Checkpoint merging is evaluated only at the end, so it has no churn rate.

Method	Churn c (%)				Final accuracy (%)				vs. SFT	
	0.8B	4B	9B	G-E2B	0.8B	4B	9B	G-E2B	Δc (%)	Δacc (pts)
SFT	14.0 ±0.5	12.8 ±0.6	14.1 ±0.7	12.7 ±0.7	20.9 ±1.2	30.3 ±1.0	36.1 ±1.0	26.4 ±0.7	–	–
+ SAM	11.4 ±0.9	9.9 ±0.5	**13.0** ±0.8	9.0 ±0.5	20.3 ±1.4	28.8 ±1.6	33.2 ±0.8	24.7 ±1.1	-19.2 ±5.1	-1.7 ±1.6
+ SAM from ep. 6	**9.4** ±0.4	**9.7** ±0.5	**13.3** ±0.7	**8.5** ±0.6	**26.2** ±1.0	**35.8** ±1.0	**36.3** ±0.3	**31.7** ±0.8	**-23.6** ±4.2	**+4.0** ±1.3
+ weight decay	14.0 ±0.5	12.7 ±0.7	13.8 ±0.6	12.6 ±0.7	20.2 ±0.9	28.8 ±0.7	35.8 ±1.2	25.7 ±1.5	-0.9 ±4.7	-0.8 ±1.5
+ replay	14.6 ±0.4	12.7 ±0.3	13.7 ±0.8	13.0 ±0.5	20.5 ±0.7	30.3 ±0.2	36.1 ±0.9	25.8 ±0.6	+0.7 ±4.1	-0.2 ±1.2
+ SDFT	14.0 ±0.6	12.4 ±0.5	14.0 ±0.6	12.7 ±0.7	21.3 ±0.2	30.4 ±1.5	36.4 ±1.3	26.6 ±0.3	-1.0 ±4.5	+0.2 ±1.4
+ PCGrad	13.5 ±0.7	12.8 ±0.9	14.2 ±0.4	12.8 ±0.6	21.1 ±1.3	30.1 ±1.1	36.7 ±0.7	26.7 ±1.2	-0.9 ±5.0	+0.2 ±1.5
+ lr 5×10^{-6}	14.4 ±0.5	12.7 ±1.0	14.2 ±0.7	12.9 ±0.7	21.2 ±0.9	30.3 ±0.8	36.2 ±1.3	26.7 ±1.7	+1.1 ±5.5	+0.2 ±1.6
+ lr 5×10^{-5}	14.3 ±0.7	12.8 ±0.7	13.8 ±0.4	12.6 ±0.5	21.0 ±1.7	30.1 ±0.4	35.7 ±0.9	26.4 ±1.1	-0.2 ±4.4	-0.2 ±1.5
+ batch size 128	13.9 ±0.5	12.9 ±0.6	13.9 ±0.6	12.8 ±0.8	21.3 ±1.0	30.3 ±0.7	36.0 ±0.8	26.9 ±1.0	-0.1 ±4.7	+0.2 ±1.3
+ batch size 256	13.9 ±0.7	13.0 ±0.5	13.8 ±0.9	12.9 ±0.7	21.0 ±1.1	30.1 ±1.1	**37.1** ±1.4	26.4 ±1.1	+0.1 ±5.3	+0.2 ±1.6
+ checkpoint merging	–	–	–	–	21.4 ±0.7	31.8 ±0.8	36.2 ±0.8	27.1 ±0.8	–	+0.7 ±1.3

Table 14: **SFT ablation on PopQA.** Churn rate and final accuracy of Qwen3.5-0.8B/4B/9B and Gemma4-E2B (G-E2B) after 12 epochs of SFT. Δc: relative change in churn against plain SFT; Δacc: change in final accuracy in points; both averaged over the four models. Grey $\pm$: standard deviation over three training seeds; in the Δ columns it is pooled over models as the square root of the mean variance. Bold: best value in each column; shaded: the method with the lowest churn and the method with the highest accuracy, averaged over models. Checkpoint merging is evaluated only at the end, so it has no churn rate.

Method	Churn c (%)				Final accuracy (%)				vs. SFT	
	0.8B	4B	9B	G-E2B	0.8B	4B	9B	G-E2B	Δc (%)	Δacc (pts)
SFT	12.7 ±0.7	14.4 ±0.7	12.3 ±0.4	12.7 ±0.6	19.1 ±1.1	30.0 ±0.8	39.8 ±0.4	25.9 ±1.1	–	–
+ SAM	9.0 ±0.8	**9.7** ±0.8	**11.4** ±0.4	9.7 ±0.8	18.2 ±0.8	28.4 ±0.8	37.2 ±1.3	25.1 ±0.8	-23.6 ±5.4	-1.5 ±1.3
+ SAM from ep. 6	**8.5** ±0.5	**10.5** ±0.7	**11.6** ±0.6	**7.9** ±0.4	**24.6** ±1.2	**35.7** ±0.9	**39.6** ±1.1	**30.1** ±0.6	**-26.1** ±4.5	**+3.8** ±1.3
+ weight decay	12.4 ±0.7	14.2 ±0.4	12.5 ±0.8	12.8 ±0.5	18.8 ±0.7	30.0 ±1.0	39.3 ±1.0	25.0 ±0.6	-0.4 ±5.0	-0.4 ±1.2
+ replay	12.8 ±0.3	14.8 ±0.6	12.2 ±0.6	12.6 ±0.5	18.7 ±0.4	30.0 ±0.8	**40.7** ±1.0	25.5 ±0.7	+0.6 ±3.9	0.0 ±1.2
+ SDFT	12.6 ±0.3	14.7 ±0.9	12.2 ±0.1	12.8 ±0.4	18.8 ±1.0	30.2 ±0.9	39.0 ±1.2	25.8 ±0.7	+0.2 ±3.9	-0.2 ±1.3
+ PCGrad	12.8 ±0.6	14.0 ±0.5	12.8 ±1.0	12.8 ±0.7	19.2 ±1.3	29.6 ±1.2	39.1 ±0.4	25.8 ±1.2	+0.5 ±5.6	-0.2 ±1.4
+ lr 5×10^{-6}	12.5 ±0.5	14.7 ±0.6	12.0 ±0.9	12.5 ±0.8	19.0 ±1.1	29.9 ±0.9	39.2 ±1.1	26.3 ±1.4	-0.6 ±5.5	-0.1 ±1.4
+ lr 5×10^{-5}	12.7 ±0.6	14.1 ±0.6	12.0 ±0.8	12.7 ±0.7	18.7 ±0.8	29.7 ±1.5	40.0 ±0.7	25.8 ±0.9	-1.2 ±5.2	-0.2 ±1.3
+ batch size 128	12.7 ±0.9	14.4 ±0.6	12.3 ±0.7	12.6 ±0.4	19.1 ±1.4	29.8 ±0.8	39.5 ±1.6	26.2 ±1.1	-0.1 ±5.3	-0.1 ±1.5
+ batch size 256	12.2 ±0.2	14.4 ±0.3	12.2 ±1.0	12.7 ±0.5	18.9 ±1.3	29.6 ±0.5	40.5 ±0.6	25.8 ±1.3	-1.2 ±4.6	0.0 ±1.3
+ checkpoint merging	–	–	–	–	18.7 ±1.0	30.8 ±1.1	38.1 ±0.5	26.2 ±1.4	–	-0.3 ±1.4

Table 15: **SFT ablation on EntityQuestions.** Churn rate and final accuracy of Qwen3.5-0.8B/4B/9B and Gemma4-E2B (G-E2B) after 12 epochs of SFT. Δc: relative change in churn against plain SFT; Δacc: change in final accuracy in points; both averaged over the four models. Grey $\pm$: standard deviation over three training seeds; in the Δ columns it is pooled over models as the square root of the mean variance. Bold: best value in each column; shaded: the method with the lowest churn and the method with the highest accuracy, averaged over models. Checkpoint merging is evaluated only at the end, so it has no churn rate.

Method	Churn c (%)				Final accuracy (%)				vs. SFT	
	0.8B	4B	9B	G-E2B	0.8B	4B	9B	G-E2B	Δc (%)	Δacc (pts)
SFT	13.0 ±0.5	13.0 ±0.7	12.4 ±0.8	12.8 ±0.6	28.8 ±0.9	40.1 ±1.0	46.9 ±0.9	37.9 ±0.6	–	–
+ SAM	**7.4** ±0.4	**9.5** ±0.7	**11.4** ±0.6	**8.1** ±0.8	**27.6** ±1.1	**37.9** ±0.6	**44.8** ±1.2	**35.4** ±1.4	**-28.9** ±5.0	**-2.0** ±1.4
+ SAM from ep. 6	**9.0** ±0.8	**8.8** ±0.7	**11.6** ±0.8	**9.3** ±0.7	**34.5** ±1.5	**47.1** ±0.5	**46.5** ±1.0	**44.3** ±0.9	**-24.5** ±6.0	**+4.7** ±1.4
+ weight decay	12.7 ±0.5	13.1 ±0.9	12.2 ±0.5	12.6 ±0.5	27.5 ±0.9	39.7 ±0.7	46.2 ±0.6	37.8 ±1.9	-1.2 ±4.9	-0.7 ±1.4
+ replay	13.0 ±0.8	12.7 ±0.7	12.4 ±0.9	13.0 ±0.8	29.3 ±0.4	39.9 ±1.0	47.4 ±1.2	38.1 ±1.3	-0.2 ±6.2	+0.2 ±1.3
+ SDFT	13.2 ±0.7	12.6 ±0.3	12.0 ±0.7	12.6 ±0.7	28.7 ±1.0	40.0 ±1.3	46.6 ±0.9	38.2 ±1.2	-1.4 ±4.9	-0.1 ±1.4
+ PCGrad	12.9 ±0.4	13.2 ±0.3	12.9 ±0.4	12.7 ±0.3	28.1 ±1.0	40.5 ±0.8	46.7 ±1.2	38.8 ±0.6	+1.0 ±3.0	+0.1 ±1.3
+ lr 5×10^{-6}	12.6 ±0.6	13.1 ±0.5	12.4 ±0.7	12.8 ±0.9	29.7 ±0.8	40.3 ±0.3	45.8 ±1.2	37.9 ±0.6	-0.7 ±5.1	0.0 ±1.2
+ lr 5×10^{-5}	12.7 ±0.9	12.9 ±0.5	12.1 ±0.9	13.1 ±0.5	29.1 ±1.3	40.6 ±1.1	47.1 ±1.0	37.5 ±1.2	-0.7 ±5.7	+0.1 ±1.4
+ batch size 128	13.2 ±0.3	12.9 ±0.5	12.3 ±0.7	12.8 ±0.4	28.4 ±0.6	40.1 ±0.8	46.9 ±0.7	38.8 ±1.1	-0.1 ±3.9	+0.1 ±1.2
+ batch size 256	12.8 ±0.5	12.7 ±0.8	12.3 ±0.6	12.6 ±0.5	29.0 ±0.5	39.7 ±1.1	46.3 ±0.6	38.2 ±1.2	-1.6 ±4.7	-0.2 ±1.3
+ checkpoint merging	–	–	–	–	28.6 ±0.8	38.1 ±1.6	**49.7** ±1.3	37.7 ±1.5	–	+0.1 ±1.6

Table 16: **SFT ablation on WebQuestions.** Churn rate and final accuracy of Qwen3.5-0.8B/4B/9B and Gemma4-E2B (G-E2B) after 12 epochs of SFT. Δc: relative change in churn against plain SFT; Δacc: change in final accuracy in points; both averaged over the four models. Grey $\pm$: standard deviation over three training seeds; in the Δ columns it is pooled over models as the square root of the mean variance. Bold: best value in each column; shaded: the method with the lowest churn and the method with the highest accuracy, averaged over models. Checkpoint merging is evaluated only at the end, so it has no churn rate.

Method	Model	Flip-rate ratio $\mathcal{C}_t/\mathcal{U}_t$ ρ=0.001	ρ=0.003	ρ=0.01	ρ=0.03
RLVR	Qwen3-1.7B	16.2 ±0.5	6.8 ±0.2	3.7 ±0.1	1.9 ±0.1
	Qwen3-4B	11.0 ±0.4	6.9 ±0.2	12.6 ±0.4	3.7 ±0.1
	Qwen3-8B	8.6 ±0.3	9.7 ±0.1	4.5 ±0.2	2.6 ±0.0
SFT	Qwen3-1.7B	15.4 ±0.3	7.1 ±0.2	3.2 ±0.1	1.8 ±0.0
	Qwen3-4B	11.3 ±0.2	7.2 ±0.1	10.6 ±0.4	3.3 ±0.0
	Qwen3-8B	8.9 ±0.3	8.2 ±0.2	4.4 ±0.2	2.8 ±0.1
SDFT	Qwen3-1.7B	17.9 ±0.5	6.5 ±0.2	3.3 ±0.1	1.7 ±0.1
	Qwen3-4B	7.8 ±0.1	7.3 ±0.2	3.6 ±0.1	2.9 ±0.1
	Qwen3-8B	14.7 ±0.6	12.0 ±0.2	5.2 ±0.2	3.3 ±0.1
OPD	Qwen3-1.7B	15.0 ±0.3	5.5 ±0.2	2.7 ±0.0	3.2 ±0.1
	Gemma4-E2B	12.2 ±0.5	7.0 ±0.4	3.1 ±0.1	2.7 ±0.1
	Qwen3.5-2B	11.2 ±0.1	13.5 ±0.1	4.4 ±0.2	3.3 ±0.1
Mean over runs		**12.5** ±3.3	**8.1** ±2.4	**5.1** ±3.2	**2.8** ±0.6

Table 17: **Sensitivity to perturbation (ID dataset).** Flip rate of churning over unforgettable queries when Gaussian noise of RMS ρ is added to every weight of the final checkpoint; mean ± 1 s.d. over seeds, last row over the 12 runs. Plotted in Figure 7.

Model		ID	MATH500	Olympiad	JEE	SciBench	GPQA-D	CruxEval	HE+	MBPP+	LCB	OOD mean
Qwen3-1.7B	Base	24.7	67.6	30.2	26.4	30.9	39.7	18.7	53.3	51.4	15.7	37.1
	Final	28.9 ±1.7	69.8 ±2.7	34.0 ±4.3	28.5 ±1.9	34.7 ±1.4	41.9 ±1.2	22.4 ±1.8	57.3 ±2.8	53.4 ±1.3	21.9 ±1.4	40.4
	Best	29.8 ±1.4	72.0 ±1.9	36.1 ±3.2	31.8 ±1.6	35.0 ±1.4	47.5 ±1.0	22.4 ±1.7	61.0 ±2.1	53.4 ±1.1	21.9 ±1.1	42.3
	Union	48.1 ±1.3	86.6 ±2.1	59.8 ±2.2	73.6 ±1.3	59.4 ±0.9	79.3 ±0.6	42.5 ±1.1	65.2 ±2.1	64.0 ±1.0	30.5 ±0.9	62.3
	Majority	28.2 ±1.5	70.2 ±1.6	32.0 ±3.2	24.1 ±1.4	31.9 ±1.2	39.9 ±1.1	19.2 ±1.1	59.1 ±2.1	52.1 ±0.8	17.8 ±1.0	38.5
	Gap	**19.3** ±2.2	**16.8** ±3.4	**25.8** ±4.9	**45.0** ±2.3	**24.8** ±1.7	**37.4** ±1.3	**20.1** ±2.1	**7.9** ±3.5	**10.6** ±1.7	**8.6** ±1.6	**21.9**
Qwen3-4B	Base	38.6	75.9	40.4	42.1	52.9	44.5	65.7	70.3	57.0	36.4	53.9
	Final	41.6 ±1.0	79.0 ±0.8	45.0 ±0.2	47.8 ±1.1	57.4 ±2.0	46.0 ±0.5	69.3 ±1.5	72.6 ±0.9	65.6 ±5.2	38.7 ±1.1	57.9
	Best	43.9 ±0.8	81.4 ±0.6	46.1 ±0.2	50.3 ±1.1	57.4 ±1.7	48.5 ±0.3	70.0 ±1.4	73.8 ±0.8	67.2 ±5.0	39.9 ±0.8	59.4
	Union	58.9 ±0.8	91.2 ±0.5	65.1 ±0.1	83.1 ±0.9	76.1 ±1.4	81.8 ±0.3	81.1 ±0.8	76.2 ±0.5	73.5 ±2.2	50.5 ±0.6	75.4
	Majority	41.8 ±0.7	82.6 ±0.7	44.2 ±0.2	47.6 ±0.9	55.5 ±1.5	40.9 ±0.3	69.1 ±1.3	72.0 ±0.7	65.3 ±3.9	37.2 ±0.8	57.2
	Gap	**17.3** ±1.3	**12.2** ±0.9	**20.2** ±0.3	**35.3** ±1.4	**18.7** ±2.4	**35.9** ±0.6	**11.8** ±1.7	**3.7** ±1.1	**7.9** ±5.6	**11.7** ±1.3	**17.5**
Qwen3-8B	Base	49.4	82.5	51.2	59.3	61.1	49.6	71.4	77.2	46.9	42.8	60.2
	Final	50.2 ±1.4	84.8 ±0.7	55.3 ±2.3	62.7 ±2.0	60.7 ±0.8	52.0 ±1.4	75.9 ±2.7	80.5 ±1.7	47.1 ±2.6	46.6 ±3.1	62.8
	Best	51.9 ±1.3	86.2 ±0.6	55.3 ±1.9	62.7 ±1.8	63.1 ±0.7	56.6 ±1.2	76.9 ±2.5	81.1 ±1.3	55.6 ±2.1	50.7 ±2.7	65.4
	Union	73.0 ±0.9	93.8 ±0.4	69.7 ±1.7	90.1 ±1.1	82.5 ±0.5	88.9 ±0.9	89.7 ±1.4	85.4 ±1.2	74.6 ±1.4	75.1 ±2.0	83.3
	Majority	48.7 ±1.0	85.4 ±0.6	52.2 ±1.8	58.1 ±1.3	61.7 ±0.7	53.5 ±1.1	76.6 ±2.2	79.3 ±1.4	47.9 ±1.8	46.8 ±1.9	62.4
	Gap	**22.8** ±1.7	**9.0** ±0.8	**14.4** ±2.9	**27.4** ±2.3	**21.8** ±1.0	**36.9** ±1.6	**13.8** ±3.0	**4.9** ±2.1	**27.5** ±3.0	**28.6** ±3.7	**20.5**
Qwen3.5-2B	Base	22.9	60.6	26.9	22.6	38.3	30.2	36.0	53.9	42.2	13.6	36.1
	Final	28.1 ±1.4	65.6 ±2.5	31.9 ±2.7	29.9 ±5.7	42.6 ±2.5	32.3 ±1.7	43.5 ±3.7	55.5 ±1.4	46.6 ±2.4	14.5 ±0.3	40.3
	Best	28.5 ±1.0	67.2 ±2.4	34.9 ±2.6	34.6 ±5.1	46.1 ±2.1	34.8 ±1.4	47.4 ±3.3	56.1 ±1.2	49.2 ±2.4	17.4 ±0.3	43.1
	Union	49.2 ±1.0	87.6 ±1.4	57.1 ±1.3	76.1 ±4.1	74.9 ±1.9	79.3 ±1.2	82.9 ±1.6	58.5 ±0.7	55.0 ±1.3	31.1 ±0.2	66.9
	Majority	28.0 ±1.2	67.2 ±2.1	32.2 ±2.2	29.3 ±4.7	43.0 ±2.0	27.8 ±1.4	46.2 ±3.3	54.9 ±0.9	47.9 ±1.9	14.3 ±0.2	40.3
	Gap	**21.1** ±1.7	**22.0** ±2.9	**25.2** ±3.0	**46.2** ±7.0	**32.2** ±3.2	**47.0** ±2.1	**39.4** ±4.0	**3.0** ±1.6	**8.5** ±2.7	**16.6** ±0.4	**26.7**
Qwen3.5-4B	Base	43.2	78.8	51.2	71.7	65.2	57.2	72.4	72.5	63.0	48.6	64.5
	Final	48.8 ±2.5	86.4 ±2.9	55.2 ±0.4	75.7 ±1.0	69.0 ±1.5	61.6 ±2.5	82.9 ±5.3	79.9 ±0.2	67.7 ±1.9	51.1 ±1.3	69.9
	Best	49.9 ±2.0	87.2 ±2.9	56.4 ±0.3	76.9 ±0.7	69.8 ±1.1	64.1 ±2.3	83.4 ±5.2	80.5 ±0.1	69.6 ±1.8	52.4 ±1.0	71.1
	Union	67.5 ±1.0	93.0 ±1.6	70.2 ±0.2	94.2 ±0.8	83.2 ±1.0	90.4 ±1.1	94.6 ±2.3	82.3 ±0.1	73.5 ±1.1	82.0 ±0.8	84.8
	Majority	48.5 ±1.9	87.4 ±1.8	56.1 ±0.3	79.6 ±0.8	70.2 ±1.2	61.1 ±1.7	84.3 ±3.4	79.9 ±0.1	69.0 ±1.3	49.1 ±1.1	70.7
	Gap	**18.7** ±2.7	**6.6** ±3.3	**15.0** ±0.5	**18.4** ±1.2	**14.2** ±1.8	**28.8** ±2.7	**11.6** ±5.8	**2.4** ±0.2	**5.8** ±2.2	**30.9** ±1.5	**14.9**
Gemma4-E2B	Base	43.1	67.9	38.6	41.4	48.6	35.8	67.5	66.8	69.5	47.6	53.7
	Final	46.2 ±0.4	82.2 ±7.4	53.1 ±0.7	51.7 ±4.0	53.4 ±3.9	39.9 ±0.3	78.5 ±4.0	74.4 ±2.5	72.0 ±1.2	49.1 ±0.7	61.6
	Best	47.1 ±0.3	84.0 ±5.2	56.1 ±0.5	54.2 ±3.5	56.8 ±3.4	48.0 ±0.2	79.2 ±3.5	76.8 ±1.9	72.8 ±1.0	50.3 ±0.5	64.2
	Union	69.1 ±0.3	93.4 ±5.8	69.6 ±0.5	81.9 ±2.0	79.4 ±1.6	83.3 ±0.1	93.8 ±2.7	81.7 ±1.6	77.2 ±0.7	71.6 ±0.3	81.3
	Majority	45.2 ±0.3	81.6 ±5.3	53.3 ±0.6	50.9 ±2.5	53.9 ±2.7	41.4 ±0.2	79.4 ±2.5	73.2 ±2.1	70.4 ±1.0	47.9 ±0.5	61.3
	Gap	**22.9** ±0.5	**11.2** ±9.4	**16.5** ±0.9	**30.3** ±4.5	**26.0** ±4.2	**43.4** ±0.3	**15.3** ±4.8	**7.3** ±3.0	**5.3** ±1.4	**22.5** ±0.7	**19.8**
Gemma4-E4B	Base	54.6	77.8	53.4	57.9	58.1	58.1	77.4	82.3	73.9	57.0	66.2
	Final	56.7 ±1.1	86.8 ±5.8	57.9 ±0.4	65.0 ±1.5	63.3 ±0.7	57.1 ±1.3	82.3 ±3.4	84.8 ±0.1	74.9 ±0.7	57.3 ±2.5	69.9
	Best	59.6 ±0.8	88.4 ±5.0	62.5 ±0.3	68.5 ±1.3	66.2 ±0.7	60.6 ±1.1	84.1 ±2.5	84.8 ±0.0	74.9 ±0.6	64.4 ±1.8	72.7
	Union	75.6 ±0.6	94.2 ±2.7	74.9 ±0.2	89.5 ±0.7	81.5 ±0.4	88.4 ±0.6	94.9 ±2.3	89.6 ±0.0	80.2 ±0.3	79.6 ±1.1	85.9
	Majority	54.1 ±0.8	88.0 ±4.2	59.3 ±0.3	66.4 ±1.1	66.7 ±0.5	57.1 ±1.2	84.8 ±2.9	82.9 ±0.0	73.3 ±0.5	61.6 ±2.1	71.1
	Gap	**18.9** ±1.3	**7.4** ±6.4	**17.1** ±0.4	**24.5** ±1.6	**18.2** ±0.9	**31.3** ±1.4	**12.6** ±4.1	**4.9** ±0.1	**5.3** ±0.7	**22.3** ±2.7	**16.0**

Table 18: **RLVR, full statistics.** Per model and dataset: Base (accuracy of the base model), Final, Best (single best checkpoint), Union, Majority (per-query majority vote over all $K=20$ checkpoints), and Gap. $\pm$ is the standard deviation over three training seeds.

Model		ID	MATH500	Olympiad	JEE	SciBench	GPQA-D	CruxEval	HE+	MBPP+	LCB	OOD mean
	Base	9.6	82.0	33.0	44.0	41.0	28.3	74.0	33.0	25.0	53.3	46.0
	Final	6.4 ±0.3	76.7 ±0.4	28.0 ±0.7	21.0 ±0.1	28.0 ±0.1	26.3 ±0.6	62.0 ±1.2	6.0 ±3.5	12.0 ±4.2	34.7 ±0.4	32.7
Qwen3-1.7B	Best	8.8 ±0.2	80.7 ±0.3	35.0 ±0.6	44.0 ±0.0	32.0 ±0.1	27.3 ±0.5	70.0 ±1.1	37.0 ±3.0	30.0 ±3.5	50.7 ±0.3	45.2
	Union	22.0 ±0.1	96.0 ±0.2	50.0 ±0.5	78.0 ±0.0	51.0 ±0.0	75.2 ±0.4	87.0 ±0.6	78.0 ±2.6	65.0 ±2.1	64.0 ±0.2	71.6
	Gap	**15.6** ±0.3	**19.3** ±0.4	**22.0** ±0.9	**57.0** ±0.1	**23.0** ±0.1	**49.0** ±0.8	**25.0** ±1.3	**72.0** ±4.3	**53.0** ±4.7	**29.3** ±0.4	**38.8**
	Base	4.4	89.3	47.0	66.0	54.0	46.0	81.0	34.0	59.0	65.3	60.2
	Final	14.2 ±2.4	84.1 ±2.5	42.4 ±0.9	69.5 ±3.2	49.7 ±1.4	42.8 ±1.3	80.7 ±3.2	7.6 ±1.1	13.4 ±1.0	64.9 ±1.0	50.6
Qwen3-4B	Best	15.1 ±2.3	91.3 ±2.3	44.0 ±0.7	73.0 ±3.0	54.0 ±1.1	42.8 ±1.1	87.0 ±2.6	35.3 ±1.1	21.3 ±1.0	67.6 ±0.8	57.4
	Union	24.3 ±1.5	96.7 ±1.8	56.6 ±0.5	82.4 ±2.2	77.5 ±1.1	68.9 ±0.6	93.0 ±1.6	68.2 ±0.8	72.8 ±0.8	82.7 ±0.8	77.6
	Gap	**10.1** ±2.8	**12.6** ±3.1	**14.2** ±1.0	**13.0** ±3.9	**27.8** ±1.8	**26.2** ±1.5	**12.3** ±3.6	**60.6** ±1.4	**59.5** ±1.3	**17.8** ±1.2	**27.1**
	Base	14.4	90.7	45.0	79.0	46.0	52.5	88.0	18.0	54.0	60.0	59.2
	Final	11.1 ±0.7	86.3 ±1.2	48.0 ±2.2	71.8 ±2.9	61.0 ±1.6	41.6 ±0.1	85.7 ±0.6	11.1 ±0.6	8.0 ±0.4	60.1 ±3.7	52.6
Qwen3-8B	Best	13.2 ±0.6	87.8 ±1.0	48.0 ±1.9	83.9 ±2.1	61.0 ±1.6	43.9 ±0.1	89.6 ±0.4	27.0 ±0.4	21.0 ±0.3	65.1 ±2.8	58.6
	Union	25.6 ±0.4	95.3 ±0.7	59.9 ±1.6	87.2 ±1.7	72.0 ±0.8	72.2 ±0.1	95.6 ±0.3	63.6 ±0.3	59.4 ±0.2	78.6 ±2.4	76.0
	Gap	**14.5** ±0.8	**9.0** ±1.4	**11.8** ±2.8	**15.5** ±3.4	**11.0** ±1.8	**30.6** ±0.1	**9.9** ±0.7	**52.5** ±0.7	**51.5** ±0.4	**18.5** ±4.4	**23.4**
	Base	2.0	75.3	32.0	34.0	41.0	46.0	63.0	30.0	26.0	21.3	41.0
	Final	7.2 ±2.7	64.0 ±1.2	28.0 ±1.3	30.0 ±0.9	35.0 ±0.8	28.8 ±0.8	62.0 ±4.1	32.0 ±2.5	9.0 ±0.8	13.3 ±0.6	33.6
Qwen3.5-2B	Best	8.0 ±2.0	64.7 ±0.9	29.0 ±1.3	36.0 ±0.9	42.0 ±0.6	38.4 ±0.6	67.0 ±3.8	32.0 ±1.8	27.0 ±0.7	22.7 ±0.4	39.9
	Union	18.0 ±1.1	88.7 ±0.5	47.0 ±0.8	68.0 ±0.5	72.0 ±0.5	84.3 ±0.4	84.0 ±2.4	51.0 ±1.0	46.0 ±0.6	30.7 ±0.3	63.5
	Gap	**10.8** ±2.9	**24.7** ±1.3	**19.0** ±1.6	**38.0** ±1.0	**37.0** ±1.0	**55.6** ±0.9	**22.0** ±4.8	**19.0** ±2.7	**37.0** ±1.0	**17.3** ±0.7	**30.0**
	Base	5.2	56.0	25.0	31.0	29.0	31.3	75.0	41.0	19.0	34.7	38.0
	Final	12.4 ±5.4	77.3 ±2.3	36.0 ±4.2	61.0 ±5.1	44.0 ±5.7	41.4 ±4.3	71.0 ±2.9	15.0 ±1.3	15.0 ±1.9	22.7 ±1.6	42.6
Qwen3.5-4B	Best	15.2 ±4.0	79.3 ±2.0	39.0 ±3.2	64.0 ±4.1	50.0 ±5.7	51.5 ±3.4	79.0 ±2.5	35.0 ±1.2	30.0 ±1.8	36.0 ±1.4	51.5
	Union	32.4 ±4.3	96.0 ±1.1	64.0 ±3.3	96.0 ±2.0	78.0 ±3.2	77.8 ±2.8	96.0 ±1.4	79.0 ±0.8	74.0 ±1.4	60.0 ±1.2	80.1
	Gap	**20.0** ±6.9	**18.7** ±2.6	**28.0** ±5.4	**35.0** ±5.5	**34.0** ±6.6	**36.4** ±5.1	**25.0** ±3.2	**64.0** ±1.5	**59.0** ±2.3	**37.3** ±2.0	**37.5**
	Base	8.0	66.7	29.0	45.0	35.0	36.4	74.0	59.0	50.0	42.7	48.6
	Final	20.2 ±1.5	84.4 ±1.6	41.2 ±4.5	85.7 ±5.1	53.0 ±3.6	56.9 ±5.1	75.2 ±1.9	19.2 ±0.9	25.6 ±1.3	36.9 ±2.7	53.1
Qwen3.5-9B	Best	23.7 ±1.3	85.8 ±1.1	44.3 ±4.2	88.2 ±4.7	60.7 ±3.4	67.5 ±3.7	82.2 ±1.8	43.9 ±0.7	45.6 ±1.0	52.7 ±2.3	63.4
	Union	47.7 ±1.1	98.6 ±0.7	76.9 ±2.7	99.6 ±3.6	81.8 ±1.7	79.4 ±3.9	99.1 ±1.6	93.2 ±0.6	89.8 ±1.0	83.1 ±2.1	89.1
	Gap	**27.5** ±1.9	**14.2** ±1.7	**35.7** ±5.2	**13.9** ±6.3	**28.8** ±4.0	**22.5** ±6.4	**23.9** ±2.5	**74.0** ±1.1	**64.2** ±1.6	**46.2** ±3.5	**35.9**
	Base	20.8	10.7	3.0	24.0	17.0	53.0	88.0	31.0	49.0	68.0	38.2
	Final	0.4 ±0.7	35.3 ±6.1	23.0 ±5.0	58.0 ±1.0	57.0 ±6.0	50.0 ±1.4	86.0 ±5.7	18.0 ±0.3	25.0 ±0.3	25.3 ±0.9	42.0
Gemma4-E2B	Best	20.0 ±0.6	44.0 ±5.6	24.0 ±4.3	60.0 ±1.0	64.0 ±5.1	56.1 ±1.2	92.0 ±4.2	31.0 ±0.3	43.0 ±0.3	60.0 ±0.7	52.7
	Union	20.0 ±0.5	75.3 ±3.8	46.0 ±2.4	84.0 ±0.7	80.0 ±2.6	86.4 ±0.6	97.0 ±2.9	90.0 ±0.2	76.0 ±0.2	66.7 ±0.7	77.9
	Gap	**19.6** ±0.8	**40.0** ±7.2	**23.0** ±5.6	**26.0** ±1.2	**23.0** ±6.5	**36.4** ±1.5	**11.0** ±6.4	**72.0** ±0.4	**51.0** ±0.4	**41.3** ±1.1	**36.0**
	Base	12.4	8.7	4.0	22.0	17.0	42.4	84.0	62.0	52.0	65.3	39.7
	Final	0.0 ±2.1	49.3 ±1.5	15.0 ±6.1	34.0 ±2.9	50.0 ±1.0	31.8 ±1.0	66.0 ±1.6	14.0 ±3.7	15.0 ±0.8	24.0 ±0.5	33.2
Gemma4-E4B	Best	14.4 ±1.6	54.7 ±1.5	23.0 ±6.1	42.0 ±2.2	53.0 ±0.8	42.9 ±0.9	82.0 ±1.5	70.0 ±3.1	46.0 ±0.6	57.3 ±0.4	52.3
	Union	15.2 ±1.0	85.3 ±0.8	43.0 ±4.6	69.0 ±1.3	76.0 ±0.6	76.8 ±0.6	93.0 ±1.1	93.0 ±2.7	80.0 ±0.3	64.0 ±0.3	75.6
	Gap	**15.2** ±2.3	**36.0** ±1.7	**28.0** ±7.6	**35.0** ±3.2	**26.0** ±1.2	**44.9** ±1.2	**27.0** ±2.0	**79.0** ±4.5	**65.0** ±0.9	**40.0** ±0.6	**42.3**

Table 19: **SFT, full statistics.** Per model and dataset: Base (accuracy of the base model), Final, Best (single best checkpoint), Union, and Gap. ± is the standard deviation over three training seeds.

Model		ID	MATH500	Olympiad	JEE	SciBench	GPQA-D	CruxEval	HE+	MBPP+	LCB	OOD mean
	Base	9.6	82.0	33.0	44.0	41.0	28.3	74.0	33.0	25.0	53.3	46.0
	Final	9.2 ±1.8	86.7 ±1.6	41.0 ±3.5	36.0 ±1.3	49.0 ±1.8	32.8 ±1.8	62.0 ±0.1	19.0 ±1.9	11.0 ±1.6	49.3 ±1.4	43.0
Qwen3-1.7B	Best	12.8 ±1.7	89.3 ±1.3	43.0 ±2.6	48.0 ±1.2	51.0 ±1.7	37.9 ±1.3	67.0 ±0.1	36.0 ±1.7	35.0 ±1.2	57.3 ±1.2	51.6
	Union	28.0 ±1.4	96.7 ±0.9	58.0 ±1.9	83.0 ±0.7	68.0 ±0.9	78.8 ±1.1	88.0 ±0.1	82.0 ±0.9	78.0 ±0.8	74.7 ±0.6	78.6
	Gap	**18.8** ±2.2	**10.0** ±1.9	**17.0** ±4.0	**47.0** ±1.5	**19.0** ±2.0	**46.0** ±2.1	**26.0** ±0.2	**63.0** ±2.1	**67.0** ±1.8	**25.3** ±1.5	**35.6**
	Base	4.4	89.3	47.0	66.0	54.0	46.0	81.0	34.0	59.0	65.3	60.2
	Final	13.2 ±4.8	92.7 ±1.5	47.0 ±0.6	67.0 ±0.1	70.0 ±5.2	48.5 ±4.5	82.0 ±0.4	17.0 ±0.0	21.0 ±1.0	60.0 ±0.8	56.1
Qwen3-4B	Best	16.8 ±3.5	94.7 ±1.1	53.0 ±0.5	69.0 ±0.0	72.0 ±4.1	54.0 ±4.2	84.0 ±0.3	49.0 ±0.0	49.0 ±0.8	68.0 ±0.7	65.9
	Union	32.8 ±2.9	98.7 ±0.8	67.0 ±0.3	88.0 ±0.0	88.0 ±2.7	79.8 ±3.2	96.0 ±0.2	91.0 ±0.0	81.0 ±0.7	84.0 ±0.5	85.9
	Gap	**19.6** ±5.6	**6.0** ±1.7	**20.0** ±0.7	**21.0** ±0.1	**18.0** ±5.8	**31.3** ±5.6	**14.0** ±0.5	**74.0** ±0.0	**60.0** ±1.2	**24.0** ±1.0	**29.8**
	Base	14.4	90.7	45.0	79.0	46.0	52.5	88.0	18.0	54.0	60.0	59.2
	Final	18.4 ±4.8	94.7 ±3.2	56.0 ±3.5	74.0 ±0.5	73.0 ±2.2	58.6 ±4.7	85.0 ±1.2	7.0 ±6.4	9.0 ±0.9	65.3 ±3.4	58.1
Qwen3-8B	Best	22.4 ±4.6	96.0 ±2.5	59.0 ±2.7	82.0 ±0.4	79.0 ±1.9	59.6 ±4.4	92.0 ±0.9	38.0 ±4.5	55.0 ±0.8	70.7 ±2.6	70.1
	Union	39.2 ±2.4	98.7 ±1.4	72.0 ±1.5	97.0 ±0.3	88.0 ±1.8	83.8 ±2.2	97.0 ±0.8	87.0 ±3.3	80.0 ±0.4	82.7 ±1.9	87.4
	Gap	**20.8** ±5.4	**4.0** ±3.5	**16.0** ±3.8	**23.0** ±0.5	**15.0** ±2.8	**25.2** ±5.2	**12.0** ±1.4	**80.0** ±7.2	**71.0** ±1.0	**17.3** ±3.9	**29.3**
	Base	2.0	75.3	32.0	34.0	41.0	46.0	63.0	30.0	26.0	21.3	41.0
	Final	4.4 ±2.1	76.0 ±0.9	30.0 ±1.7	29.0 ±4.0	52.0 ±3.9	32.8 ±0.8	60.0 ±3.4	17.0 ±1.2	6.0 ±0.6	14.7 ±1.1	35.3
Qwen3.5-2B	Best	7.6 ±1.6	78.0 ±0.9	36.0 ±1.6	44.0 ±3.9	56.0 ±3.0	40.9 ±0.8	70.0 ±2.5	27.0 ±0.9	28.0 ±0.5	21.3 ±1.1	44.6
	Union	19.2 ±1.4	94.7 ±0.6	51.0 ±0.7	80.0 ±2.5	82.0 ±2.8	84.8 ±0.7	88.0 ±2.5	57.0 ±0.6	52.0 ±0.3	34.7 ±0.5	69.4
	Gap	**14.8** ±2.5	**18.7** ±1.1	**21.0** ±1.8	**51.0** ±4.7	**30.0** ±4.8	**52.0** ±1.1	**28.0** ±4.2	**40.0** ±1.3	**46.0** ±0.7	**20.0** ±1.2	**34.1**
	Base	5.2	56.0	25.0	31.0	29.0	31.3	75.0	41.0	19.0	34.7	38.0
	Final	8.8 ±1.0	89.3 ±0.0	50.0 ±2.4	71.0 ±5.7	61.0 ±1.3	63.1 ±3.3	86.0 ±2.9	17.0 ±5.2	18.0 ±3.7	34.7 ±0.8	54.5
Qwen3.5-4B	Best	10.4 ±0.9	91.3 ±0.0	50.0 ±1.9	79.0 ±4.0	77.0 ±1.2	66.2 ±2.9	89.0 ±2.4	49.0 ±3.6	46.0 ±2.7	54.7 ±0.6	66.9
	Union	17.6 ±0.7	97.3 ±0.0	66.0 ±1.7	95.0 ±2.9	86.0 ±1.0	87.4 ±2.2	98.0 ±1.4	86.0 ±3.9	80.0 ±2.2	78.7 ±0.3	86.0
	Gap	**8.8** ±1.2	**8.0** ±0.0	**16.0** ±3.0	**24.0** ±6.4	**25.0** ±1.6	**24.2** ±4.0	**12.0** ±3.2	**69.0** ±6.5	**62.0** ±4.3	**44.0** ±0.8	**31.6**
	Base	8.0	66.7	29.0	45.0	35.0	36.4	74.0	59.0	50.0	42.7	48.6
	Final	8.4 ±1.1	92.7 ±1.4	50.0 ±0.2	84.0 ±5.0	74.0 ±5.5	68.7 ±1.0	86.0 ±3.8	26.0 ±1.4	21.0 ±0.1	46.7 ±1.7	61.0
Qwen3.5-9B	Best	12.0 ±0.8	94.0 ±1.2	55.0 ±0.2	86.0 ±4.1	76.0 ±4.6	71.2 ±0.7	92.0 ±3.0	44.0 ±1.1	53.0 ±0.1	57.3 ±1.4	69.8
	Union	19.6 ±0.7	97.3 ±1.0	67.0 ±0.1	98.0 ±3.7	87.0 ±2.3	91.9 ±0.6	99.0 ±2.1	94.0 ±0.9	80.0 ±0.0	73.3 ±1.3	87.5
	Gap	**11.2** ±1.3	**4.7** ±1.8	**17.0** ±0.2	**14.0** ±6.2	**13.0** ±6.0	**23.2** ±1.1	**13.0** ±4.3	**68.0** ±1.7	**59.0** ±0.1	**26.7** ±2.1	**26.5**
	Base	12.4	8.7	4.0	22.0	17.0	42.4	84.0	62.0	52.0	65.3	39.7
	Final	10.4 ±1.0	6.7 ±0.2	0.0 ±0.3	28.0 ±0.7	11.0 ±1.8	31.3 ±1.1	84.0 ±0.1	17.0 ±1.0	16.0 ±0.8	30.7 ±1.0	25.0
Gemma4-E2B	Best	13.6 ±0.8	11.3 ±0.2	4.0 ±0.3	32.0 ±0.6	21.0 ±1.8	41.4 ±0.9	88.0 ±0.1	49.0 ±1.0	44.0 ±0.6	56.0 ±0.8	38.5
	Union	19.2 ±0.6	17.3 ±0.1	6.0 ±0.1	62.0 ±0.4	42.0 ±1.1	78.3 ±0.5	95.0 ±0.0	91.0 ±0.8	76.0 ±0.3	73.3 ±0.7	60.1
	Gap	**8.8** ±1.2	**10.7** ±0.2	**6.0** ±0.3	**34.0** ±0.8	**31.0** ±2.1	**47.0** ±1.2	**11.0** ±0.1	**74.0** ±1.3	**60.0** ±0.8	**42.7** ±1.2	**35.2**
	Base	20.8	10.7	3.0	24.0	17.0	53.0	88.0	31.0	49.0	68.0	38.2
	Final	17.2 ±1.3	16.7 ±2.7	7.0 ±1.0	36.0 ±4.0	37.0 ±1.0	52.5 ±1.3	92.0 ±2.9	21.0 ±6.0	19.0 ±0.7	34.7 ±0.0	35.1
Gemma4-E4B	Best	18.8 ±1.3	18.0 ±2.0	12.0 ±0.7	47.0 ±2.8	42.0 ±0.7	57.6 ±1.2	93.0 ±2.7	72.0 ±4.3	41.0 ±0.5	69.3 ±0.0	50.2
	Union	22.8 ±0.6	24.7 ±1.4	21.0 ±0.8	70.0 ±3.1	70.0 ±0.5	83.8 ±0.9	96.0 ±1.2	91.0 ±3.5	81.0 ±0.4	78.7 ±0.0	68.5
	Gap	**5.6** ±1.4	**8.0** ±3.1	**14.0** ±1.2	**34.0** ±5.0	**33.0** ±1.1	**31.3** ±1.6	**4.0** ±3.1	**70.0** ±6.9	**62.0** ±0.8	**44.0** ±0.0	**33.4**

Table 20: **SDFT, full statistics.** Per model and dataset: Base (accuracy of the base model), Final, Best (single best checkpoint), Union, and Gap. ± is the standard deviation over three training seeds.

Model		ID	MATH500	Olympiad	JEE	SciBench	GPQA-D	CruxEval	HE+	MBPP+	LCB	OOD mean
	Base	23.7	68.1	30.4	19.9	17.6	9.4	66.1	46.8	55.8	28.1	38.0
	Final	25.9 ±3.0	69.4 ±3.0	31.3 ±0.7	20.0 ±0.2	19.8 ±2.7	10.6 ±0.8	65.8 ±0.8	50.6 ±1.8	54.8 ±0.3	29.6 ±0.9	39.1
Qwen3-8B→1.7B	Best	26.9 ±2.2	73.2 ±2.4	34.4 ±0.6	23.3 ±0.2	23.0 ±2.0	12.1 ±0.6	66.9 ±0.7	51.8 ±1.4	57.4 ±0.2	30.2 ±0.8	41.4
	Union	42.8 ±1.3	86.2 ±1.8	46.0 ±0.3	42.7 ±0.1	36.0 ±1.5	24.2 ±0.4	83.6 ±0.5	71.3 ±0.8	70.9 ±0.1	42.0 ±0.6	55.9
	Gap	**16.9** ±3.3	**16.8** ±3.5	**14.7** ±0.7	**22.7** ±0.2	**16.2** ±3.1	**13.6** ±0.9	**17.9** ±0.9	**20.7** ±1.9	**16.1** ±0.3	**12.4** ±1.1	**16.8**
	Base	26.2	72.0	41.3	40.2	51.5	31.9	52.8	39.5	30.2	14.7	41.5
	Final	28.6 ±0.6	75.6 ±0.5	40.2 ±0.9	35.0 ±1.0	50.9 ±1.5	27.8 ±0.9	61.5 ±2.1	39.6 ±1.9	36.5 ±3.0	18.1 ±2.2	42.8
Qwen3.5-9B→2B	Best	29.5 ±0.5	76.4 ±0.4	43.8 ±0.7	42.9 ±0.8	52.7 ±1.1	33.3 ±0.9	62.0 ±2.0	42.7 ±1.9	39.9 ±2.8	19.2 ±2.2	45.9
	Union	49.0 ±0.3	90.2 ±0.2	59.6 ±0.4	70.9 ±0.5	73.6 ±0.6	59.6 ±0.5	79.4 ±1.7	56.7 ±1.1	54.0 ±2.2	33.4 ±1.4	64.2
	Gap	**20.4** ±0.7	**14.6** ±0.5	**19.4** ±1.0	**35.9** ±1.1	**22.7** ±1.6	**31.8** ±1.0	**17.9** ±2.7	**17.1** ±2.2	**17.5** ±3.8	**15.3** ±2.6	**21.4**
	Base	43.9	60.5	35.4	24.7	44.0	33.0	76.4	56.8	51.9	48.8	47.9
	Final	46.0 ±2.6	62.2 ±5.0	36.5 ±0.4	32.0 ±4.2	48.7 ±4.7	38.4 ±6.3	80.2 ±3.1	59.8 ±1.9	54.8 ±1.1	49.0 ±0.7	51.3
Gemma4-E4B→E2B	Best	47.3 ±1.9	66.4 ±4.3	40.9 ±0.4	34.0 ±3.0	52.2 ±4.4	44.9 ±5.7	81.4 ±2.7	62.8 ±1.6	55.6 ±0.9	51.8 ±0.6	54.4
	Union	72.1 ±2.0	90.4 ±3.7	69.9 ±0.3	66.6 ±3.0	75.9 ±3.0	86.4 ±4.8	92.9 ±1.2	69.5 ±0.8	65.6 ±0.6	71.8 ±0.5	76.6
	Gap	**26.1** ±3.3	**28.2** ±6.2	**33.4** ±0.5	**34.6** ±5.2	**27.2** ±5.5	**48.0** ±7.9	**12.6** ±3.4	**9.8** ±2.1	**10.8** ±1.3	**22.7** ±0.8	**25.3**

Table 21: **OPD, full statistics.** Per model and dataset: Base (accuracy of the base model), Final, Best (single best checkpoint), Union, and Gap. ± is the standard deviation over three training seeds.

G.2 PER-MODEL RADAR PLOTS

Figures 19–22 break Figure 3 down by model. Each axis is the forgetting gap on one OOD dataset, in points; the ID gap and the OOD mean are given below each panel. The numbers are the Gap rows of Tables 18–21.

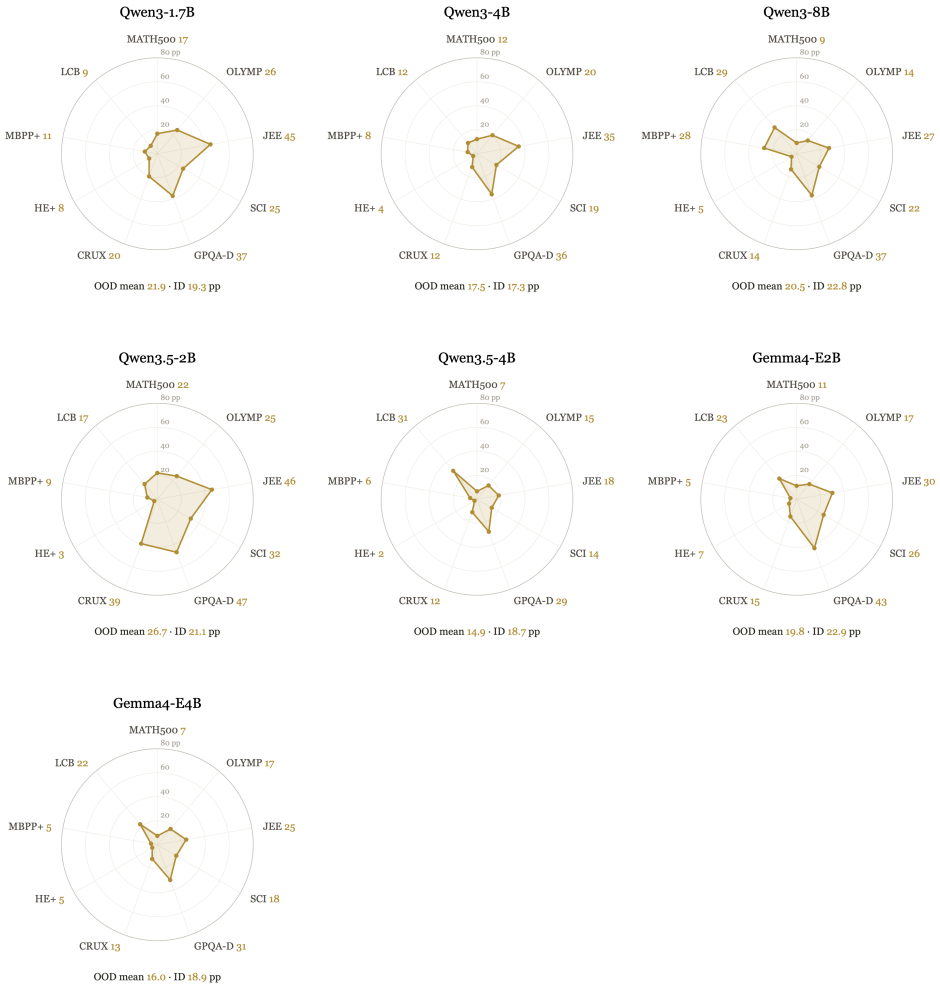

Figure 19: **Forgetting gap under RLVR, per model.** Seven models; axes are the nine OOD datasets.



Figure 20: **Forgetting gap under SFT, per model.** Eight models; as Figure 19.

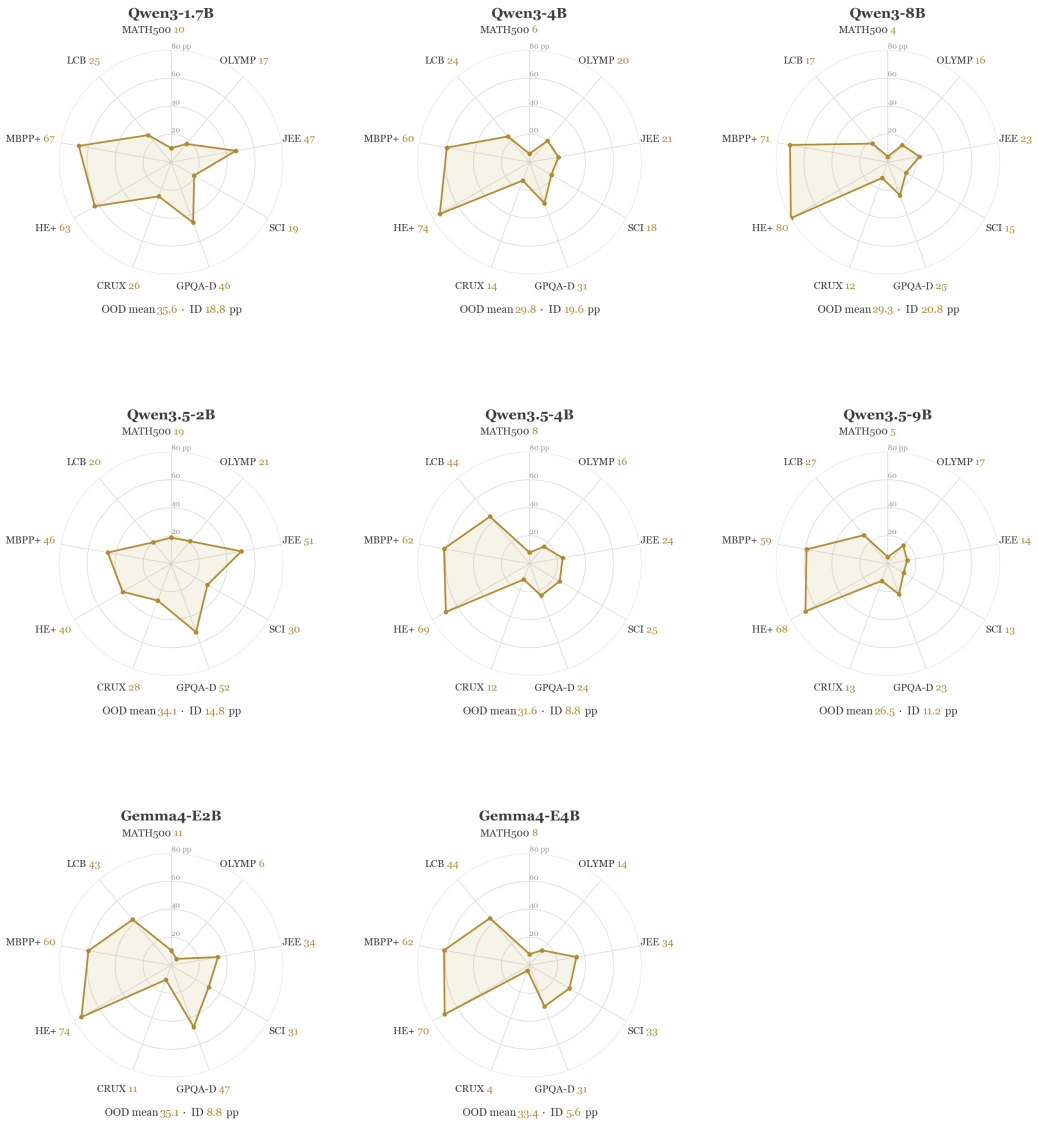

Figure 21: **Forgetting gap under SDFT, per model.** Eight models; as Figure 19.

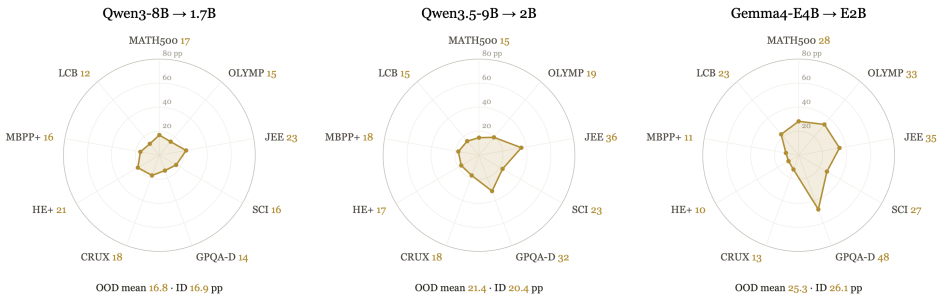

Figure 22: **Forgetting gap under OPD, per teacher→student pair.** As Figure 19.

G.3 Forgetting gap changes little with model size

Figure 23 plots the forgetting gap against model size, one panel per OOD dataset and one line per method, with the same size classes as Figure 9. Because the gap of a single run is one number, we take the gap at each of the last ten checkpoints (updates 275–500, where the gap has plateaued; Figures 24–27) of every model in a class as the measurements. Going from ∼2B to ∼8B, the final-checkpoint gap averaged over the nine datasets changes by -2.3 points under RLVR ($22.8 \rightarrow 20.5$), -2.2 under SFT ($37.0 \rightarrow 34.8$), and -7.0 under SDFT ($34.9 \rightarrow 27.9$); the OPD students are all ∼2B. Fitting each model's final gap against $\log_2$ size for the 27 method–dataset pairs gives a negative slope in 17 cases and a positive one in 10; only three are significant at $p < 0.05$, all under SDFT (MATH500, JEEBench, GPQA-Diamond). On the code datasets the SFT and SDFT gaps stay at 50–67 points at every size, and the RLVR gap on MBPP+ grows with size.

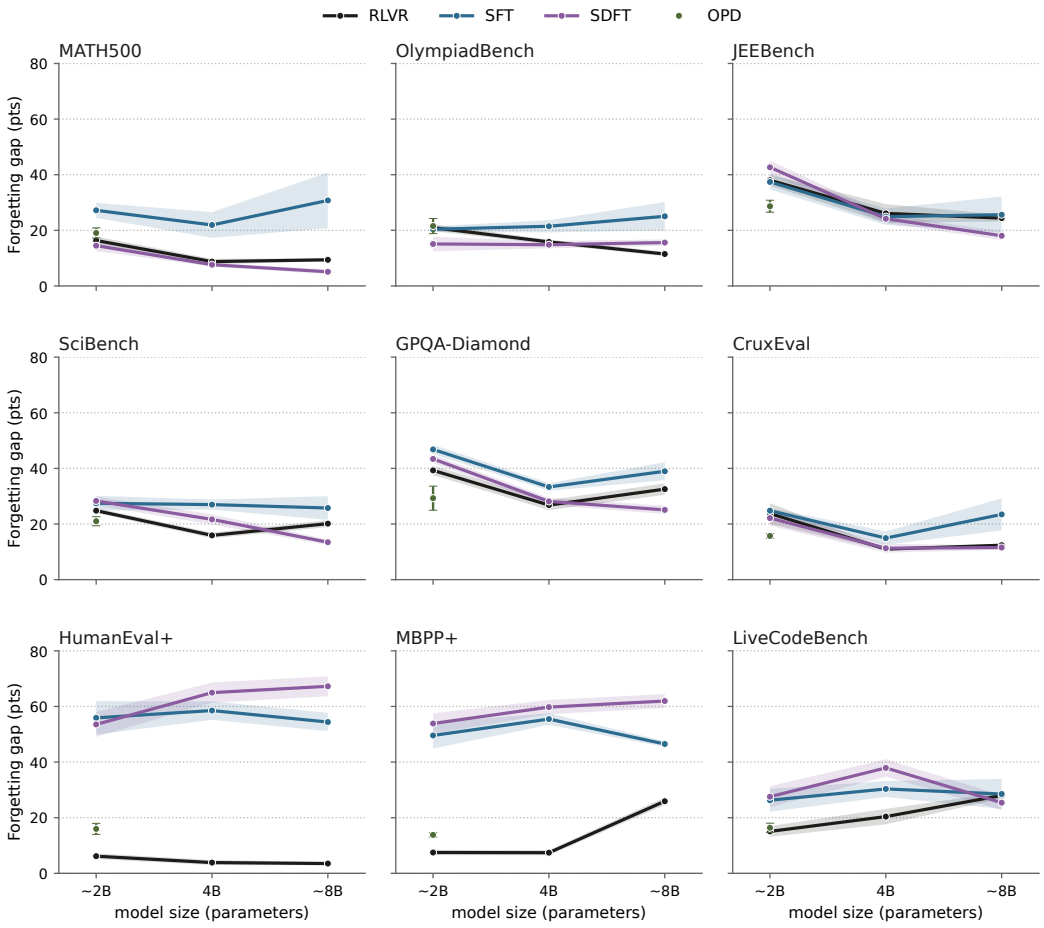

Figure 23: **Forgetting gap versus model size, per OOD dataset.** Each point is the mean forgetting gap (Equation 7, evaluated at each of the last ten checkpoints) over all models in a size class (∼2B: 1.7–2B models; 4B; ∼8B: 8–9B models); bands and whiskers are 95% confidence intervals over those checkpoints. OPD is a single point because every student is ∼2B.

G.4 Forgetting gap over training

Figures 24–27 show, per model and dataset, the accuracy of the current checkpoint next to the fraction of queries solved by at least one checkpoint so far; the vertical distance between the two curves is the forgetting gap at that checkpoint.

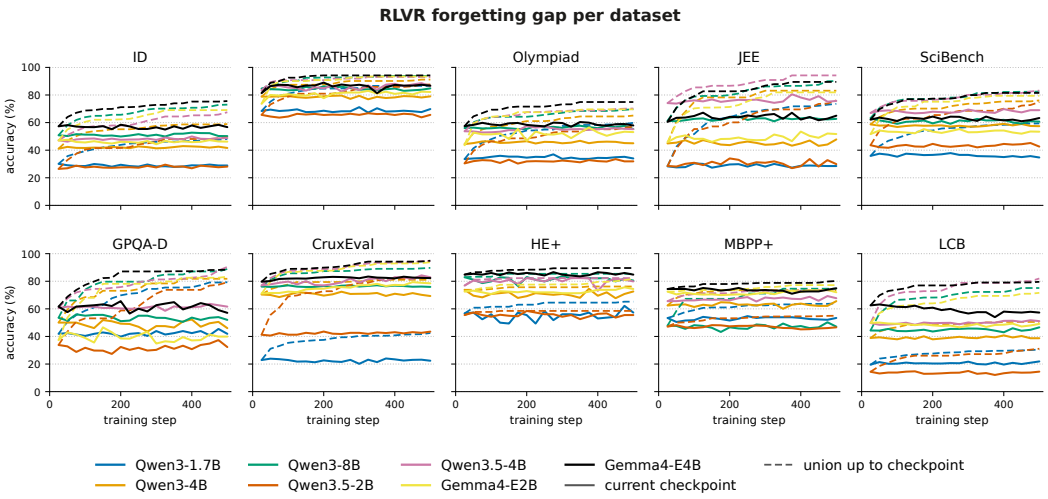

Figure 24: **Forgetting gap over training under RLVR**, per dataset and model. Solid: accuracy of the current checkpoint; dashed: fraction of queries solved by at least one checkpoint so far.

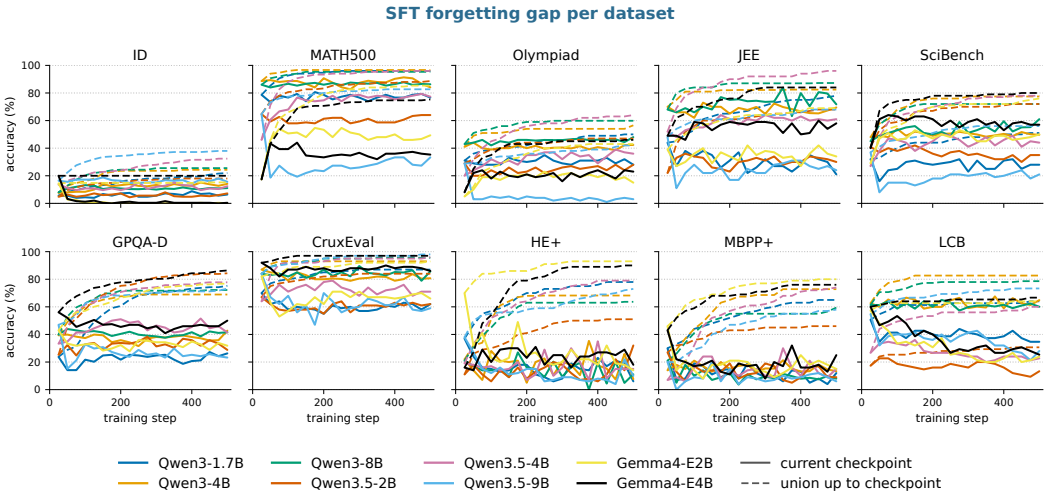

Figure 25: **Forgetting gap over training under SFT.** As Figure 24.

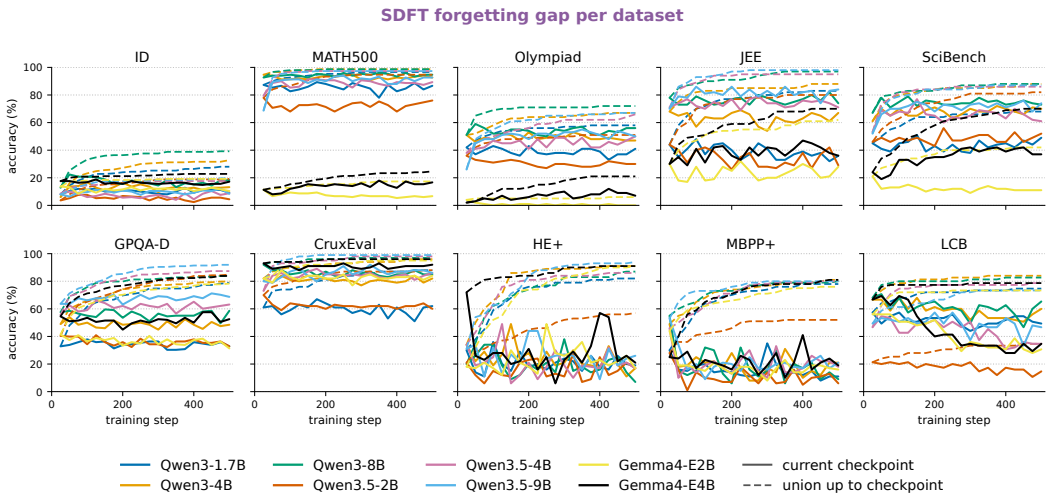

Figure 26: **Forgetting gap over training under SDFT.** As Figure 24.

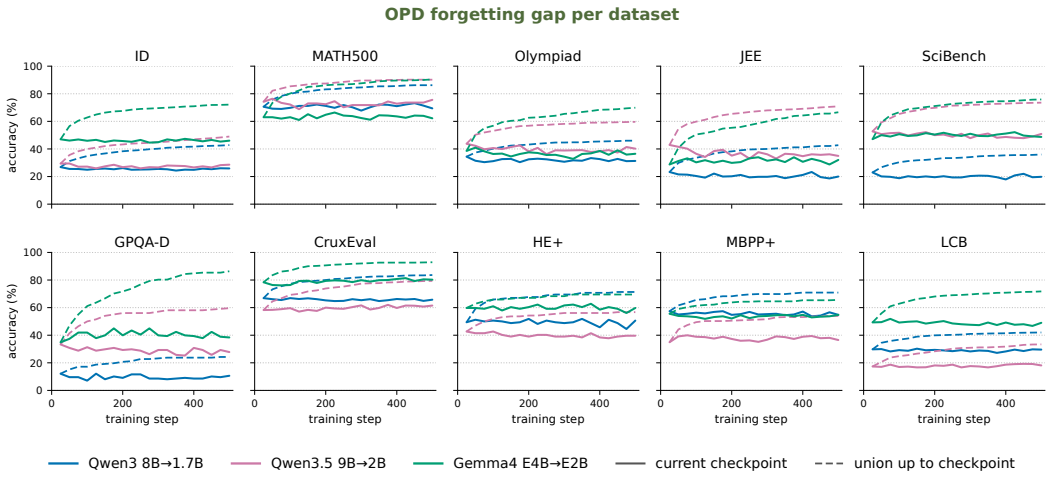

Figure 27: **Forgetting gap over training under OPD.** As Figure 24; one pair of curves per teacher→student pair.

H CORRECT FINAL ANSWERS ARE RECOVERED NEGATIVE FLIPS

For each run we take the queries that the final checkpoint θ_T solves and, over the 20 checkpoints measured during training, count how many times each of them flipped from correct to incorrect. We report the ratio

$$\frac{\#\{q \in S(\theta_T) \,:\, q \text{ flips C}\to\text{W at least } k \text{ times during training}\}}{|S(\theta_T)|}, \tag{9}$$

together with the fraction that flips exactly k times (Table 22). For k=3, the ratio is 18.4% under RLVR, 36.7% under SFT, 34.7% under SDFT, and 24.2% under OPD, 28.5% on average; that is, 28.5% of final accuracy, averaged over the four methods, comes from queries that flipped at least three times. A query that has been lost and regained three or more times is not stably solved: its correctness at θ_T reflects which side of the edge it happened to be on when training stopped, so more than a quarter of final accuracy rests on answers that the next update could flip.

Method		$k=1$	$k=2$	$k=3$	$k=4$	$k=5$
RLVR	$\geq k$	**74.0** ±6.7	**40.8** ±9.6	**18.4** ±7.6	**7.3** ±4.4	**2.6** ±2.2
	$= k$	33.2 ±3.8	22.4 ±2.5	11.1 ±3.2	4.7 ±2.3	1.8 ±1.2
SFT	$\geq k$	**90.8** ±2.8	**65.8** ±4.8	**36.7** ±3.7	**16.3** ±1.9	**5.9** ±1.1
	$= k$	25.1 ±2.4	29.0 ±1.5	20.4 ±2.2	10.4 ±1.2	4.2 ±0.6
SDFT	$\geq k$	**86.4** ±3.3	**60.7** ±5.5	**34.7** ±4.7	**16.5** ±2.6	**6.7** ±1.1
	$= k$	25.7 ±2.6	26.1 ±0.9	18.1 ±2.2	9.9 ±1.6	4.4 ±0.7
OPD	$\geq k$	**83.3** ±9.5	**52.5** ±10.4	**24.2** ±5.4	**8.2** ±2.5	**2.0** ±0.9
	$= k$	30.8 ±3.8	28.3 ±6.8	16.1 ±3.7	6.1 ±1.6	1.7 ±0.7

Table 22: **Many correct final answers are recovered negative flips.** Among the queries the final checkpoint θ_T answers correctly, the percentage that went through at least k ($\geq k$) or exactly k ($= k$) correct→incorrect flips over the 20 checkpoints of training before ending correct. Pooled over all evaluated datasets (ID and OOD); mean over runs with the half-width of the 95% t-interval across runs. 100 minus the ≥ 1 entry is the share of correct final answers that were never lost during training.

I AI USAGE

We used AI assistants to aid and polish the writing of this paper, as well as to help find related work. All research ideas, experiments, analyses, and results are the authors' own, and the authors take full responsibility for the content of the paper.