# CONTINUALLY POST-TRAINING SMALL MODELS VIA SYNTHETIC SELF-DISTILLATION

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Continual post-training faces a dual challenge of acquisition and retention: models must learn new capabilities while preserving prior knowledge. While Self-Distillation Fine-Tuning (SDFT) mitigates this using self-supervision with in-context demonstrations, it relies on strong in-context learning and instruction-following abilities. Consequently, SDFT fails on smaller models critical for low-latency, on-device use. To enable continual learning at smaller scales, we first identify the root cause of failure. We show that poor acquisition and forgetting correlate linearly with the post-trained model's distribution shift, revealing the necessity of in-distribution learning. We thus introduce Synthetic Self-Distillation (SYNTHSELF), imitating SDFT's in-distribution rewrite by outsourcing the task to a minimally larger, same-family aligner. Coupled with off-policy supervision and rejection sampling, this provides the high-quality supervision small models require without breaking their original distribution. Evaluations spanning **4** acquisition datasets (including a mixture) and **14** retention benchmarks show SYNTHSELF consistently outperforms **6** methods (including off- and on-policy distillation) across **3** small model scales. Extensive further analyses confirm SYNTHSELF paves the way toward effective small-model continual learning.

## 1 INTRODUCTION

Continual post-training of language models (LMs) is a common setting where practitioners aim to inject new capabilities into already-trained models for scenarios such as customizing for specific use-cases and enabling human-like lifelong learning (Shi et al., 2024; Wang et al., 2024a; Zheng et al., 2025a;b; Srijith et al., 2025; Chen et al., 2026a). Yet, this process is notoriously challenging on both the acquisition (injecting new abilities) and retention (maintaining existing capabilities) axes (Harmon et al., 2026; Luo et al., 2025; Chen et al., 2026b). The failure to retain existing capabilities, widely known as catastrophic forgetting (McCloskey & Cohen, 1989; French, 1999; Kirkpatrick et al., 2017), is particularly well-documented in settings where a supervised fine-tuning (SFT) dataset is used for teacher forcing (Kalajdzievski, 2024; Luo et al., 2025; Lai et al., 2026; Chen et al., 2026b; Shenfeld et al., 2026b).

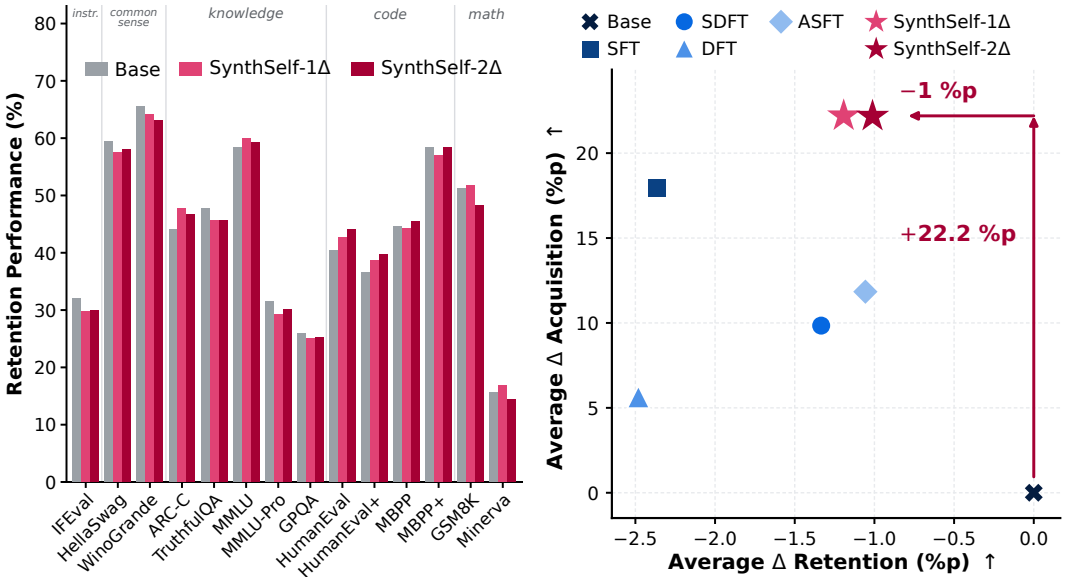
A recent work showed that self-distillation fine-tuning (SDFT) is an effective way to inject SFT demonstrations in an on-policy manner (Shenfeld et al., 2026a), outperforming recent off-policy continual fine-tuning methods such as re-invoke (Lu & Lab, 2025) and Dynamic Fine-tuning (DFT; Wu et al. (2026)) on both acquisition and retention axes. SDFT leverages the in-context learning (ICL) and instruction-following (IF) abilities of post-trained models to provide demonstrations in-context to the self-teacher model, and applies on-policy distillation (OPD; Ross et al. (2011), Agarwal et al. (2024)) on itself, supervising on an implicit self-rewrite of the demonstrations.

However, as models struggle with ICL and IF at smaller scales, SDFT remains infeasible in this regime (see Fig. 3a). We aim to address this limitation, as small models have become critical for applications such as on-device deployment, latency-specialized models, draft models, verifiers, and large-small collaborative methods (Cobbe et al., 2021; Chen et al., 2023; Abdin et al., 2024; Liu et al., 2024; Lu et al., 2025; Belcak et al., 2025; Cai et al., 2026).

**Contribution.** To better understand how to leverage the key principles of SDFT, we first examine why it succeeds in continual post-training (§ 2.2.1). We reveal that the distribution shift of the post-



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.




(a) Averaged across 3 base models, continual post-training via SYNTHSELF on the medical domain results in near-zero forgetting across all 14 benchmarks.

(b) Across 3 base models, 4 acquisition benchmarks, and 14 retention tasks, SYNTHSELF maximizes performance gains with minimal forgetting.

Figure 1: **Overview of SYNTHSELF performance across three model scales (`0.5`, `1.5`, and `3B`).**

trained base model linearly correlates with poor acquisition and forgetting (**Fig. 2**), indicating that staying in-distribution is crucial for continual learning. While the primary advantage of SDFT over off-policy methods is its ability to staying in distribution, evaluating its self-generated offline dataset reveals that SDFT's self-supervision quality is inadequate at smaller model scales (**Fig. 3**).

In response, we propose Synthetic Self-Distillation (SYNTHSELF), which *imitates* self-distillation's implicit in-distribution rewrite by outsourcing the distribution-aligned rewrite to a larger model (§ 2.4). Our method distinctly differs from SDFT in the following ways: **(i)** choosing off-policy over on-policy supervision allows us to apply rejection sampling, which in turn improves supervision quality for small models that are sensitive to data quality; and **(ii)** using minimally larger, same-family aligners allows us to simultaneously **(ii.a.)** provide high-quality supervision (**Tab. 2**), **(ii.b.)** remain in-distribution relative to the base model (**Fig. 4**), and **(ii.c.)** maintain low computational costs. We empirically validate SYNTHSELF (§ 3), demonstrating that:

1. SYNTHSELF achieves the best acquisition across **3** domains (science, tool-use, and medicine) and a 4-subject science mixture while preserving prior capabilities on **14** benchmarks, consistently across **3** base model sizes (`0.5`, `1.5`, `3B`), and against **4** relevant baselines (**Fig. 1, Tab. 1**).

2. Controlling for the confounding factor of teacher model size, we demonstrate that SYNTHSELF outperforms both on- and off-policy distillation in acquisition by **+5.7%p** to **+16.9%p**, averaged over all four settings and three base scales (**Tab. 1**, **Fig. 5**).

3. We show that utilizing a smaller (weaker) in-distribution aligner is more effective than deploying a larger (stronger) out-of-distribution teacher, yielding a **69.9%** reduction in alignment computational cost (FLOPs) (**Tab. 2**).

4. We validate that our improvement over SDFT is not due to prompt differences using prompt-matched and diverse prompts, achieving **+18.96%p** and **+1.68%p** gains in acquisition and retention over prompt-matched SDFT (**Fig. 6**).

5. Finally, our ablation studies confirm that combining rejection sampling with off-policy supervision is a meaningful component of our methodology (**Fig. 5**).

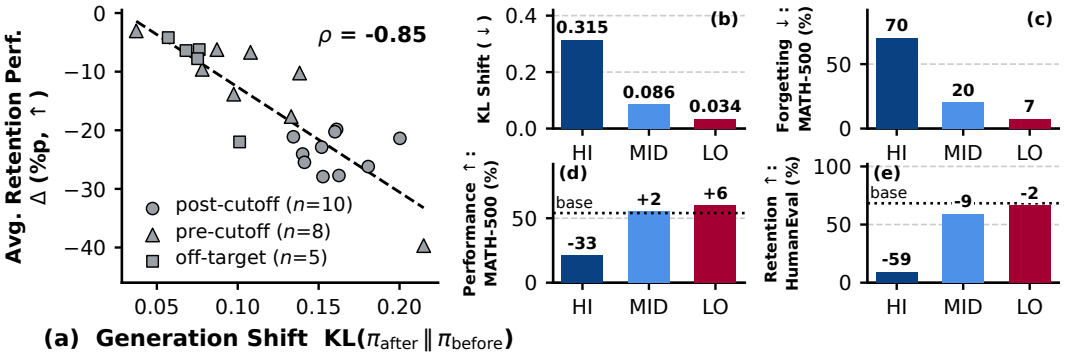



Figure 2: **Distribution shift negatively correlates with ideal continual learning. (a) Experiment 1:** Post-training distribution shift strongly correlates with forgetting existing capabilities across GPQA-Diamond, ARC-C, MMLU-Pro, HumanEval, and MBPP, regardless of data source or type. **(b) - (e) Experiment 2:** Transforming the same dataset from its original OOD form (HI) to more in-distribution forms (MID, LO) monotonically improves **(c)** target domain (math) retention, **(d)** target domain performance, and **(e)** cross-domain (coding) retention.

## 2 Method

### 2.1 Problem Setting: Continual Fine-tuning

**Continual fine-tuning.** Given a post-trained policy $\pi$, we continually inject new knowledge from a dataset $\mathcal{D}_{\text{train}} = \{(x_i, y_i)\}_{i=1}^{N}$ by further training $\pi$. In doing so, we aim to maximize **(i)** acquisition: performance on a same-domain test set $\mathcal{D}_{\text{test}}$, while maintaining **(ii)** retention: performance on existing capabilities, evaluated via $\mathcal{D}_{\text{ret}}$.

**Self-distillation fine-tuning.** Given a demonstration pair $(x, y)$ and an SDFT prompt $P^{\text{SDFT}}$ that places the demonstration in-context, SDFT trains the policy $\pi_\theta$ parameterized by $\theta$ on-policy toward the frozen model's self-rewrite of that demonstration,

$$\mathcal{L}(\theta) \;=\; D_{\text{KL}}\big(\pi_\theta(\cdot \mid x) \,\big\|\, \pi\big(\cdot \mid P^{\text{SDFT}}(x, y)\big)\big) \;=\; \mathbb{E}_{y' \sim \pi_\theta(\cdot|x)}\left[\log \frac{\pi_\theta(y' \mid x)}{\pi(y' \mid P^{\text{SDFT}}(x, y))}\right], \quad (1)$$

where $y'$ denotes a completion sampled from the current policy. This supervision preserves the semantic content of $y$ while remaining within the model's native distribution, allowing for effective continual learning.

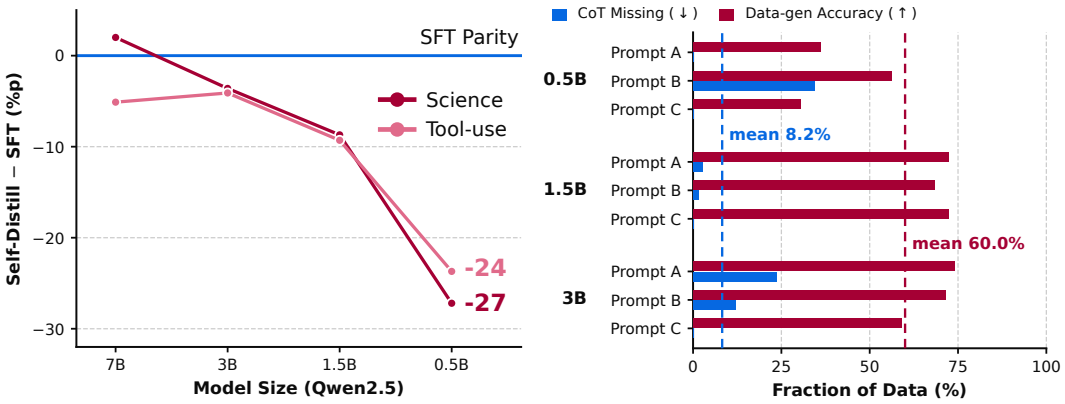
### 2.2 In-distribution Demonstration Injection

#### 2.2.1 In-distribution learning is key for continual learning

Building on the observation that SDFT preserves the native distribution, we hypothesize that training on out-of-distribution (OOD) targets is the primary driver of poor acquisition and forgetting. To test this, we continually fine-tune `Qwen2.5-3B-Instruct` (Qwen et al., 2025), measuring the resulting policy shift via $D_{\text{KL}}(\pi_{\text{after}} \,\|\, \pi_{\text{before}})$. To avoid data contamination, we use this older model to evaluate against post-release datasets. Further details are available in Appendix B.

**Experiment 1: Distribution shift correlates with forgetting across datasets.** **[Setup]** First, we collect 23 diverse math training datasets, size-matched to 2.5M completion tokens. Datasets are grouped by domain and release date relative to the base model: **(1)** pre-cutoff, **(2)** post-cutoff: long chain-of-thought (CoT) math reasoning, and **(3)** off-target: post-cutoff code or general instruction data.[1] Forgetting is measured as the mean performance drop ($\Delta$) across five non-math benchmarks: GPQA-Diamond (Rein et al., 2024), ARC-C (Clark et al., 2018), MMLU-Pro (Wang et al., 2024b),

[1] Although a dataset's release date does not definitively guarantee whether its tokens were seen during training, it serves as a reasonable proxy for distinguishing between older and newer data. Our overall results and analyses remain robust regardless of this specific grouping.




(a) SDFT acquisition performance relative to SFT declines monotonically as model size decreases.

(b) For example, in the tool-use domain, self-rewriting results in reduced accuracy and missing CoTs.

Figure 3: **SDFT fails at smaller model sizes as its ability to self-teach deteriorates.**

HumanEval (Chen et al., 2021), and MBPP (Odena et al., 2021). **[Results]** Fig. 2 (a) demonstrates that the magnitude of forgetting (y-axis) is strongly correlated with the distribution shift of the policy $\pi$ after fine-tuning (Spearman $\rho = -0.85$), suggesting that updates causing significant distribution shifts are detrimental to continual learning.

**Experiment 2: Distribution shift correlates with poor acquisition and forgetting within a dataset.** **[Setup]** Second, to isolate distribution shift within an underlying dataset, we hold the corpus fixed using OpenR1-Math-220k (Face, 2025), a long-CoT dataset. Using identical prompts, we construct three 2M completion-token datasets that vary only in target text shift: (**HI**) high shift: full R1 trace with reasoning, (**MID**) mid shift: R1 final solution without long-CoT, and (**LO**) low shift: the model's own verified-correct solution. LO is curated via self-rejection sampling (Singh et al., 2024). The base model decodes a single solution per problem under the same prompt, which is then verified for correctness. **[Results]** Fig. 2 (c) - (e) show that higher distribution shift tracks with in-domain forgetting and catastrophic forgetting, whereas low-shift self-generated samples improve math acquisition performance while preserving existing coding capabilities.

### 2.3 Limitations of Self-Distillation

**Self-supervision fails at small model scales.** Although on-policy self-distillation (SDFT; Shenfeld et al. (2026a)) successfully preserves the base distribution during continual post-training, its efficacy drops for smaller models. Fig. 3a demonstrates that SDFT's performance relative to SFT decreases monotonically with student size. We attribute this to poor self-supervision quality. To evaluate this, we generate an offline dataset to approximate on-policy teacher quality. Fig. 3b confirms that smaller models yield weaker supervision: regardless of the prompt (Appendix A), average accuracy falls to 60%, and CoT traces are dropped 8.2% of the time.

### 2.4 Off-policy In-distribution distillation as Synthetic Self-distillation

To bridge this gap, SynthSelf is designed to **(1)** remain in-distribution with respect to the base model $\pi$, and **(2)** provide high-quality supervision. We achieve this via two core mechanisms: **(i)** utilizing off-policy distillation to enable rejection sampling, and **(ii)** selecting a larger aligner model that simultaneously **(ii.a.)** yields high-quality supervision, **(ii.b.)** stays in-distribution relative to the base model, and **(ii.c.)** incurs low computational overhead.

**(i) Off-policy supervision enables rejection sampling.** By generating offline datasets for SFT rather than training on-policy, we can filter for accuracy through rejection sampling. This step is essential because small models are highly sensitive to supervision quality.

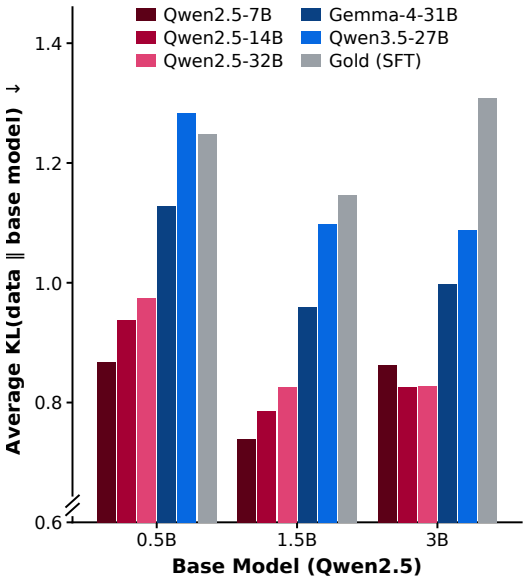



Figure 4: **Same-family rewrites provide more in-distribution off-policy data relative to Gold SFT.** Conversely, stronger, cross-family models fail to meaningfully improve distributional alignment with the base model.

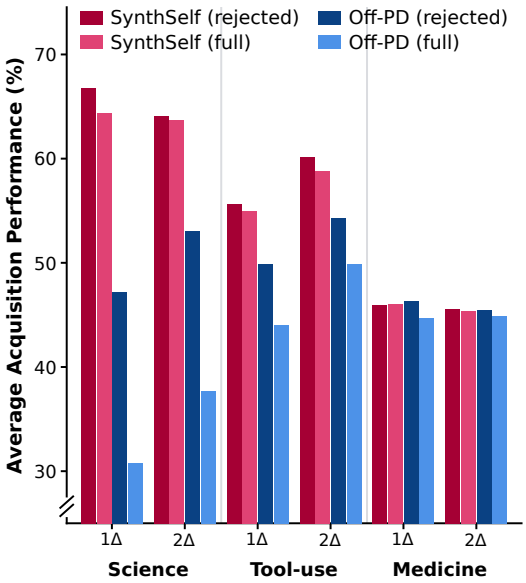



Figure 5: **An ablation study demonstrates that off-policy rejection sampling is a valuable component.** Furthermore, SYNTHSELF consistently outperforms standard off-policy distillation, even when controlling for model scale.

**(ii) Identifying a low-cost, competent in-distribution aligner.** We formalize the dual objectives of **(ii.a.)** generating high-quality supervision and **(ii.b.)** maintaining in-distribution alignment. Let $\mathcal{A} = \{\pi_A^{(1)}, \ldots, \pi_A^{(K)}\}$ be a pool of candidate aligners. We aim to select an aligner $\pi_A^\star \in \mathcal{A}$ that minimizes the distributional gap to the base model, subject to quality constraints (Eq. 6).

Given a demonstration pair $(x_i, y_i)$ and a rewrite prompt $P^{\text{rewrite}}$ (Box A), the aligner's rewrite distribution is:

$$q_i(y) \;\coloneqq\; \pi_A\big(y \mid P^{\text{rewrite}}(x_i, y_i)\big)\,. \tag{2}$$

This distribution translates the knowledge from $(x_i, y_i)$ into the aligner's native distribution. Appendix D shows an example: the rewrite keeps the gold answer and format while restating the reasoning in the aligner's own phrasing. We decode greedily, $\hat{y}_{i,t} = \arg\max_{v \in \mathcal{V}} \pi_A\big(v \mid P^{\text{rewrite}}(x_i, y_i), \hat{y}_{i,<t}\big)$, ensuring a deterministic rewrite $\hat{y}_i$. We measure supervision quality **(ii.a.)** directly on these rewrites:

$$R(\pi_A) \;=\; \mathbb{E}_i\big[r(x_i, \hat{y}_i)\big], \qquad \hat{y}_i \;=\; \text{greedy}(q_i), \tag{3}$$

where a verifiable reward $r(x_i, \hat{y}_i) = \mathbb{I}[a(\hat{y}_i) = a(y_i)]$ checks the extracted answer $a(\cdot)$ against the gold demonstration.

The student is trained to imitate the point mass $\delta_{\hat{y}_i}$ at the mode of $q_i$. The divergence against this point mass simplifies to the negative log-likelihood (NLL) under the student $\pi$, serving as our metric for in-distribution alignment **(ii.b.)**:

$$\mathcal{L}(\pi_A) \;=\; \mathbb{E}_i\big[D_{\text{KL}}\big(\delta_{\hat{y}_i} \,\|\, \pi(\cdot \mid x_i)\big)\big] \;=\; -\,\mathbb{E}_i \sum_{t=1}^{|\hat{y}_i|} \log \pi\big(\hat{y}_{i,t} \mid x_i, \hat{y}_{i,<t}\big). \tag{4}$$

Standard SFT on gold demonstrations follows an identical objective, replacing $\delta_{\hat{y}_i}$ with $\delta_{y_i}$ to provide a baseline divergence:

$$\mathcal{L}_{\text{gold}} \;=\; \mathbb{E}_i\big[D_{\text{KL}}\big(\delta_{y_i} \,\|\, \pi(\cdot \mid x_i)\big)\big] \;=\; -\,\mathbb{E}_i \sum_{t=1}^{|y_i|} \log \pi\big(y_{i,t} \mid x_i, y_{i,<t}\big). \tag{5}$$

**Algorithm 1:** SYNTHSELF (post-aligner selection)

**Input:** Demonstrations $\mathcal{D}_{\text{train}} = \{(x_i, y_i)\}_{i=1}^{N}$, post-trained base model $\pi_\theta$, aligner $\pi_A^\star$, rewrite instruction $P^{\text{rewrite}}$, answer extractor $a(\cdot)$

1 $\hat{\mathcal{D}} \leftarrow \emptyset$ ; // Initialize offline dataset
2 **foreach** $(x_i, y_i) \in \mathcal{D}_{train}$ **do**
/* **Synthetic In-distribution Rewrite** */
3 $\hat{y}_i \leftarrow \arg\max_y \ \pi_A\big(y \mid P^{\text{rewrite}}(x_i, y_i)\big)$ ; // Greedy decoding
/* **Rejection Sampling** */
4 **if** $a(\hat{y}_i) = a(y_i)$ **then**
5 $\hat{\mathcal{D}} \leftarrow \hat{\mathcal{D}} \cup \{(x_i, \hat{y}_i)\}$ ; // Same answer, but in-distribution
6 **else**
7 **discard** $(x_i, \hat{y}_i)$
8 **end**
9 **end**
/* **Off-policy learning via SFT** */
10 $\hat{\theta} \leftarrow \arg\min_\theta \ \mathbb{E}_{(x,\hat{y})\sim\hat{\mathcal{D}}}[-\log \pi_\theta(\hat{y} \mid x)]$ ;
11 **return** $\pi_{\hat{\theta}}$ ;

Consequently, we formulate aligner selection as:

$$\pi_A^\star \ = \ \arg\min_{\pi_A \in \mathcal{A}} \ \mathcal{L}(\pi_A) \qquad \text{s.t.} \qquad \mathcal{L}(\pi_A) \ < \ \mathcal{L}_{\text{gold}} \quad \text{and} \quad R(\pi_A) \ > \ R(\pi). \tag{6}$$

In short, we select the aligner closest in distribution to the student. The constraints guarantee its rewrites offer a distributional advantage over standard SFT ($\mathcal{L}(\pi_A) < \mathcal{L}_{\text{gold}}$) and strictly exceed the base model's supervision quality ($R(\pi_A) > R(\pi)$).

In practice, scaling up within the same model family perfectly satisfies Eq. 6. Fig. 4 shows that same-family aligners (`Qwen2.5-7B`, `14B`, `32B`) achieve the lowest average KL, falling well below $\mathcal{L}_{\text{gold}}$, whereas stronger cross-family models fail to meaningfully reduce the NLL (averaged across three domains detailed in § 3). Furthermore, same-family aligners yield higher data accuracy than the base model, satisfying $R(\pi_A) > R(\pi)$ (Tab. 2). Finally, to satisfy the low-cost constraint **(ii.c.)**, we instantiate $\pi_A^\star$ using minimally larger `Qwen2.5-7B` and `14B` models, ensuring minimal computational overhead.

**Algorithmic implementation.** Using this minimally larger same-family aligner $\pi_A$, we implement SYNTHSELF (see Alg. 1). We first initialize an empty offline dataset $\hat{\mathcal{D}}$ (Line 1). Next, processing the data in batches, we generate an in-distribution rewrite $\hat{y}_i$ for each sample $i$ via the aligner $\pi_A$ using greedy decoding (Lines 2 and 3). Samples that satisfy the rejection sampling are then added to $\hat{\mathcal{D}}$ (Lines 4 to 8). Finally, we perform SFT on this filtered dataset $\hat{\mathcal{D}}$ to update the base model parameters (Line 10).

## 3 EMPIRICAL STUDY

We evaluate SYNTHSELF on the three domains from SDFT (Shenfeld et al., 2026a), alongside a 4-subject science mixture to verify generalization. In §3.3, we further analyze our approach by: **(i)** demonstrating that distribution matching outweighs aligner capability; **(ii)** benchmarking against SDFT on matched and diverse prompts; and **(iii)** ablating rejection sampling to confirm the value of off-policy supervision.

### 3.1 EXPERIMENTAL PROTOCOL

**Baselines.** We compare our method against four baselines: standard SFT, SDFT (Shenfeld et al., 2026a), DFT (Wu et al., 2026), which modifies the objective function to imitate on-policy learning, naturally inducing in-distribution learning, and ASFT (Zhu et al., 2026), which improves upon DFT by further mitigating distributional shift. To control for the larger aligner model size, we also

| Model | | Baseline | | | | | SYNTHSELF (ours) | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Base | SFT | SDFT | DFT | ASFT | SS-1Δ | Δ Off-PD | Δ On-PD | SS-2Δ | Δ Off-PD | Δ On-PD |
| **SCIENCE: CHEMISTRY** | | | | | | | | | | | | |
| 0.5B | **Acq.** | 25.4 | 56.8 | 38.5 | 35.5 | 45.6 | 63.7 | +21.1 | +34.5 | 59.4 | +10.1 | +21.5 |
| | **Ret.** | 31.8 | 28.2 | 30.3 | 28.7 | 30.3 | 27.8 | -2.4 | +1.2 | 29.5 | -0.8 | +5.1 |
| 1.5B | **Acq.** | 31.0 | 64.9 | 53.6 | 56.8 | 48.5 | 67.1 | +16.4 | +33.3 | 65.3 | +12.0 | +22.9 |
| | **Ret.** | 43.9 | 44.1 | 41.8 | 43.8 | 45.2 | 45.0 | +0.4 | +1.9 | 44.6 | +0.0 | +2.5 |
| 3B | **Acq.** | 30.2 | 68.8 | 65.9 | 60.9 | 58.6 | 69.6 | +21.5 | +38.7 | 67.7 | +11.2 | +21.9 |
| | **Ret.** | 52.9 | 52.6 | 51.5 | 51.3 | 51.6 | 52.7 | -0.7 | +2.9 | 51.8 | -2.3 | -1.3 |
| **Avg.** | **Acq.** | 28.9 | 63.5 | 52.7 | 51.1 | 50.9 | 66.8 | +19.7 | +35.5 | 64.1 | +11.1 | +22.1 |
| | **Ret.** | 42.8 | 41.6 | 41.2 | 41.2 | 42.4 | 41.8 | -0.9 | +2.0 | 42.0 | -1.0 | +2.1 |
| **SCIENCE MIXTURE (4 SUBJECTS)** | | | | | | | | | | | | |
| 0.5B | **Acq.** | 23.0 | 48.7 | 25.7 | 32.8 | 35.5 | 49.8 | +7.2 | +20.0 | 43.0 | +0.4 | +9.6 |
| | **Ret.** | 31.8 | 30.0 | 29.5 | 29.1 | 29.7 | 30.8 | -0.7 | +2.4 | 28.8 | -2.5 | +1.8 |
| 1.5B | **Acq.** | 30.4 | 55.5 | 40.2 | 52.8 | 48.9 | 58.9 | +12.6 | +18.7 | 57.0 | +7.0 | +14.3 |
| | **Ret.** | 43.9 | 42.8 | 43.1 | 43.3 | 44.3 | 43.3 | -1.5 | -1.1 | 44.5 | -2.3 | +0.9 |
| 3B | **Acq.** | 43.8 | 61.1 | 60.6 | 50.9 | 55.1 | 63.2 | +7.2 | +17.4 | 67.7 | +10.4 | +16.6 |
| | **Ret.** | 52.9 | 52.9 | 51.5 | 50.7 | 51.6 | 51.6 | -2.5 | +1.4 | 52.9 | +0.0 | +2.0 |
| **Avg.** | **Acq.** | 32.4 | 55.1 | 42.2 | 45.5 | 46.5 | 57.3 | +9.0 | +18.7 | 55.9 | +6.0 | +13.5 |
| | **Ret.** | 42.8 | 41.9 | 41.3 | 41.0 | 41.9 | 41.9 | -1.6 | +0.9 | 42.1 | -1.6 | +1.5 |
| **TOOL-USE** | | | | | | | | | | | | |
| 0.5B | **Acq.** | 23.7 | 40.2 | 24.7 | 30.9 | 42.3 | 43.3 | -1.0 | +7.2 | 56.7 | +10.3 | +6.2 |
| | **Ret.** | 31.8 | 25.9 | 30.5 | 26.0 | 29.0 | 28.3 | -1.0 | +0.5 | 29.9 | -0.5 | +3.2 |
| 1.5B | **Acq.** | 43.3 | 57.7 | 51.5 | 59.8 | 54.6 | 61.9 | +11.3 | +15.5 | 55.7 | -2.1 | -6.2 |
| | **Ret.** | 43.9 | 39.9 | 42.6 | 41.6 | 41.6 | 42.3 | -1.1 | -3.0 | 42.4 | -1.4 | +0.3 |
| 3B | **Acq.** | 33.0 | 66.0 | 57.7 | 47.4 | 58.8 | 61.9 | +7.2 | +13.4 | 68.0 | +9.3 | +9.3 |
| | **Ret.** | 52.9 | 50.9 | 52.2 | 49.6 | 53.2 | 51.0 | +0.5 | +0.9 | 51.4 | +0.4 | +2.1 |
| **Avg.** | **Acq.** | 33.3 | 54.6 | 44.7 | 46.0 | 51.9 | 55.7 | +5.8 | +12.0 | 60.1 | +5.8 | +3.1 |
| | **Ret.** | 42.8 | 38.9 | 41.8 | 39.1 | 41.3 | 40.5 | -0.5 | -0.5 | 41.2 | -0.5 | +1.8 |
| **MEDICINE** | | | | | | | | | | | | |
| 0.5B | **Acq.** | 26.7 | 16.0 | 26.7 | 2.0 | 18.2 | 32.7 | -3.0 | -1.5 | 34.2 | -0.5 | -2.0 |
| | **Ret.** | 31.8 | 28.4 | 30.5 | 28.0 | 29.4 | 31.0 | -0.8 | +3.7 | 29.9 | -2.4 | +2.1 |
| 1.5B | **Acq.** | 42.8 | 41.3 | 28.8 | 27.7 | 36.7 | 47.8 | +0.5 | -0.2 | 48.3 | -0.3 | -3.0 |
| | **Ret.** | 43.9 | 41.4 | 42.5 | 40.8 | 42.8 | 43.7 | -2.8 | +4.5 | 43.3 | -2.6 | +2.2 |
| 3B | **Acq.** | 57.5 | 48.8 | 54.8 | 20.7 | 50.2 | 57.3 | +1.3 | +5.2 | 54.0 | +1.0 | +0.7 |
| | **Ret.** | 52.9 | 48.7 | 52.2 | 51.7 | 52.8 | 52.3 | -3.0 | +1.4 | 52.8 | -0.6 | -0.6 |
| **Avg.** | **Acq.** | 42.3 | 35.4 | 36.8 | 16.8 | 35.0 | 45.9 | -0.4 | +1.2 | 45.5 | +0.1 | -1.4 |
| | **Ret.** | 42.8 | 39.5 | 41.7 | 40.2 | 41.7 | 42.4 | -2.2 | +3.2 | 42.0 | -1.9 | +1.2 |
| **AVERAGE** | | | | | | | | | | | | |
| **Avg.** | **Acq.** | 34.2 | 52.2 | 44.1 | 39.8 | 46.1 | 56.4 | +8.5 | +16.9 | 56.4 | +5.7 | +9.3 |
| | **Ret.** | 42.8 | 40.5 | 41.5 | 40.4 | 41.8 | 41.7 | -1.3 | +1.4 | 41.8 | -1.2 | +1.7 |

Table 1: **SYNTHSELF leads acquisition (Acq.) in all 4 benchmarks and ties the best baseline on 14 retention tasks (Ret.), over 4 baselines, off-/on-policy distillation, and three base models.** Green and red background shading on Avg. rows is min-max normalized across each row. The 14 retention tasks span instruction following, commonsense, diverse knowledge, code, and math. The mixture contains Chemistry, Biology, Materials, and Physics in both the train and test splits. Δ Off-PD and Δ On-PD are SYNTHSELF minus off- or on-policy distillation (%p).

compare SYNTHSELF against off- and on-policy distillation (Agarwal et al., 2024). Here, off-policy distillation is equivalent to Alg. 1, but conditions the generation on the original query $x_i$ rather than the rewrite prompt $P^{\text{rewrite}}(x_i, y_i)$. We tune standard SFT minimally and fix these hyperparameters across all methods for fairness. To ensure a strong baseline, we additionally optimize SDFT's skip-token hyperparameter; which discards $T$ initial formatting tokens. Conversely, SYNTHSELF is tuning-free and introduces zero additional hyperparameters.

**Datasets and models.** Following Shenfeld et al. (2026a), we evaluate on three domains: science (SciKnowEval Chemistry L-3; Feng et al. (2025)), tool-use (ToolAlpaca; Tang et al. (2023)),

| Aligner | Acquisition ↑ SCI. | Acquisition ↑ TOOL | KL (NLL) ↓ SCI. | KL (NLL) ↓ TOOL | Data-gen Accuracy (%) ↑ SCI. | Data-gen Accuracy (%) ↑ TOOL | Data-gen PFLOPs ↓ SCI. | Data-gen PFLOPs ↓ TOOL |
|---|---|---|---|---|---|---|---|---|
| **More In-distribution Aligners** | | | | | | | | |
| `Qwen2.5-7B` | **66.8** | 55.7 | **0.995** | **0.650** | 99.0 | 84.1 | **13.9** | **39.0** |
| `Qwen2.5-14B` | 64.1 | **60.1** | 1.033 | 0.666 | 99.1 | 81.8 | 36.6 | 76.0 |
| **Stronger but More Out-of-distribution Aligners** | | | | | | | | |
| `Qwen3.5-27B` | 65.1 | 50.5 | 1.440 | 0.873 | 90.3 | 95.8 | 70.0 | 142.1 |
| `Gemma-4-31B` | 60.8 | 56.0 | 1.245 | 0.812 | **99.8** | **99.3** | 148.2 | 167.9 |
| **Average** | | | | | | | | |
| ID | 65.5 | 57.9 | 1.014 | 0.658 | 99.0 | 82.9 | 25.2 | 57.5 |
| OOD | 63.0 | 53.3 | 1.342 | 0.843 | 95.0 | 97.5 | 109.1 | 155.0 |
| $\Delta$ = ID − OOD | +2.5 | +4.6 | -0.328 | -0.185 | +4.0 | -14.6 | -83.8 | -97.5 |

Table 2: **SYNTHSELF performs better with in-distribution aligners.** Averaged over the `Qwen2.5-0.5B`, `1.5B`, and `3B` base models, using aligners that are more in-distribution (measured by NLL) is a stronger predictor of performance than the strength of the aligner (measured by data-generation accuracy). Data-generation accuracy represents the reward $R(\pi_A)$ in Eq. 3. It is additionally advantageous that smaller, same-family aligners are the most compute-efficient when generating off-policy data. **Bold** indicates best value across each column.

and medicine (MedMCQA; Pal et al. (2022)). We expand to all four SciKnowEval subjects for our mixture experiment. We substitute the medical dataset used by Shenfeld et al. (2026a) with MedMCQA to avoid LLM-as-a-judge evaluation and improve reproducibility. We adopt the `Qwen2.5-Instruct` family as base models, unlike later versions like `Qwen3` (Yang et al., 2025) exhaustively trained on open-source data; this choice reduces contamination and ensures demonstration novelty. Further experimental details are in Appendix C.

### 3.2 RESULTS

SYNTHSELF achieves the highest acquisition across all four benchmarks, including the science mixture, while matching the strongest baselines on 14 retention tasks (Tab. 1, Fig. 1). We abbreviate SYNTHSELF as SS, with 1$\Delta$ and 2$\Delta$ denoting aligners one and two sizes larger, respectively (i.e., `Qwen2.5-7B` and `Qwen2.5-14B`). Interestingly, recent methods fail to surpass SFT in acquisition at these smaller scales. DFT's performance in the medical domain deteriorates especially significantly; we have double-checked that these evaluation values are correct. However, consistent with the discussion in § 2.2.1, the on-policy and in-distribution mechanisms of SDFT and ASFT successfully improve retention over SFT.

Moreover, we validate that the performance gains of SYNTHSELF stem primarily from in-distribution demonstration injection, rather than merely leveraging a larger aligner (Tab. 1). Off- and on-policy distillation (Off-/On-PD) use the identical aligner, differing only in conditioning on $x_i$ rather than on $P^{\text{rewrite}}(x_i, y_i)$. Averaged across all four settings, SYNTHSELF nonetheless improves acquisition by +5.7%p to +16.9%p while maintaining comparable retention. As expected, demonstration injection drives these acquisition gains, whereas all three methods (SYNTHSELF, Off-PD, and On-PD) similarly preserve the base model distribution to sustain retention capabilities. Lastly, we observe that SYNTHSELF does not yield significant improvements over Off-/On-PD in the medical domain. This is likely due to the aligner models already having perfect knowledge of the training data, making demonstration injection largely redundant. Despite this, SYNTHSELF remains highly competitive, successfully outperforming all remaining baselines.

### 3.3 FURTHER EXPERIMENTAL ANALYSES

**Distribution matching outweighs aligner capability, and leads to data generation cost savings.** We validate our hypothesis that lower-cost, in-distribution aligners are more effective than stronger out-of-distribution (OOD) alternatives. Although larger aligners like `Qwen3.5-27B` and `Gemma-4-31B` (Team et al., 2026) generate higher-quality supervision (measured via data-

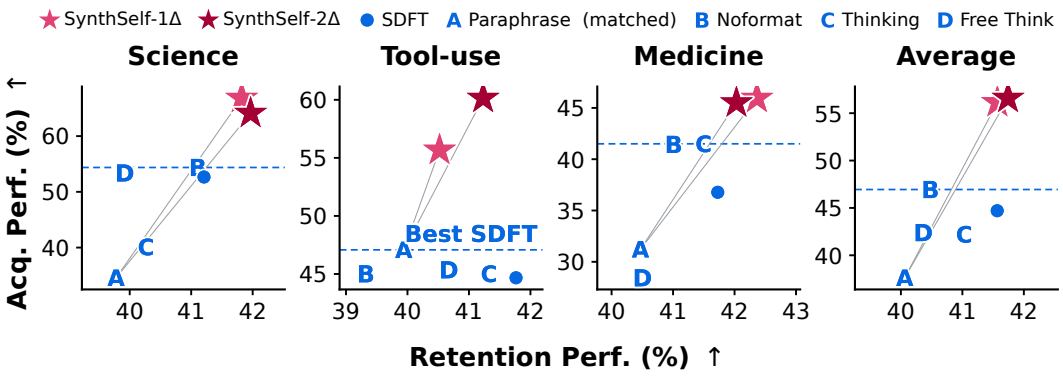



Figure 6: **SYNTHSELF outperforms SDFT across diverse prompt variations, including a version prompt-matched to SYNTHSELF.** On average, SYNTHSELF outperforms prompt-matched SDFT by +18.96%p and +1.68%p on acquisition and retention, respectively.

generation accuracy), they introduce severe distributional shift relative to the base model. Consequently, they yield similar or worse acquisition performance (Tab. 2). Moreover, using a same-family aligner reduces data generation inference costs by 62.9% to 76.9% FLOPs.

**SYNTHSELF outperforms prompt-matched self-distillation.** We confirm that SYNTHSELF's gains over SDFT do not arise from differences in supervision prompts (Fig. 6). Evaluating four SDFT prompt variants, including one matched to SYNTHSELF, we find our method consistently achieves higher acquisition across all three domains and higher average retention than every variant. Full prompts are provided in Appendix A.

**Rejection sampling ablation.** Finally, we ablate the offline data generation that enables rejection sampling (Alg. 1, Lines 4–7). As shown in Fig. 5, this filtering mechanism consistently improves both SYNTHSELF and Off-PD. We hypothesize such filtering is especially critical for smaller base models, which are highly sensitive to noisy supervision.

## 4 DISCUSSION

**Additional related work.** Huang et al. (2026) highlights the impact of distribution drift in SFT using code-generation data on `Qwen3-8B`. However, they do not directly address continual learning or small language models, restricting their evaluation to limited scales and domains. Furthermore, Li et al. (2026) combines mechanisms that we instead isolate: SDFT-style self-rewrites with similarity filtering, cross-family fusion, and an on-policy KL regularizer; covered individually by our SDFT, on-policy distillation, and ASFT baselines. Crucially, we differ in the supervision source. Because their rewrites originate from the student, they inherit the small-scale degradation shown in Fig. 3. Our same-family aligner explicitly resolves this limitation.

**Training compute savings and further performance headroom.** Rejection sampling reduces training compute by shrinking the dataset $|\hat{D}|$ (e.g., by 17.1% for tool use; Tab. 2). Despite training on fewer demonstrations than SFT, SYNTHSELF improves acquisition by +4.2%p on average, proving these gains stem from supervision quality rather than quantity. To further boost performance, practitioners can scale $k$ in pass@$k$, trading inference compute for larger coverage supervision.

**Limitation and future directions.** SYNTHSELF requires a larger, same-family aligner. Fortunately, most major open-weight families provide a spectrum of model sizes, naturally satisfying this requirement since smaller variants are often distilled from their larger counterparts (e.g., `Qwen2.5`, `Qwen3.5`, `OLMo 2` (Walsh et al., 2025), `Llama 3` (Grattafiori et al., 2024), and `Gemma 4`). Nevertheless, future work could explore adapting out-of-family aligners to produce in-distribution supervision. One promising direction is adding the base model's answer in-context with an appropriate prompt to generate an in-distribution rewrite.

### AI USE STATEMENT

We used generative AI to assist with experimental code generation and reviewing writing for grammar and spelling. Furthermore, language models generated the reasoning demonstrations for three SciKnowEval subjects. The Chemistry demonstrations were generated with `GPT-4o` by Shenfeld et al. (2026a). AI was not used for conceptualization, methodology design, data analysis, or result interpretation. All reported numbers were produced directly by our experimental pipeline, not by a language model. The authors reviewed all AI-assisted scripts and text, and we take full responsibility for the final content of this work.

## REFERENCES

Marah I Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Hassan Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Harkirat Behl, Alon Benhaim, Misha Bilenko, Johan Bjorck, Sébastien Bubeck, Martin Cai, Caio César Teodoro Mendes, Weizhu Chen, Vishrav Chaudhary, Parul Chopra, Allie Del Giorno, Gustavo de Rosa, Matthew Dixon, Ronen Eldan, Dan Iter, Abhishek Goswami, Suriya Gunasekar, Emman Haider, Junheng Hao, Russell J. Hewett, Jamie Huynh, Mojan Javaheripi, Xin Jin, Piero Kauffmann, Nikos Karampatziakis, Dongwoo Kim, Mahmoud Khademi, Lev Kurilenko, James R. Lee, Yin Tat Lee, Yuanzhi Li, Chen Liang, Weishung Liu, Xihui (Eric) Lin, Zeqi Lin, Piyush Madan, Arindam Mitra, Hardik Modi, Anh Nguyen, Brandon Norick, Barun Patra, Daniel Perez-Becker, Thomas Portet, Reid Pryzant, Heyang Qin, Marko Radmilac, Corby Rosset, Sambudha Roy, Olli Saarikivi, Amin Saied, Adil Salim, Michael Santacroce, Shital Shah, Ning Shang, Hiteshi Sharma, Xia Song, Olatunji Ruwase, Xin Wang, Rachel Ward, Guanhua Wang, Philipp Witte, Michael Wyatt, Can Xu, Jiahang Xu, Weijian Xu, Sonali Yadav, Fan Yang, Ziyi Yang, Donghan Yu, Chengruidong Zhang, Cyril Zhang, Jianwen Zhang, Li Lyna Zhang, Yi Zhang, Yunan Zhang, and Xiren Zhou. Phi-3 technical report: A highly capable language model locally on your phone. Technical Report MSR-TR-2024-12, Microsoft, August 2024. URL https://www.microsoft.com/en-us/research/publication/phi-3-technical-report-a-highly-capable-language-model-locally-on-your-phone/.

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos Garea, Matthieu Geist, and Olivier Bachem. On-policy distillation of language models: Learning from self-generated mistakes. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=3zKtaqxLhW.

Peter Belcak, Greg Heinrich, Shizhe Diao, Yonggan Fu, Xin Dong, Saurav Muralidharan, Yingyan Celine Lin, and Pavlo Molchanov. Small language models are the future of agentic ai, 2025. URL https://arxiv.org/abs/2506.02153.

Qinnan Cai, Yibo Zhao, and Xiang Li. Think big, search small: Where capacity matters in hierarchical search agents? *arXiv preprint arXiv:2607.07548*, 2026.

Charlie Chen, Sebastian Borgeaud, Geoffrey Irving, Jean-Baptiste Lespiau, Laurent Sifre, and John Jumper. Accelerating large language model decoding with speculative sampling, 2023. URL https://arxiv.org/abs/2302.01318.

Hongyang Chen, Zhongwu Sun, Hongfei Ye, Kunchi Li, and Xuemin Lin. Beyond static models: An evolving framework for continual learning in large language models across training stages, 2026a. URL https://arxiv.org/abs/2603.12658.

Howard Chen, Noam Razin, Karthik R Narasimhan, and Danqi Chen. Retaining by doing: The role of on-policy data in mitigating forgetting. In *Forty-third International Conference on Machine Learning*, 2026b. URL https://openreview.net/forum?id=ODTM64azGa.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, Nick Ryder, Mikhail Pavlov, Alethea Power, Lukasz Kaiser, Mohammad Bavarian, Clemens Winter, Philippe Tillet, Felipe Petroski Such, Dave Cummings, Matthias Plappert, Fotios Chantzis, Elizabeth Barnes, Ariel Herbert-Voss, William Hebgen Guss, Alex Nichol, Alex

Paino, Nikolas Tezak, Jie Tang, Igor Babuschkin, Suchir Balaji, Shantanu Jain, William Saunders, Christopher Hesse, Andrew N. Carr, Jan Leike, Josh Achiam, Vedant Misra, Evan Morikawa, Alec Radford, Matthew Knight, Miles Brundage, Mira Murati, Katie Mayer, Peter Welinder, Bob McGrew, Dario Amodei, Sam McCandlish, Ilya Sutskever, and Wojciech Zaremba. Evaluating large language models trained on code. 2021.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*, 2018.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems, 2021. URL https://arxiv.org/abs/2110.14168.

Hugging Face. Open r1: A fully open reproduction of deepseek-r1, January 2025. URL https://github.com/huggingface/open-r1.

Kehua Feng, Xinyi Shen, Weijie Wang, Xiang Zhuang, Yuqi Tang, Qiang Zhang, and Keyan Ding. Sciknoweval: Evaluating multi-level scientific knowledge of large language models, 2025. URL https://arxiv.org/abs/2406.09098.

Robert M French. Catastrophic forgetting in connectionist networks. *Trends in cognitive sciences*, 3(4):128–135, 1999.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei

Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

Jackson Harmon, Andreas Hochlehnert, Matthias Bethge, and Ameya Prabhu. Mapping post-training forgetting in language models at scale. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=qCIg2WGudx.

Zixian Huang, Kaichen Yang, Xu Huang, Feiyang Hao, Qiming Ge, Bowen Li, He Du, Kai Chen, and Qipeng Guo. How to fine-tune a reasoning model? a teacher–student cooperation framework to synthesize student-consistent SFT data. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=tylCinzjUF.

Damjan Kalajdzievski. Scaling laws for forgetting when fine-tuning large language models, 2024. URL https://arxiv.org/abs/2401.05605.

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. doi: 10.1073/pnas.1611835114. URL https://www.pnas.org/doi/abs/10.1073/pnas.1611835114.

Song Lai, Haohan Zhao, Rong Feng, Changyi Ma, Wenzhuo Liu, Hongbo Zhao, Xi Lin, Dong Yi, Qingfu Zhang, Hongbin Liu, Gaofeng Meng, and Fei Zhu. Reinforcement fine-tuning naturally mitigates forgetting in continual post-training, 2026. URL https://arxiv.org/abs/2507.05386.

Yuduo Li, Xiaofeng Shi, Qian Kou, Longbin Yu, and Hua Zhou. Raft: Data refinement and adaptive distillation for domain fine-tuning with alleviated forgetting, 2026. URL https://arxiv.org/abs/2606.00147.

Zechun Liu, Changsheng Zhao, Forrest Iandola, Chen Lai, Yuandong Tian, Igor Fedorov, Yunyang Xiong, Ernie Chang, Yangyang Shi, Raghuraman Krishnamoorthi, Liangzhen Lai, and Vikas Chandra. MobileLLM: Optimizing sub-billion parameter language models for on-device use cases. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 32431–32454. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/liu24ce.html.

Kevin Lu and Thinking Machines Lab. On-policy distillation. *Thinking Machines Lab: Connectionism*, 2025. doi: 10.64434/tml.20251026. https://thinkingmachines.ai/blog/on-policy-distillation.

Zhenyan Lu, Xiang Li, Dongqi Cai, Rongjie Yi, Fangming Liu, Xiwen Zhang, Nicholas D. Lane, and Mengwei Xu. Small language models: Survey, measurements, and insights, 2025. URL https://arxiv.org/abs/2409.15790.

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, 2025.

Michael McCloskey and Neal J. Cohen. Catastrophic interference in connectionist networks: The sequential learning problem. volume 24 of *Psychology of Learning and Motivation*, pp. 109–165. Academic Press, 1989. doi: https://doi.org/10.1016/S0079-7421(08)60536-8. URL https://www.sciencedirect.com/science/article/pii/S0079742108605368.

Augustus Odena, Charles Sutton, David Martin Dohan, Ellen Jiang, Henryk Michalewski, Jacob Austin, Maarten Paul Bosma, Maxwell Nye, Michael Terry, and Quoc V. Le. Program synthesis with large language models. In *n/a*, pp. n/a, n/a, 2021. n/a.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In Gerardo Flores, George H Chen, Tom Pollard, Joyce C Ho, and Tristan Naumann (eds.), *Proceedings of the Conference on Health, Inference, and Learning*, volume 174 of *Proceedings of Machine Learning Research*, pp. 248–260. PMLR, 07–08 Apr 2022. URL https://proceedings.mlr.press/v174/pal22a.html.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report, 2025. URL https://arxiv.org/abs/2412.15115.

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*, 2024. URL https://openreview.net/forum?id=Ti67584b98.

Stephane Ross, Geoffrey Gordon, and Drew Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. In Geoffrey Gordon, David Dunson, and Miroslav Dudík (eds.), *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, volume 15 of *Proceedings of Machine Learning Research*, pp. 627–635, Fort Lauderdale, FL, USA, 11–13 Apr 2011. PMLR. URL https://proceedings.mlr.press/v15/ross11a.html.

Idan Shenfeld, Mehul Damani, Jonas Hübotter, and Pulkit Agrawal. Self-distillation enables continual learning. In *Forty-third International Conference on Machine Learning*, 2026a. URL https://openreview.net/forum?id=qA6FgH0nnZ.

Idan Shenfeld, Jyothish Pari, and Pulkit Agrawal. RL's razor: Why online reinforcement learning forgets less. In *The Fourteenth International Conference on Learning Representations*, 2026b. URL https://openreview.net/forum?id=7HNRYT4V44.

Haizhou Shi, Zihao Xu, Hengyi Wang, Weiyi Qin, Wenyuan Wang, Yibin Wang, Zifeng Wang, Sayna Ebrahimi, and Hao Wang. Continual learning of large language models: A comprehensive survey. *arXiv preprint arXiv:2404.16789*, 2024.

Avi Singh, John D Co-Reyes, Rishabh Agarwal, Ankesh Anand, Piyush Patil, Xavier Garcia, Peter J Liu, James Harrison, Jaehoon Lee, Kelvin Xu, Aaron T Parisi, Abhishek Kumar, Alexander A Alemi, Alex Rizkowsky, Azade Nova, Ben Adlam, Bernd Bohnet, Gamaleldin Fathy Elsayed, Hanie Sedghi, Igor Mordatch, Isabelle Simpson, Izzeddin Gur, Jasper Snoek, Jeffrey Pennington, Jiri Hron, Kathleen Kenealy, Kevin Swersky, Kshiteej Mahajan, Laura A Culp, Lechao Xiao, Maxwell Bileschi, Noah Constant, Roman Novak, Rosanne Liu, Tris Warkentin, Yamini Bansal, Ethan Dyer, Behnam Neyshabur, Jascha Sohl-Dickstein, and Noah Fiedel. Beyond human data: Scaling self-training for problem-solving with language models. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=lNAyUngGFK. Expert Certification.

P. K. Srijith, Shrey Satapara, and Sarath Chandar. Continual learning in large language models: Foundations to frontiers. In Benjamin Heinzerling and Lun-Wei Ku (eds.), *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics: Tutorial Abstract*, pp. 6–17, Mumbai, India, December 2025. Association for Computational Linguistics. ISBN 979-8-89176-302-9. doi: 10.18653/v1/2025.ijcnlp-tutorials.2. URL https://aclanthology.org/2025.ijcnlp-tutorials.2/.

Qiaoyu Tang, Ziliang Deng, Hongyu Lin, Xianpei Han, Qiao Liang, Boxi Cao, and Le Sun. Toolalpaca: Generalized tool learning for language models with 3000 simulated cases, 2023. URL https://arxiv.org/abs/2306.05301.

Gemma Team, Sherif El Abd, Vaibhav Aggarwal, Robin Algayres, Alek Andreev, Olivier Bachem, Ian Ballantyne, Cormac Brick, Victor Cărbune, Michelle Casbon, Mayank Chaturvedi, Aditya Chawla, Victor Cotruta, Alice Coucke, Phil Culliton, Robert Dadashi, Lucas Dixon, Mohamed Elhawaty, Utku Evci, Clément Farabet, Johan Ferret, Filippo Galgani, Sertan Girgin, Jean-Bastien Grill, Maarten Grootendorst, Jiaxian Guo, Cassidy Hardin, Yanzhang He, Steven M. Hernandez, Omri Homburger, Léonard Hussenot, Juyeong Ji, Armand Joulin, Aishwarya Kamath, Parnian Kassraie, Olivier Lacombe, Preethi Lahoti, Gaël Liu, Gus Martins, Luciano Martins, Tatiana

Matejovicova, Ramona Merhej, Nikola Momchev, Sneha Mondal, Ryan Mullins, Sindhu Raghuram Panyam, Shreya Pathak, Sarah Perrin, André Susano Pinto, Etienne Pot, Angéline Pouget, Alexandre Ramé, Sabela Ramos, Douglas Reid, David Rim, Morgane Rivière, Karsten Roth, Louis Rouillard, Omar Sanseviero, Pier Giuseppe Sessa, Shane Settle, Danila Sinopalnikov, Sara Smoot, Piotr Stanczyk, Andreas Steiner, Lawrence Stewart, Ilya Tolstikhin, Michael Tschannen, Anton Tsitsulin, Nino Vieillard, Renjie Wu, Pingmei Xu, Haichuan Yang, Edouard Yvinec, Biao Zhang, Li Zhang, Joe Zou, Nicolas Aagnes, Abdelrahman Abdelhamed, Jakub Adamek, Shivani Agrawal, Shubham Agrawal, Ibrahim Alabdulmohsin, Jean Baptiste Alayrac, Uri Alon, Chandramouli Amarnath, Ankesh Anand, Chrysovalantis Anastasiou, Setareh Ariafar, François-Xavier Aubet, Kyriakos Axiotis, Federico Barbero, Joelle Barral, Alexei Bendebury, Urs Bergmann, Stanley Bileschi, Kat Black, Mathieu Blondel, Sebastian Borgeaud, Arthur Bražinskas, Ryan Burnell, Robert Busa-Fekete, Mu Cai, Daniele Calandriello, Glenn Cameron, Charlotte Caucheteux, Rahma Chaabouni, Garima Chadha, Jetha Chan, Blake Jianhang Chen, Jesse Chen, Lin Chen, Xu Chen, Derek Cheng, Tzu hsiang Chien, Nikolai Chinaev, Yi Chou, Zhaohui Chu, Benjamin Coleman, Pooja Consul, Sam Conway-Rahman, Scott Crowell, Dylan Cutler, Vivek Dani, Samira Daruki, Anil Das, Daniel Deutsch, Nishanth Dikkala, Li Ding, Qiuhan Ding, Shenil Dodhia, Konstantin Donhauser, Tulsee Doshi, Anca Dragan, Alex Druinsky, Sahil Dua, Zoltan Egyed, Danielle Eisenbud, Daniel Eppens, Cindy Fan, Bahare Fatemi, Yassir Fathullah, Vlad Feinberg, Milen Ferev, Sebastian Flennerhag, Takumi Fujimoto, João Gabriel Oliveira, Isaac Galatzer-Levy, João Gante, Simon Geisler, Soham Ghosal, Antonious M. Girgis, Tamara von Glehn, Alec Go, Alhaad Gokhale, Alex Grills, Yiming Gu, Mayank Gupta, Pramod Gupta, Guru Guruganesh, Raia Hadsell, Hamza Harkous, Jitendra Harlalka, Demis Hassabis, Anja Hauth, Joe Heyward, Arian Hosseini, Chih-Yang Hsia, I-Hung Hsu, Xiaopeng Huang, Yangsibo Huang, Kevin Hui, Adrian Hutter, Te I, Fotis Iliopoulos, Advait Jain, Ganesh Jawahar, Ziwei Ji, Qilin Jin, Melvin Johnson, Kandarp Joshi, Arun Kandoor, Wang-Cheng Kang, Koray Kavukcuoglu, Mehran Kazemi, Kathleen Kenealy, Amr Khalifa, Phoebe Kirk, Ivan Korotkov, Suraj Kothawade, Vitaly Kovalev, Neel Kovelamudi, Adam Kraft, Ravin Kumar, Vivek Kumar, Harish Kuppam, Justin Lannin, Chen-Yu Lee, Seungji Lee, Dmitry Lepikhin, Alon Levkovitch, Dongdong Li, Qiujia Li, Valentin Liévin, Ethan Lin, Ziqian Lin, Casper Liu, Tianlin Liu, Tianqi Liu, Xin Liu, Ivan Lobov, Mayank Lunayach, Min Ma, Gagan Madan, Andrii Maksai, Eric Malmi, Michal Matuszak, Daniel McDuff, Gaurav Menghani, Maciej Mikuła, Daniil Mirylenka, Karolis Misiunas, Vedant Misra, Andreea Mitran, Kareem Mohamed, Maksim Mukha, Eric Noland, James O'Donnell, Brendan O'Donoghue, Kate Olszewska, Bernett Orlando, Wanqiong Pan, Rina Panigrahy, Unnati Parekh, Nicolas Perez-Nieves, Chunjong Park, Eric Paskie, Liqian Peng, Bryce Petrini, Slav Petrov, Jonas Pfeiffer, Bilal Piot, Martyna Plomecka, Siim Poder, Octavio Ponce, Arijit Pramanik, David Racz, Anish Rajan, Michelle Ramanovich, Anand Rao, Marvin Ritter, Vitor Rodrigues, Evan Rosen, Mikołaj Rybiński, Noveen Sachdeva, Michaël E. Sander, Rohit Sathyanarayana, Sagar Savla, Samuel Schmidgall, Tal Schuster, George Scrivener, Benoit Seguin, Andrew Sellergren, Aliaksei Severyn, Izhak Shafran, Dhruv Shah, Bobak Shahriari, Yuan Shangguan, Ashish Shenoy, Pradeep Shenoy, Rakesh Shivanna, Pauline Sho, Lucas Spangher, Wojciech Stokowiec, Tim Strother, Yao Su, Yinghao Sun, Mukund Sundararajan, Andrea Tacchetti, Mor Hazan Taege, Pouya Tafti, Jean Tarbouriech, Chetan Tekur, Shantanu Thakoor, Rahul Thapa, Madeleine Traverse, Lenart Treven, Tao Tu, Chien Te Tung, Çağlar Ünlü, Petar Veličković, Malini Pooni Venkat, Sagar Gubbi Venkatesh, Vidya Venkiteswaran, Francesco Visin, Alex Vitvitskyi, Kiran Vodrahalli, Weiyi Wang, Xin Wang, Tris Warkentin, Jan Wassenberg, John Wieting, Cindy Wu, Lechao Xiao, Hao Xu, Yuhui Xu, Fuzhao Xue, Arun Yadav, Jun Yan, Antoine Yang, Lin Yang, Ming-Hsuan Yang, Ziyu Ying, Jae Hyeon Yoo, Morteza Zadimoghaddam, Sajjad Zafar, Fred Zhang, Jiageng Zhang, Jianyi Zhang, Xiaofan Zhang, Chao Zhao, David Zhou, and Chen Zou. Gemma 4 technical report, 2026. URL https://arxiv.org/abs/2607.02770.

Evan Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Shane Arora, Akshita Bhagia, Yuling Gu, Shengyi Huang, Matt Jordan, Nathan Lambert, Dustin Schwenk, Oyvind Tafjord, Taira Anderson, David Atkinson, Faeze Brahman, Christopher Clark, Pradeep Dasigi, Nouha Dziri, Allyson Ettinger, Michal Guerquin, David Heineman, Hamish Ivison, Pang Wei Koh, Jiacheng Liu, Saumya Malik, William Merrill, Lester James Validad Miranda, Jacob Morrison, Tyler Murray, Crystal Nam, Jake Poznanski, Valentina Pyatkin, Aman Rangapur, Michael Schmitz, Sam Skjonsberg, David Wadden, Christopher Wilhelm, Michael Wilson, Luke Zettlemoyer, Ali Farhadi, Noah A. Smith, and Hannaneh Hajishirzi. 2 OLMo 2 furious (COLM's version). In *Second Conference on Language Modeling*, 2025. URL https://openreview.net/forum?

id=2ezugTT9kU.

Yifan Wang, Yafei Liu, Chufan Shi, Haoling Li, Chen Chen, Haonan Lu, and Yujiu Yang. InsCL: A data-efficient continual learning paradigm for fine-tuning large language models with instructions. In Kevin Duh, Helena Gomez, and Steven Bethard (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 663–677, Mexico City, Mexico, June 2024a. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.37. URL https://aclanthology.org/2024.naacl-long.37/.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, Tianle Li, Max Ku, Kai Wang, Alex Zhuang, Rongqi Fan, Xiang Yue, and Wenhu Chen. MMLU-pro: A more robust and challenging multi-task language understanding benchmark. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024b. URL https://openreview.net/forum?id=y10DM6R2r3.

Yongliang Wu, Yizhou Zhou, Zhou Ziheng, Yingzhe Peng, Xinyu Ye, Xinting Hu, Wenbo Zhu, Lu Qi, Ming-Hsuan Yang, and Xu Yang. On the generalization of SFT: A reinforcement learning perspective with reward rectification. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=Lv7PjbcaMi.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

Junhao Zheng, Xidi Cai, Shengjie Qiu, and Qianli Ma. Spurious forgetting in continual learning of language models. In *The Thirteenth International Conference on Learning Representations*, 2025a. URL https://openreview.net/forum?id=ScI7IlKGdI.

Junhao Zheng, Shengjie Qiu, Chengming Shi, and Qianli Ma. Towards lifelong learning of large language models: A survey. *ACM Comput. Surv.*, 57(8), March 2025b. ISSN 0360-0300. doi: 10.1145/3716629. URL https://doi.org/10.1145/3716629.

He Zhu, Junyou Su, Peng Lai, Ren Ma, Wenjia Zhang, Linyi Yang, and Guanhua Chen. Anchored supervised fine-tuning. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=PORko7QT64.

# APPENDIX

## A PROMPTS

Prompt A is $P^{\text{rewrite}}$, the instruction behind every trained SYNTHSELF cell. Prompts B, C and D are the rewrite variants used in the self-rewrite study (§ 2.2.1) and the prompt-matched SDFT ablation (§ 3.3). {q} and {gold} are the question and the gold response.

**Prompt A: Paraphrase (matched) = $P^{\text{rewrite}}$**

```
Rewrite the following response in your own words.  Keep the SAME
final answer and the SAME format exactly.
Question:
{q}
Response to rewrite:
{gold}
```

**Prompt B: No Format**

```
Rewrite the following response in your own words.  Keep the SAME
final answer.
Question:
{q}
Response to rewrite:
{gold}
End with exactly this format:  <reasoning>...</reasoning>
<answer>LETTER</answer>
```

**Prompt C: Thinking**

```
Rewrite the following response in your own words, including the
thinking process.  Keep the SAME final answer.
Question:
{q}
Response to rewrite:
{gold}
End with exactly this format:  <reasoning>...</reasoning>
<answer>LETTER</answer>
```

**Prompt D: Free Think**

```
Rewrite the following response in your own words, adding your
thinking process before the final answer.  Keep the SAME final
answer.
Question:
{q}
Response to rewrite:
{gold}
```

**Prompt B: No Format, on tool-use**

```
Rewrite the following response in your own words.  Keep the SAME
final answer.
```

```
Question:
{q}
Response to rewrite:
{gold}
End with exactly this format:  Action:  <tool>
Action Input:  <json>
```

Prompt C: Thinking, on tool-use

```
Rewrite the following response in your own words, including the
thinking process.  Keep the SAME final answer.
Question:
{q}
Response to rewrite:
{gold}
End with exactly this format:  Action:  <tool>
Action Input:  <json>
```

# B Setup for the In-distribution Experiments

Both experiments continue fine-tuning `Qwen2.5-3B-Instruct` in `bf16` with LoRA ($r$=16, $\alpha$=32, all linear layers) for 1 epoch at a constant learning rate of $10^{-4}$ with AdamW, seed 1, and differ only in the training data. Decoding is greedy. Tab. 3 lists every dataset Experiment 1 trains on, with its source and upload date.

# C Further Experimental Details

## C.1 Hardware

All experiments run on A100 nodes. For other hardware such as CPU, RAM and disk we use commonly available ones, as these did not induce any bottlenecks.

## C.2 Datasets

All datasets are set to default values following the SDFT paper. Tab. 4 gives the Science Mixture split. SciKnowEval supplies only an answer key for its multiple-choice rows, so the `<reasoning>` demonstrations every arm trains on or rewrites are model-generated: GPT-4o for Chemistry, released by Shenfeld et al. (2026a), and `gpt-5.6-luna` for the other three subjects. **Gold** is therefore the subset of the L-3 pool a correct demonstration exists for, split 80/20 into training and test, so the evaluation prompts are filtered the same way, though every arm is scored on the same pool.

## C.3 Training Configuration

Tab. 5 lists the settings. Every cell is a single seed 42 chosen *a priori*, and never changed after experimentation.

| Training Dataset (HuggingFace Path) | Uploaded |
|---|---|
| **Pre-cutoff**: math released before the cutoff | |
| `microsoft/orca-math-word-problems-200k` | 2024-03 |
| `openai/gsm8k` (`main`) | 2022-04 |
| `AI-MO/NuminaMath-1.5` | 2025-02 |
| `EleutherAI/hendrycks_math` (`algebra`) | 2023-09 |
| `TIGER-Lab/MathInstruct` | 2023-09 |
| `deepmind/aqua_rat` (`raw`) | 2022-03 |
| `meta-math/MetaMathQA` | 2023-09 |
| `nvidia/OpenMathInstruct-1` | 2024-02 |
| **Post-cutoff**: long-CoT reasoning released after the cutoff | |
| `zwhe99/DeepMath-103K` | 2025-04 |
| `nvidia/OpenMathReasoning` (split `cot`) | 2025-04 |
| `simplescaling/s1K-1.1` | 2025-02 |
| `GAIR/LIMO` | 2025-02 |
| `open-thoughts/OpenThoughts-114k` | 2025-01 |
| `qihoo360/Light-R1-SFTData` | 2025-03 |
| `bespokelabs/Bespoke-Stratos-17k` | 2025-01 |
| `open-r1/OpenR1-Math-220k` | 2025-02 |
| `open-thoughts/OpenThoughts2-1M` | 2025-04 |
| `a-m-team/AM-DeepSeek-R1-Distilled-1.4M` (`am_0.5M`) | 2025-03 |
| **Off-target**: code or general instruction data, also post-cutoff | |
| `OpenCoder-LLM/opc-sft-stage2` (`educational_instruct`) | 2024-11 |
| `EricLu/SCP-116K` | 2025-01 |
| `allenai/tulu-3-sft-mixture` | 2024-11 |
| `KodCode/KodCode-V1` | 2025-02 |
| `nvidia/OpenCodeReasoning` (`split_0`) | 2025-04 |

Table 3: **The training data.** The 23 corpora Experiment 1 trains on, each taken as published and named by the HuggingFace path it is loaded from. Groupings are domain and upload date against the base model's 2024-09 cutoff, and the dates are each repository's `createdAt` read from the HuggingFace API.

| | | Per-subject domain | | | SCIENCE-MIX | |
|---|---|---|---|---|---|---|
| **Subject** | **L-3** | **Gold** | **Train** | **Test** | **Train** | **Test** |
| `Chemistry` | 3,542 | 2,674 | 2,139 | 535 | 668 | 133 |
| `Biology` | 679 | 356 | 285 | 71 | 285 | 71 |
| `Material` | 1,283 | 969 | 775 | 194 | 668 | 133 |
| `Physics` | 1,498 | 1,114 | 891 | 223 | 668 | 133 |
| **Total** | **7,002** | **5,113** | **4,090** | **1,023** | **2,289** | **470** |

Table 4: **Train/test split of the SCIENCE-MIX corpus.** SCIKNOWEVAL L-3 gold, split 80/20 at seed 0.

| Setting | Value |
|---|---|
| Adaptation | full fine-tuning of all parameters |
| Learning rate | $5 \times 10^{-5}$ |
| Epochs | 2 |
| Effective batch | 16 sequences |
| Optimizer | paged AdamW, 8-bit |
| Schedule | cosine; warmup ratio 0.1 |
| Gradient clipping | $\|g\| \leq 1.0$ |
| Seed | 42 |

Table 5: **Configuration for every cell in Tabs. 1 and 2.** Unlisted values are library defaults.

## D QUALITATIVE EXAMPLE OF ALIGNER REWRITES

A qualitative example is provided in Box D.

Sample: Science, aligner `Qwen2.5-7B`

**Question** (the task prompt, shown to the student at train and test time)

```
How do you calculate the half life of carbon 14?

A: 5.73 x 10^3 yr
B: 4.56 x 10^3 yr
C: 7.89 x 10^3 yr
D: 3.21 x 10^3 yr
```

**Gold** (the original demonstration; what plain SFT trains on)

```
<reasoning>
The half-life of carbon-14 is a well-established scientific fact
used in radiocarbon dating.  The correct value is approximately
5730 years.  This corresponds to option A: 5.73 x 10^3 yr.
</reasoning>
<answer>
A
</answer>
```

**Rewrite** (what SYNTHSELF trains on)

```
<reasoning>
The half-life of carbon-14 is known to be about 5730 years, which
matches option A: 5.73 x 10^3 yr.
</reasoning>
<answer>
A
</answer>
```