

# 000 001 002 003 004 005 006 007 008 009 010 011 012 013 014 015 016 017 018 019 020 021 022 023 024 025 026 027 028 029 030 031 032 033 034 035 036 037 038 039 040 041 042 043 044 045 046 047 048 049 050 051 052 053 MAPPING POST-TRAINING FORGETTING IN LANGUAGE MODELS AT SCALE

Anonymous authors

Paper under double-blind review

## ABSTRACT

Scaled post-training now drives many of the largest capability gains in language models (LMs), yet its effect on pretrained knowledge remains poorly understood. Not all forgetting is equal: Forgetting one fact (e.g., a U.S. president or an API call) does not “average out” by recalling another. Hence, we propose a sample-wise paradigm to measure what is forgotten and when backward transfer occurs. Our metric counts  $1 \rightarrow 0$  transitions (correct before post-training, incorrect after) to quantify forgetting and  $0 \rightarrow 1$  transitions to quantify backward transfer. Traditional task averages conflate these effects and obscure large changes. For multiple-choice benchmarks, we add chance-adjusted variants that subtract the expected contribution of random guessing from pre- and post-training accuracies. We apply this framework across post-training stages, model sizes, and data scales. Our large-scale analysis across nearly 30 model pairs and 100 sub-benchmarks with up to 32,768 generated tokens per sample shows that: (1) Domain-continual pretraining induces moderate forgetting with low-to-moderate backward transfer; (2) RL/SFT post-training applied to base models and Instruction tuning yields moderate-to-large backward transfer on math and logic with overall low-to-moderate forgetting; (3) Applying RL/SFT to instruction-tuned models is sensitive on data scale: at small scales, both forgetting and backward transfer are small; at larger scales, effects are mixed and warrant further study with better controls; (4) Model merging does not reliably mitigate forgetting. Overall, our framework offers a practical yardstick for mapping how post-training alters pretrained knowledge at scale – enabling progress towards generally capable AI systems.

## 1 INTRODUCTION

Scaling post-training has become the dominant driver of capability gains in modern language models (LMs) (Jaech et al., 2024). Practitioners now iterate through multi-step post-training pipelines often at data scales that rival early pretraining (Tie et al., 2025). The implicit bet is that each step in the pipeline accumulates new capabilities, with dramatic improvements in areas like coding, math, tool use and safety, without sacrificing the broad world knowledge. In contrast, it is considered common knowledge in continual learning that this sequential training would lead to catastrophic forgetting (see Table 1). We test this assumption: as we scale post-training, do we erode the very breadth of world knowledge that pretraining painstakingly compresses into the weights? If the implicit assumption does not hold, we risk trading generalist competence for narrow specialization, undermining progress toward generally capable models.

Measuring forgetting in modern post-training pipelines is tricky. Classical evaluations compare aggregate test accuracy before and after training (Luo et al., 2025), implicitly treating a benchmark as a single task with fungible i.i.d. samples (e.g., classifying images of cats). Pretrained knowledge violates this assumption. Knowing one U.S. president does not compensate for forgetting another; recalling a NumPy broadcasting rule does not offset losing a specific cloud-API syntax. In short, knowledge samples are not fungible: Each carries unique value for quantifying pretraining knowledge. Aggregation can hide substantial losses. Hence, we measure forgetting and backward transfer in a sample-wise manner, rather than at the task level as proposed by Lopez-Paz & Ranzato (2017).

Specifically, we define *forgetting* as items that are answered correctly before a post-training stage but incorrectly afterward (the  $1 \rightarrow 0$  transitions), and *backward transfer* as items that are answered incorrectly before but correctly after post-training (the  $0 \rightarrow 1$  transitions). A further complication is that most knowledge-intensive LLM evaluation benchmarks are multiple-choice. Random guessing

054 inflates accuracy and can create illusory transitions: an apparent “1 → 0” may simply be a lucky  
 055 guess that later becomes an incorrect answer, even when the underlying knowledge did not change;  
 056 likewise for 0 → 1 transitions. When the answer is only among few options (e.g., 4), performance by  
 057 random guessing can account for a substantial share of observed transitions, distorting both level and  
 058 trend estimates of forgetting. Thus a principled metric should (i) resolve outcomes at the *item* level  
 059 and (ii) explicitly correct for chance.

060 To account for these considerations we introduce chance-adjusted metrics for forgetting ( $F_{\text{true}}$ ) and  
 061 backward transfer ( $BT_{\text{true}}$ ), which correct for transitions expected under random choice. They do  
 062 not need logits or repeated sampling, measurable using the number of choices in benchmark and  
 063 marginal accuracy of the model pre- and post- training, making them practical at scale. Intuitively,  
 064 chance-adjusted forgetting asks: among items the model genuinely knew before, what fraction  
 065 became wrong beyond chance? Conversely, chance-adjusted backward asks: among items the model  
 066 genuinely did not correctly solve, what fraction became correct beyond chance?

067 Our primary contribution is a large-scale study measuring forgetting caused by post-training across  
 068 post-training pipelines. By evaluating the models on the same set of samples before and after each  
 069 stage, we obtain a map of what was retained, what was forgotten, and where losses concentrate.  
 070 We seek to answer three questions: (i) Where in the pipeline is forgetting most pronounced (e.g.,  
 071 instruction tuning vs. reasoning-focused training)?, (ii) What kinds of pretraining knowledge are  
 072 most affected (culture vs. logic)?, and (iii) How much knowledge is forgotten or re-elicited? We have  
 073 the following key findings:

#### Key Findings

- **Domain-Continual Pretraining** induces low to moderate forgetting across most categories; backward transfer is limited. Forgetting effects marginally decrease with increasing model scale.
- **Instruction-Tuning and SFT/RL from base models** yield low to moderate forgetting, with spikes in the Culture and Knowledge categories, but moderate to high (for SFT/RL from Base) backward-transfer gains in the Math and Logic categories across model families; Forgetting and backtransfer decrease as parameters increase. Reasoning training yields similar forgetting and larger backward transfer than instruction tuning.
- **SFT/RL Reasoning Post-Training from instruct models** have data-scale dependent behaviour: For the low-data regime, it yields low forgetting and backward transfer. For the high-data regime, no dominant factor robustly described the forgetting and backward transfer dynamics.
- **Model Merging** does not reliably mitigate forgetting across post-training pipelines (yet).

090 **Table 1: Catastrophic forgetting literature across LLM post-training stages.** Continual learning  
 091 literature indicates extensive forgetting across the post-training pipeline. However, we find far less  
 092 forgetting when testing widely used post-training pipelines, indicating an important gap existing  
 093 between continual learning setups and how people post-train language models.

095 Stage	096 Name	097 Level	098 Summary
099 CPT (§3.1)	100 Investigating Continual Pretraining 101 in LLMs: Insights and Implications (Yıldız et al., 2024)	Med	102 Most models show continual im- 103 provement; only Llama-2 models 104 degrade.
	105 Examining Forgetting in Continual 106 Pre-training of Aligned LLMs (Li & 107 Lee, 2024a)		108 Continual pre-training degrades ca- 109 pabilities, alignment and alters out- 110 put behavior.
SFT/DPO (§3.2)	Mitigating Forgetting in LLM Su- 103 pervised Fine-Tuning and Preference 104 Learning (Fernando et al., 2024)	Low	105 Combining SFT and DPO sequen- 106 tially leads to forgetting and a poor 107 balance between goals (~ 2% on 108 MMLU).

(Continued on next page)

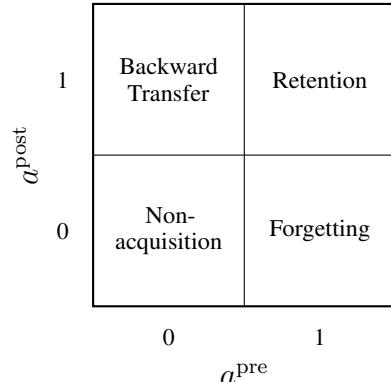
108  
109  
110  
(Continued from previous page)

Stage	Name	Level	Summary
SFT (\\$3.3)	Interpretable Catastrophic Forgetting of LLM Fine-tuning via Instruction Vector (Jiang et al., 2024)	High	Fine-tuning on TRACE shows declines primarily from lost instruction-following ability.
	An Empirical Study of Catastrophic Forgetting in LLMs During Continual Fine-tuning (Luo et al., 2025)	High	Forgetting of domain knowledge, reasoning intensifies as model scale increases ( $\sim 10\%$ MMLU drop).
	Catastrophic Forgetting in LLMs: A Comparative Analysis Across Language Tasks (Haque, 2025)	High	Severity varies by architecture and pre-training quality; some models degrade sharply while others barely change.
	Mitigating Catastrophic Forgetting in LLMs with Self-Synthesized Rehearsal (Huang et al., 2024)	High	Sequential fine-tuning causes major forgetting; synthetic rehearsal mitigates it.
RL (\\$3.2)	Mitigating the Alignment Tax of RLHF (Lin et al., 2024)	Med	RLHF induces forgetting (“alignment tax”); model averaging reduces it.
SFT/RL (\\$3.2)	Understanding Catastrophic Forgetting in LLMs via Implicit Inference (Kotha et al., 2024)	High	Fine-tuning skews the model’s implicit task inference rather than erasing capabilities.
	Temporal Sampling for Forgotten Reasoning in LLMs (Li et al., 2025)	High	Fine-tuned LLMs often forget solutions they previously generated (“temporal forgetting”) across sizes and methods (SFT, GRPO).

136  
137  
2 MEASURING SAMPLEWISE FORGETTING AND BACKWARD TRANSFER138  
139  
140  
141  
142  
143  
144  
145  
146  
147  
148  
149  
150  
151  
To formalize these metrics, first consider an evaluation set of  $N$  multiple-choice questions with  $k$  options. For each sample  $i$ , let  $a_i^{\text{pre}}, a_i^{\text{post}} \in \{0, 1\}$  indicate correctness before and after post-training. As illustrated in Fig. 1, each sample falls into one of four quadrants based on effect of training on new task:152  
153  
154  
155  
156  
157  
(i) Retention preserves knowledge ( $1 \rightarrow 1$ ),  
(ii) Backward Transfer improves performance ( $0 \rightarrow 1$ ),  
(iii) Forgetting reduces performance ( $1 \rightarrow 0$ ), and  
(iv) non-acquisition has no effect ( $0 \rightarrow 0$ ).158  
159  
160  
161  
We define sample-wise *forgetting* and *backward transfer* as the proportions of  $1 \rightarrow 0$  and  $0 \rightarrow 1$  flips, respectively:

162  
163  
164  
$$F = \frac{1}{N} \sum_{i=1}^N \mathbf{1}\{a_i^{\text{pre}} = 1 \wedge a_i^{\text{post}} = 0\}$$

165  
166  
167  
$$BT = \frac{1}{N} \sum_{i=1}^N \mathbf{1}\{a_i^{\text{pre}} = 0 \wedge a_i^{\text{post}} = 1\}$$

168  
169  
170  
171  
However, these intuitive metrics confound genuine knowledge change with label flips caused by guessing, especially when  $k$  is small. For example, two independent random binary classifiers ( $k=2$ ) yield  $F = 0.25$  because  $0.5 \times 0.5 = 0.25$ .172  
173  
174  
Figure 1: Each sample is assigned to one of four quadrants by correctness before and after.

162  
163  
164  
165  
166  
167  
168  
169  
170  
171  
172  
173  
174  
175  
176  
177  
178  
179  
180  
181  
182  
183  
184  
185  
186  
187  
188  
189  
190  
191  
192  
193  
194  
195  
196  
197  
198  
199  
200  
201  
202  
203  
204  
205  
206  
207  
208  
209  
210  
211  
212  
213  
214  
215

**A chance baseline for flips.** To account for guessing, we assume a simple response model: on each item the model either *knows* the answer or *guesses* uniformly among the  $k$  choices. Let  $\bar{a}$  be mean accuracy on a set. Then  $\bar{a} = \bar{a}_{\text{true}} + x$ , where  $x$  is the fraction correct by chance. Since an incorrect guess occurs with probability  $(k-1)/k$ ,

$$\frac{1-\bar{a}}{x+(1-\bar{a})} = \frac{k-1}{k} \implies x = \frac{1-\bar{a}}{k-1}.$$

A  $1 \rightarrow 0$  flip due purely to chance requires (i) a pre-training correct guess and (ii) a post-training error (converse for backward transfer). Assuming independence between pre- and post-training guessing events,

$$F_{\text{chance}} = \underbrace{\frac{1-\bar{a}^{\text{pre}}}{k-1}}_{\text{correct by chance (pre)}} \cdot \underbrace{(1-\bar{a}^{\text{post}})}_{\text{incorrect (post)}}, \quad BT_{\text{chance}} = \underbrace{(1-\bar{a}^{\text{pre}})}_{\text{incorrect (pre)}} \cdot \underbrace{\frac{1-\bar{a}^{\text{post}}}{k-1}}_{\text{correct by chance (post)}}.$$

These metrics depend only on aggregate accuracies and  $k$ ; they require no logits or heavy computation.

**Chance-adjusted forgetting and backward transfer.** From these estimates we can isolate knowledge change beyond chance by subtracting the baselines from the respective forgetting/backward-transfer and clip at zero:

$$F_{\text{true}} = \max(F - F_{\text{chance}}, 0), \quad BT_{\text{true}} = \max(BT - BT_{\text{chance}}, 0).$$

For example, if accuracy drops from 80% to 70% on a 4-option MCQ test, raw forgetting is 10%, but chance-adjusted forgetting is only about 6% – showing how the correction removes the effect of lucky guesses. Clipping ensures the metric remains valid even if models perform below chance. In practice, for an accurate measure of forgetting this metric’s mean and variance statistics should be computed over multiple seeds as is described in Section B.1.

**Ceilings: how much could a model forget or improve?** Observed forgetting can be small simply because little was truly correct to begin with. The *maximum possible* forgetting equals the fraction truly correct before post-training, which we adjust for guessing and clip at 0:

$$F_{\text{max}} = \bar{a}_{\text{true}}^{\text{pre}} = \max(\bar{a}^{\text{pre}} - x^{\text{pre}}, 0) = \max\left(\frac{k\bar{a}^{\text{pre}} - 1}{k-1}, 0\right).$$

Similarly, the *maximum possible* backward transfer equals the fraction truly correct after post-training:

$$BT_{\text{max}} = \bar{a}_{\text{true}}^{\text{post}} = \max(\bar{a}^{\text{post}} - x^{\text{post}}, 0) = \max\left(\frac{k\bar{a}^{\text{post}} - 1}{k-1}, 0\right).$$

where

$$x^{\text{pre}} = \frac{1-\bar{a}^{\text{pre}}}{k-1}, \quad x^{\text{post}} = \frac{1-\bar{a}^{\text{post}}}{k-1}$$

By construction  $F_{\text{true}} \leq F_{\text{max}}$  and  $BT_{\text{true}} \leq BT_{\text{max}}$ . Reporting the adjusted metrics alongside these ceilings separates true knowledge loss/acquisition from chance and contextualizes headroom for degradation or improvement.

**Assumptions and scope.** The correction uses two assumptions: (i) when the model does not know an answer, it guesses uniformly at random; and (ii) pre- and post-training guessing events are independent. These assumptions allow dataset-level adjustments from pre- and post-training accuracies alone. Note that  $F_{\text{true}}$  could quantify failure to elicit previously accessible knowledge and need not imply that the model has lost/unlearned the underlying information. Likewise, changes in  $BT_{\text{true}}$  often reflect improved elicitation rather than newly acquired knowledge.

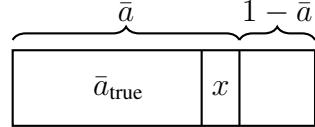


Figure 2: Accuracy  $\bar{a}$  decomposes into true knowledge  $\bar{a}_{\text{true}}$  and lucky guesses  $x$ .

216 3 WHEN, WHAT & HOW MUCH IS PRETRAINING KNOWLEDGE FORGOTTEN?  
217218 In this section, we ask three questions:  
219220 1. *When is pretraining knowledge forgotten?*221 Our analysis spans four widely used continual-training regimes: (i) domain-continual  
222 training (§3.1), (ii) instruction tuning (§3.2), (iii) light SFT/RL on reasoning traces, and (iv)  
223 large-scale SFT/RL for reasoning (§3.3). In total, we evaluate almost 30 model-training  
224 combinations chosen to reflect common practice results, providing broad coverage of how  
225 contemporary LLMs are post-trained in the wild. Each post-trained model is compared with  
226 its initial checkpoint (details in the Appendix).227 2. *What pretraining knowledge is forgotten?*228 We evaluate each model on 12 public benchmarks, collectively subdivided into close to a  
229 100 total subdomains. To summarize systematic patterns, we cluster sub-benchmarks into  
230 nine semantically coherent groups that exhibit similar forgetting trends (e.g., common sense,  
231 culture, deduction, language/communication, liberal arts, science/tech). These clusters  
232 provide a better map of which pretraining knowledge areas are most affected by a given  
233 post-training recipe.234 3. *How much pretraining knowledge is forgotten?*235 Unless stated otherwise, chance-adjusted metrics for forgetting ( $F_{true}$ ) and backward transfer  
236 ( $BT_{true}$ ) are used to quantify the severity.237 **Experimental setup.** We standardize settings across models for fair comparison. All experiments  
238 use the LightEval framework (Habib et al., 2023) and log per-sample accuracy. We apply a  
239 zero-shot chain-of-thought prompt to all models and require answers in a fixed MCQ format (see  
240 Appendix); base models receive a few-shot prompt solely to teach the format. When available<sup>1</sup>, we  
241 add chat-specific templates to be in line with best practices. We cap sequence length at 32K tokens,  
242 except for Qwen2.5-7B-Math and Qwen2.5-7B-Math-Instruct Yang et al. (2024a), which are limited  
243 to 4K<sup>2</sup>. Decoding uses temperature 0.6 with nucleus sampling ( $\text{top\_p}$ ) of 0.95. We provide additional  
244 details in the Appendix. We provide extensive quantitative results in Appendix G and provide figures  
245 and qualitative commentary in the following sections, defining moderate forgetting as  $15 \pm 5\%$ , low  
246 forgetting to be below that, and high forgetting to be above that. To facilitate reproducibility and  
247 further inquiry, we will release per-sample logs for every sub-benchmark alongside code.

248 We now showcase our results in the subsections below.

## 249 3.1 SUBAREA 1: DOMAIN-CONTINUAL PRETRAINING

250 **Motivation.** A popular class of continual learning works adapt general LLMs at the application layer  
251 for domains such as coding, mathematics, search, and tool use. As generalist LLMs are increasingly  
252 wrapped with tools and domain-specific interfaces, specialization must not erode broad pretraining  
253 knowledge. Models still need to contextualize domain outputs, communicate with diverse users,  
254 respect cultural norms, and uphold safety and ethical standards. These needs motivate our study of  
255 forgetting and backward transfer under domain-continual pretraining.256 **Setup.** We study continual pretraining that converts a general base model into a specialized one,  
257 exemplified by Qwen2.5-Coder (Hui et al., 2024) and Qwen2.5-Math (Yang et al., 2024b).<sup>3</sup> Unlike  
258 general instruction tuning or reasoning post-training, domain-continual pretraining shifts the  
259 underlying representation using large, relatively uncurated, web-scale domain corpora.260 **Main results.** Figure 3 summarizes our findings. Domain-continual pretraining induces little to  
261 moderate amounts of forgetting among all post-training methods we evaluate. Backward transfer  
262 to general abilities is weak: Gains in the specialized domain rarely improve non-target tasks.  
263 The effect spans categories of pretraining knowledge, with no single category driving it, although264  
265 <sup>1</sup>This budget was sufficient in practice; we never required more tokens.  
266267 <sup>2</sup>Because base models sometimes continue into subsequent questions, we set explicit stop sequences to end  
268 generation once a prediction is produced.269 <sup>3</sup>We treat domain-continual reasoning via SFT/RL separately in §3.3 and focus on domain-continual training  
here.

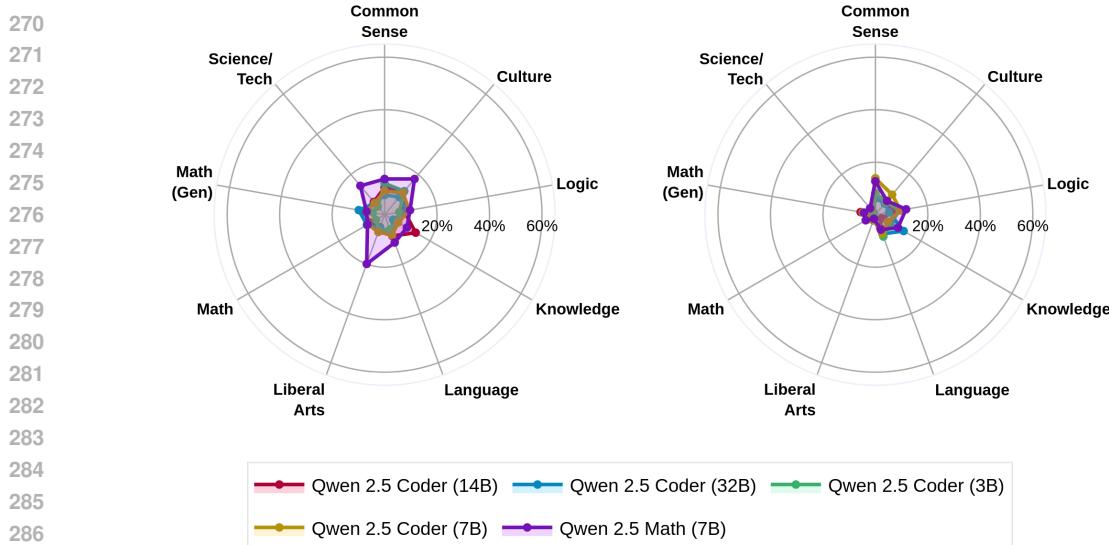


Figure 3: **Forgetting (left) and Backward Transfer (right) after domain-continual pretraining.** Forgetting is low-to-moderate and consistent across categories; backward transfer is low. Scaling model size reduces forgetting.

math-specialized models show significantly more forgetting. Lastly, larger models forget less and have marginally better backward transfer.

**Qualitative analysis.** We performed manual errors analysis, which indicates reduced instruction-following fidelity (e.g., weaker adherence to constraints, formats, and role-specific directives). Evidence of this is found in supplemental tests, where a zero-shot, chat-template evaluation is done. In this case, a coder model may, for example, answer “Who was the president of the US?” with a response followed by code, often with embedded answers, making extraction difficult of the “true answer”. While few-shot prompting alleviates this, it demonstrates a weakened instruction-following ability and less easily elicited knowledge.

#### Takeaway

Domain-continual pretraining yields low-to-moderate forgetting across categories; backward transfer is limited. Scaling model size marginally reduces forgetting. This indicates current domain-continual pretraining pipelines appear to alleviate much of the large forgetting behavior seen in previous literature.

### 3.2 SUBAREA 2: INSTRUCTION TUNING

**Motivation.** Base models often require carefully engineered prompts to elicit pretraining knowledge, limiting usability. Modern post-training pipelines therefore add instruction tuning to enable natural user interaction with minimal prompting. Most continual-learning work we surveyed focuses on mitigating forgetting in this setting. We ask: To what extent does instruction following come at the expense of previously learned knowledge?

**Setup.** We measure forgetting and backward transfer from instruction tuning in generalist models (Qwen2.5 (Yang et al., 2024a), Llama 3.1 (Dubey et al., 2024)) and domain-continual pretrained models (Qwen2.5-Coder)<sup>4</sup>.

**Results.** As shown in Figure 4, there is low to moderate forgetting across models, with spikes in the Culture and Knowledge categories. However, there is substantial backward transfer in the Math category. Furthermore, scaling model size reduces forgetting and increases backward transfer. This effect is consistent across domain-general and domain-specific base models. While most of the

<sup>4</sup>Qwen2.5-Math Instruct is surprisingly tuned with GRPO which leads to it being classified under Reasoning

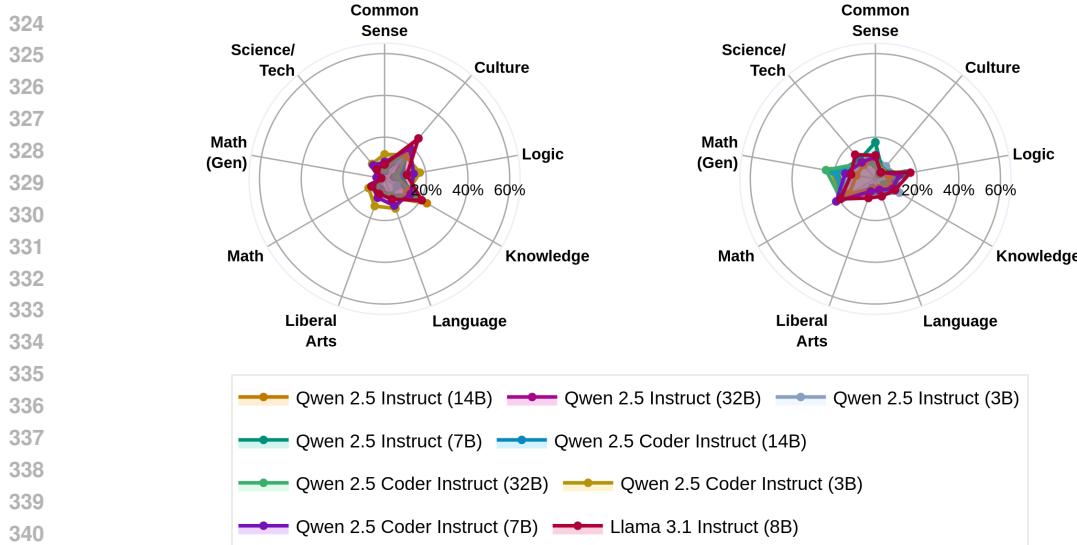


Figure 4: **Forgetting (left) and Backward Transfer (right) after instruction-tuning** yields moderate forgetting and backward transfer categories-wise. Scaling model size reduces forgetting and backward transfer.

continual learning literature focuses on reducing forgetting in this area, we note the forgetting is low to moderate with current training practices.

**Qualitative analysis.** Transfer gains likely reflect better elicitation of pretraining knowledge: Instruction-tuned models use what they already know with straightforward prompts used in benchmarks, whereas base models often require carefully crafted prompts.

### Takeaway

Instruction tuning produces low-to-moderate forgetting overall and moderate backward-transfer, particularly in math, across model families; the forgetting and back-transfer tend to decrease with increasing model scale. Shifting focus to other subareas of post-training might spur interesting research directions, but there is still progress to be made in this area.

### 3.3 SUBAREA 3: TRAINING WITH REASONING TRACES (SFT AND RL)

**Motivation.** Recent methods encourage explicit reasoning by letting models *think* on a scratchpad before answering; which is now scaled in size and trace length with RL objectives. As training domains and data grow, we measure how much such reasoning training induces forgetting to guide continual-learning practice.

**Setup.** We consider two settings: (i) starting from a base model and (ii) starting from an instruction-tuned model. For the latter, we separate light-touch post-training (small datasets) from heavy post-training. We do not separate RL from SFT as the behavior across forgetting and backward transfer is similar between the two objectives.

#### 3.3.1 TRAINING WITH REASONING TRACES FROM BASE MODELS

**Models.** We evaluate QwQ-32B (from Qwen2.5-32B Base) (Qwen Team, 2025), Qwen2.5-Math-7B-Instruct (RL post-trained with GRPO), and DeepSeek-R1-Distill models across different models (Qwen2.5 Base and Llama 8B base) (DeepSeek-AI, 2025).

**Results.** From Figure 5, we see that across scales, model families, and training types, we observe large gains, particularly in Math and Logic, in backward transfer with minimal forgetting. Forgetting is generally low, but is moderate for knowledge and large for Culture. The exception to this trend is the Qwen2.5 Math Instruct model which shows substantial forgetting across many categories. Sample-wise inspection shows this is primarily due to weak adherence to the prompt, sometimes

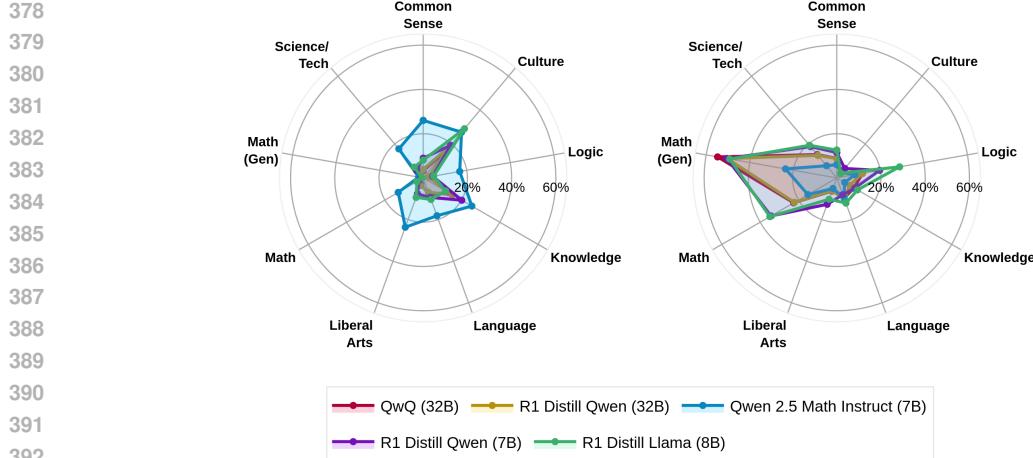


Figure 5: **Forgetting (left) and Backward Transfer (right) after reasoning training (SFT/RL) from base model.** It generally yields minimal forgetting, except in the Culture and Knowledge categories, and has moderate to high backward-transfer gains. Qwen2.5 Math Instruct (7B) is an exception to this trend, demonstrating forgetting across all categories.

outputting random multilingual text. Except for this case, when compared to instruction tuning on the same base model (Figure 4), we see similar forgetting and larger back-transfer<sup>5</sup>.

We conclude that much of the backward transfer reflects improved instruction following. To isolate reasoning effects beyond elicitation, the next sections analyze reasoning training that starts from an instruction-tuned model, for better exploration of gains. However, models with light-touch reasoning training (i.e. low data) behave differently from those trained at scale (i.e. high data). We therefore present these two cases separately.

#### Takeaway

Training with SFT/RL for reasoning results in dynamics similar to instruction tuning, but to an even greater extent: We generally observe low to moderate forgetting overall and larger category-specific backward transfer gains. Forgetting mitigation in this domain should consider broad categories of knowledge/abilities when measuring forgetting and back transfer.

#### 3.3.2 REASONING TRAINING FROM INSTRUCTION-TUNED MODELS: LOW-DATA SCENARIO

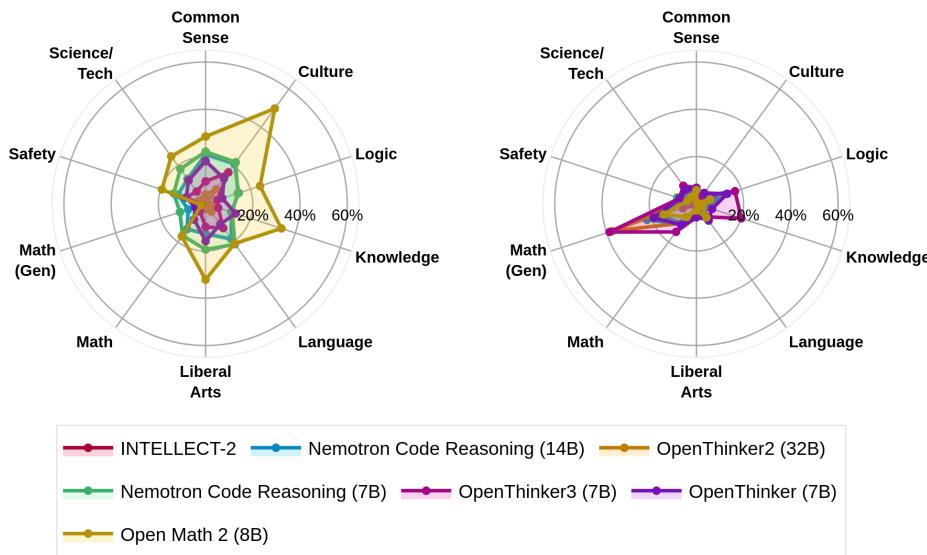
**Models.** We use the s1.1 family (7B, 14B, 32B) (Muennighoff et al., 2025) and LIMO (v1 and v2) (Ye et al., 2025) all tuned from corresponding sized Qwen instruct models.

**Results.** Figure 26 summarizes our findings. Across categories, models show minimal forgetting and low backward transfer, except generative math where large gains occurs. This makes sense, as training for a few passes on little data leaves pretraining knowledge largely intact. That is, the model does not forget much, but it also exhibits little backward transfer gains beyond the instruction-tuned baseline. Scaling model size marginally lowers forgetting, and the smaller teacher-student gap similarly tends to reduce backward transfer, with the exception of the Knowledge category.

#### Takeaway

For low-data regime, reasoning training from instruct models yields low forgetting and backward transfer. Forgetting decreases with model scale; backward transfer gains also tend to fall with a narrowing student-teacher gap. This suggests that future forgetting mitigation literature on reasoning models should focus on medium-to-large sized training datasets.

<sup>5</sup>All corresponding tables are available in Appendix G.3 for detailed comparison.

432 3.3.3 REASONING TRAINING FROM INSTRUCTION-TUNED MODELS: HIGH-DATA SCENARIO  
433452 **Figure 6: Forgetting (left) and Backward Transfer (right) after reasoning training from instruct:**  
453 **high data scenario.** No single factor robustly explains the dynamics of forgetting and backward  
454 transfer.455 **Models.** We evaluate OpenCodeReasoner and OpenMath2 (Bercovich et al., 2025), OpenThinker-7B,  
456 OpenThinker2-32B, and OpenThinker3-7B (Guha et al., 2025), and Intellect-2-32B (Prime Intellect  
457 Team et al., 2025). This spans SFT (former) and RL (Intellect-2).458 **Results.** Results vary by domain mix and model quality. The OpenThinker models generally show  
459 low-to-moderate forgetting and moderate backward transfer, perhaps due to the breadth of the  
460 training datamix, whereas OpenCodeReasoner models show consistently high forgetting with low  
461 backward transfer gains due to the narrower training data. Furthermore, we find this may be primarily  
462 due to weakened instruction-following capabilities, as sample-level inspection shows the model will  
463 refuse to answer with letters, when numbers are present as options, instead answering numerically.  
464 This is also seen with the Nemotron Code Reasoning models, where answers will often be embedded  
465 within python code. These factors can make the forgetting and backward transfer observed highly  
466 dependent on the extraction method used. We account for this through LLM as a judge in Section C.4.  
467 Scaling model size, if compared in OpenThinker models, signals improvements in both forgetting  
468 and backward transfer – as seen in most previous sections. Decentralized training (as in Intellect-2),  
469 in contrast, showed minimal forgetting or backward transfer. We conjecture that the model largely  
470 remain unchanged compared to the original model as it shows negligible gains on the optimized math  
471 benchmarks Hochlehnert et al. (2025). However, the results here remain preliminary. We do not find  
472 a single dominant factor—initialization, data regime, or scale that sufficiently explains forgetting and  
473 backward-transfer dynamics. We believe controlling the finer details which determine the quality of  
474 the trained model might lead to better conclusions.475 **Takeaway**476 No single factor robustly explains the dynamics of forgetting and backward transfer; training on  
477 a mix of domains appear to improve both forgetting and backward transfer.

## 480 4 DOES MODEL MERGING REDUCE FORGETTING?

481 **Motivation.** Recent work shows that offline model merging can combine capabilities from multiple  
482 models (Dziadzio et al., 2025). Unlike classical continual learning (De Lange et al., 2022), it  
483 requires neither the original training data nor the ability to resume training, which is practical in  
484 resource-constrained settings.

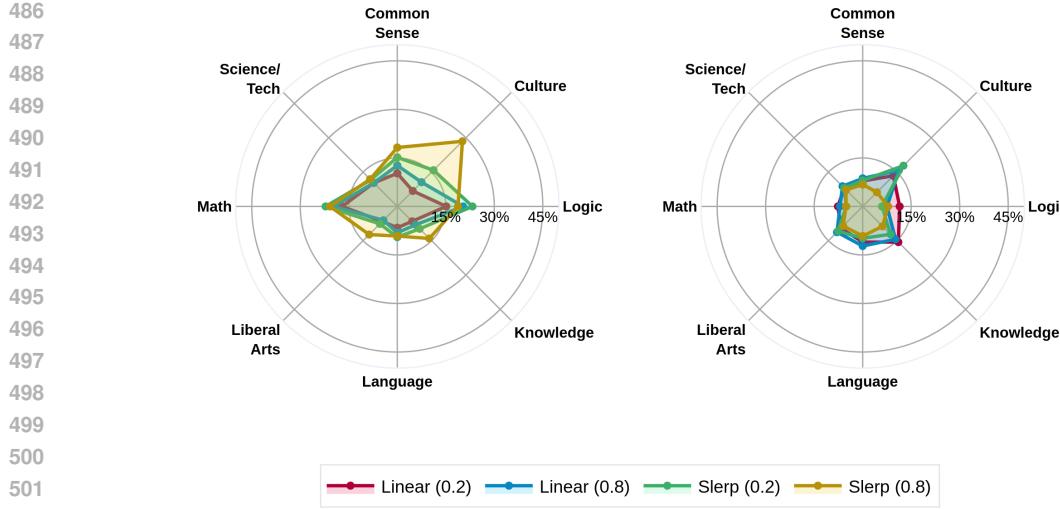


Figure 7: **Forgetting and Backward Transfer of Qwen 2.5 Base merged with Qwen 2.5 Coder (7B) relative to Qwen2.5 Coder.** Induces moderate forgetting and little backward transfer.

**Setup.** We evaluate Exponential Moving Average (EMA) merging; in the two-checkpoint case this is linear interpolation,

$$\theta_{\text{EMA}}(\alpha) = \alpha \theta_{\text{pre}} + (1 - \alpha) \theta_{\text{post}}.$$

Prior large-scale studies find these simple schemes effective for continual learning with foundation models (Roth et al., 2024). Our experiments compare linear interpolations (e.g. LERP and SLERP) across OpenThinker-7B, OpenThinker3-7B, and Qwen2.5-Coder-7B, together with their base checkpoints.

**Results.** We compare merged checkpoints to the post-trained model  $\theta_{\text{post}}$ ; results for  $\theta_{\text{pre}}$  appear in the Appendix. For Qwen2.5-Coder-7B and OpenThinker3-7B, even small mixes with the base checkpoint degrade performance, severely for the latter case (Figures 7, 15). In contrast, OpenThinker-7B shows small overall gains, accompanied by moderate forgetting (Figure 19). In our setting, merging does not mitigate forgetting. This may reflect that we merge only two checkpoints, whereas prior work often merges eight or more (Yadav et al., 2023; 2024). We further hypothesize that weight drift between our checkpoints is larger than is typical in the merging literature, which could explain these outcomes.

### Takeaway

Merging model does not yet reliably mitigate forgetting in post-training pipelines.

Merging remains promising, but further study is needed to determine when each method works, how to overcome its limitations, and whether an increased scale can compensate for these difficulties. Future works may consider the effect of the number of models merged and weight drift on reasoning models.

## 5 CONCLUSION

We present a new metric for sample-wise forgetting and backward transfer that corrects for chance in multiple-choice evaluations. Our results challenge a common claim: sequential training does not automatically erode pre-training knowledge. Forgetting depends on the post-training method and its scale. By focusing on sample-wise forgetting, we offer a clearer map of what knowledge is lost and in what stages of instruction tuning do language models lose during post-training – providing fertile ground to study how to preserve (minimize forgetting) and accumulate (higher backward transfer) knowledge while adding new capabilities by post-training. Promising ways to prevent forgetting include: (1) Designing objectives and data that explicitly penalize  $1 \rightarrow 0$  transitions; (2) Using targeted synthetic corpora or brief mid-training bursts to repair localized forgetting; (3) Adding retrieval mechanisms to reduce reliance on in-weight knowledge storage.

540 REFERENCES  
541

542 Samuel Ainsworth, Jonathan Hayase, and Siddhartha Srinivasa. Git re-basin: Merging models modulo  
543 permutation symmetries. In *The Eleventh International Conference on Learning Representations*,  
544 2023. URL <https://openreview.net/forum?id=CQsmMYmlP5T>.

545 Akhiad Bercovich, Itay Levy, Izik Golan, Mohammad Dabbah, Ran El-Yaniv, Omri Puny, Ido  
546 Galil, Zach Moshe, Tomer Ronen, Najeeb Nabwani, Ido Shahaf, Oren Tropf, Ehud Karpas, Ran  
547 Zilberman, Jiaqi Zeng, Soumye Singh, Alexander Bukharin, Yian Zhang, Tugrul Konuk, Gerald  
548 Shen, Ameya Sunil Mahabaleshwar, Bilal Kartal, Yoshi Suhara, Olivier Delalleau, Zijia Chen,  
549 Zhilin Wang, David Mosallanezhad, Adi Renduchintala, Haifeng Qian, Dima Rekesh, Fei Jia,  
550 Somshubra Majumdar, Vahid Noroozi, Wasi Uddin Ahmad, Sean Narenthiran, Aleksander Ficek,  
551 Mehrzad Samadi, Jocelyn Huang, Siddhartha Jain, Igor Gitman, Ivan Moshkov, Wei Du, Shubham  
552 Toshniwal, George Armstrong, Branislav Kisacanin, Matvei Novikov, Daria Gitman, Evelina  
553 Bakhturina, Jane Polak Scowcroft, John Kamalu, Dan Su, Kezhi Kong, Markus Kliegl, Rabeeh  
554 Karimi, Ying Lin, Sanjeev Satheesh, Jupinder Parmar, Pritam Gundecha, Brandon Norick, Joseph  
555 Jennings, Shrimai Prabhumoye, Syeda Nahida Akter, Mostofa Patwary, Abhinav Khattar, Deepak  
556 Narayanan, Roger Waleffe, Jimmy Zhang, Bor-Yiing Su, Guyue Huang, Terry Kong, Parth Chadha,  
557 Sahil Jain, Christine Harvey, Elad Segal, Jining Huang, Sergey Kashirsky, Robert McQueen, Izzy  
558 Puterman, George Lam, Arun Venkatesan, Sherry Wu, Vinh Nguyen, Manoj Kilaru, Andrew Wang,  
559 Anna Warna, Abhilash Somasudramath, Sandip Bhaskar, Maka Dong, Nave Assaf, Shahar  
560 Mor, Omer Ullman Argov, Scot Junkin, Oleksandr Romanenko, Pedro Larroy, Monika Katariya,  
561 Marco Rovinelli, Viji Balas, Nicholas Edelman, Anahita Bhiwandiwalla, Muthu Subramaniam,  
562 Smita Ithape, Karthik Ramamoorthy, Yuting Wu, Suguna Varshini Velury, Omri Almog, Joyjit  
563 Daw, Denys Fridman, Erick Galinkin, Michael Evans, Katherine Luna, Leon Derczynski, Nikki  
564 Pope, Eileen Long, Seth Schneider, Guillermo Siman, Tomasz Grzegorzek, Pablo Ribalta, Monika  
565 Katariya, Joey Conway, Trisha Saar, Ann Guan, Krzysztof Pawelec, Shyamala Prayaga, Oleksii  
566 Kuchaiev, Boris Ginsburg, Oluwatobi Olabiyi, Kari Briski, Jonathan Cohen, Bryan Catanzaro,  
567 Jonah Alben, Yonatan Geifman, Eric Chung, and Chris Alexiuk. Llama-nemotron: Efficient  
568 reasoning models, 2025. URL <https://arxiv.org/abs/2505.00949>.

569 Magdalena Biesialska, Katarzyna Biesialska, and Marta R. Costa-jussà. Continual lifelong learning  
570 in natural language processing: A survey. In Donia Scott, Nuria Bel, and Chengqing Zong  
571 (eds.), *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 6523–  
572 6541, Barcelona, Spain (Online), December 2020. International Committee on Computational  
573 Linguistics. doi: 10.18653/v1/2020.coling-main.574. URL <https://aclanthology.org/2020.coling-main.574/>.

574 Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. Piqa: Reasoning  
575 about physical commonsense in natural language. In *Thirty-Fourth AAAI Conference on Artificial  
576 Intelligence*, 2020.

577 Shiqi Chen, Jinghan Zhang, Tongyao Zhu, Wei Liu, Siyang Gao, Miao Xiong, Manling Li, and  
578 Junxian He. Bring reason to vision: Understanding perception and reasoning through model  
579 merging. In *Forty-second International Conference on Machine Learning*, 2025. URL <https://openreview.net/forum?id=ntCAP6tMoX>.

580 Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and  
581 Oyvind Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge.  
582 *arXiv:1803.05457v1*, 2018.

583 Matthias De Lange, Rahaf Aljundi, Marc Masana, Sarah Parisot, Xu Jia, Aleš Leonardis, Gregory  
584 Slabaugh, and Tinne Tuytelaars. A continual learning survey: Defying forgetting in classification  
585 tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(7):3366–3385, 2022.  
586 doi: 10.1109/TPAMI.2021.3057446.

587 DeepSeek-AI. Deepseek llm: Scaling open-source language models with longtermism. *arXiv preprint  
588 arXiv:2401.02954*, 2024. URL <https://github.com/deepseek-ai/DeepSeek-LLM>.

589 DeepSeek-AI. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement  
590 Learning, 2025. URL <https://arxiv.org/abs/2501.12948>.

594 Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha  
 595 Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn,  
 596 Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston  
 597 Zhang, Aurélien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Rozière, Bethany Biron,  
 598 Bin Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris  
 599 McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton  
 600 Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, David  
 601 Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes,  
 602 Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip  
 603 Radenovic, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme  
 604 Nail, Grégoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu,  
 605 Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel M. Kloumann, Ishan Misra, Ivan  
 606 Evtimov, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vrane, Jason Park, Jay Mahadeokar, Jeet  
 607 Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi,  
 608 Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph  
 609 Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasudevan Alwala, Kartikeya Upasani,  
 610 Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, and et al. The llama 3 herd of models. *CoRR*,  
 611 abs/2407.21783, 2024. URL <https://doi.org/10.48550/arXiv.2407.21783>.

612 Sebastian Dziadzio, Vishaal Udandarao, Karsten Roth, Ameya Prabhu, Zeynep Akata, Samuel  
 613 Albanie, and Matthias Bethge. How to merge multimodal models over time? In *ICLR 2025*  
 614 *Workshop on Modularity for Collaborative, Decentralized, and Continual Deep Learning*, 2025.

615 Heshan Fernando, Han Shen, Parikshit Ram, Yi Zhou, Horst Samulowitz, Nathalie Baracaldo, and  
 616 Tianyi Chen. Mitigating forgetting in llm supervised fine-tuning and preference learning. *arXiv*  
 617 *preprint arXiv:2410.15483*, 2024.

618 Heshan Fernando, Han Shen, Parikshit Ram, Yi Zhou, Horst Samulowitz, Nathalie Baracaldo, and  
 619 Tianyi Chen. Mitigating forgetting in llm supervised fine-tuning and preference learning, 2025.  
 620 URL <https://arxiv.org/abs/2410.15483>.

621 Robert M. French. Catastrophic forgetting in connectionist networks. *Trends in Cognitive Sciences*, 3(4):128–135, 1999. ISSN 1364-6613. doi: [https://doi.org/10.1016/S1364-6613\(99\)01294-2](https://doi.org/10.1016/S1364-6613(99)01294-2). URL <https://www.sciencedirect.com/science/article/pii/S1364661399012942>.

622 Etash Guha, Ryan Marten, Sedrick Keh, Negin Raoof, Georgios Smyrnis, Hritik Bansal, Marianna  
 623 Nezhurina, Jean Mercat, Trung Vu, Zayne Sprague, Ashima Suvarna, Benjamin Feuer, Liangyu  
 624 Chen, Zaid Khan, Eric Frankel, Sachin Grover, Caroline Choi, Niklas Muenninghoff, Shiye Su,  
 625 Wanja Zhao, John Yang, Shreyas Pimpalgaonkar, Kartik Sharma, Charlie Cheng-Jie Ji, Yichuan  
 626 Deng, Sarah Pratt, Vivek Ramanujan, Jon Saad-Falcon, Jeffrey Li, Achal Dave, Alon Albalak,  
 627 Kushal Arora, Blake Wulfe, Chinmay Hegde, Greg Durrett, Sewoong Oh, Mohit Bansal, Saadia  
 628 Gabriel, Aditya Grover, Kai-Wei Chang, Vaishaal Shankar, Aaron Gokaslan, Mike A. Merrill,  
 629 Tatsunori Hashimoto, Yejin Choi, Jenia Jitsev, Reinhard Heckel, Maheswaran Sathiamoorthy,  
 630 Alexandros G. Dimakis, and Ludwig Schmidt. Openthoughts: Data recipes for reasoning models,  
 631 2025. URL <https://arxiv.org/abs/2506.04178>.

632 Nathan Habib, Clémentine Fourrier, Hynek Kydlíček, Thomas Wolf, and Lewis Tunstall. Light-  
 633 eval: A lightweight framework for llm evaluation, 2023. URL <https://github.com/huggingface/lighteval>.

634 Naimul Haque. Catastrophic forgetting in llms: A comparative analysis across language tasks. *arXiv*  
 635 *preprint arXiv:2504.01241*, 2025.

636 Dan Hendrycks, Collin Burns, Steven Basart, Andrew Critch, Jerry Li, Dawn Song, and Jacob  
 637 Steinhardt. Aligning ai with shared human values. *Proceedings of the International Conference on*  
 638 *Learning Representations (ICLR)*, 2021a.

639 Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob  
 640 Steinhardt. Measuring massive multitask language understanding. *Proceedings of the International*  
 641 *Conference on Learning Representations (ICLR)*, 2021b.

648 Andreas Hochlehnert, Hardik Bhatnagar, Vishaal Udandarao, Samuel Albanie, Ameya Prabhu, and  
 649 Matthias Bethge. A sober look at progress in language model reasoning: Pitfalls and paths to  
 650 reproducibility. *arXiv preprint arXiv:2504.07086*, 2025.

651 Jianheng Huang, Leyang Cui, Ante Wang, Chengyi Yang, Xinting Liao, Linfeng Song, Junfeng Yao,  
 652 and Jinsong Su. Mitigating catastrophic forgetting in large language models with self-synthesized  
 653 rehearsal. *arXiv preprint arXiv:2403.01244*, 2024.

654 Binyuan Hui, Jian Yang, Zeyu Cui, Jiaxi Yang, Dayiheng Liu, Lei Zhang, Tianyu Liu, Jiajun Zhang,  
 655 Bowen Yu, Kai Dang, et al. Qwen2.5-coder technical report. *arXiv preprint arXiv:2409.12186*,  
 656 2024.

657 Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec  
 658 Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. Openai o1 system card. *arXiv preprint*  
 659 *arXiv:2412.16720*, 2024.

660 Gangwei Jiang, Zhaoyi Li, Defu Lian, and Ying Wei. Refine large language model fine-tuning via  
 661 instruction vector. *arXiv preprint arXiv:2406.12227*, 2024.

662 Suhas Kotha, Jacob Mitchell Springer, and Aditi Raghunathan. Understanding catastrophic forgetting  
 663 in language models via implicit inference. In *The Twelfth International Conference on Learning  
 664 Representations*, 2024. URL <https://openreview.net/forum?id=VrHiF2hsrm>.

665 Chen-An Li and Hung-Yi Lee. Examining forgetting in continual pre-training of aligned large  
 666 language models. *arXiv preprint arXiv:2401.03129*, 2024a.

667 Chen-An Li and Hung-Yi Lee. Examining forgetting in continual pre-training of aligned large  
 668 language models, 2024b. URL <https://arxiv.org/abs/2401.03129>.

669 Lijun Li, Bowen Dong, Ruohui Wang, Xuhao Hu, Wangmeng Zuo, Dahua Lin, Yu Qiao, and Jing  
 670 Shao. Salad-bench: A hierarchical and comprehensive safety benchmark for large language models.  
 671 *arXiv preprint arXiv:2402.05044*, 2024.

672 Yuetai Li, Zhangchen Xu, Fengqing Jiang, Bhaskar Ramasubramanian, Luyao Niu, Bill Yuchen  
 673 Lin, Xiang Yue, and Radha Poovendran. Temporal sampling for forgotten reasoning in LLMs. In  
 674 *Second Workshop on Test-Time Adaptation: Putting Updates to the Test! at ICML 2025*, 2025.  
 675 URL <https://openreview.net/forum?id=J0HWRSsPJ>.

676 Stephanie Lin, Jacob Hilton, and Owain Evans. TruthfulQA: Measuring how models mimic human  
 677 falsehoods. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds.), *Proceedings of  
 678 the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*,  
 679 pp. 3214–3252, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi:  
 10.18653/v1/2022.acl-long.229. URL [https://aclanthology.org/2022.acl-long.  
 680 229/](https://aclanthology.org/2022.acl-long.229).

681 Yong Lin, Hangyu Lin, Wei Xiong, Shizhe Diao, Jianmeng Liu, Jipeng Zhang, Rui Pan, Haoxiang  
 682 Wang, Wenbin Hu, Hanning Zhang, Hanze Dong, Renjie Pi, Han Zhao, Nan Jiang, Heng Ji, Yuan  
 683 Yao, and Tong Zhang. Mitigating the alignment tax of rlhf, 2024. URL [https://arxiv.org/  
 684 abs/2309.06256](https://arxiv.org/abs/2309.06256).

685 David Lopez-Paz and Marc'Aurelio Ranzato. Gradient episodic memory for continual learning.  
 686 *Advances in neural information processing systems*, 30, 2017.

687 Keming Lu, Bowen Yu, Fei Huang, Yang Fan, Runji Lin, and Chang Zhou. Online merging optimizers  
 688 for boosting rewards and mitigating tax in alignment. *arXiv preprint arXiv:2405.17931*, 2024.

689 Yun Luo, Zhen Yang, Fandong Meng, Yafu Li, Jie Zhou, and Yue Zhang. An empirical study of  
 690 catastrophic forgetting in large language models during continual fine-tuning. *IEEE Transactions  
 691 on Audio, Speech and Language Processing*, 2025.

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

702 Ken McRae and Amy Hetherington. Catastrophic interference is eliminated in pretrained networks.  
 703 1993. URL <https://api.semanticscholar.org/CorpusID:2129036>.  
 704

705 Sanket Vaibhav Mehta, Darshan Patil, Sarath Chandar, and Emma Strubell. An empirical investigation  
 706 of the role of pre-training in lifelong learning. *Journal of Machine Learning Research*, 24(214):  
 707 1–50, 2023. URL <http://jmlr.org/papers/v24/22-0496.html>.  
 708

709 Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi, Luke  
 710 Zettlemoyer, Percy Liang, Emmanuel Candès, and Tatsunori Hashimoto. s1: Simple test-time  
 711 scaling, 2025. URL <https://arxiv.org/abs/2501.19393>.  
 712

713 Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong  
 714 Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow  
 715 instructions with human feedback. *Advances in neural information processing systems*, 35:27730–  
 716 27744, 2022.  
 717

718 Baolin Peng, Linfeng Song, Ye Tian, Lifeng Jin, Haitao Mi, and Dong Yu. Stabilizing rlhf through  
 719 advantage model and selective rehearsal, 2023. URL <https://arxiv.org/abs/2309.10202>.  
 720

721 Kunat Pipatanakul, Pittawat Taveekitworachai, Potsawee Manakul, and Kasima Tharnpipitchai.  
 722 Adapting language-specific llms to a reasoning model in one day via model merging – an open  
 723 recipe, 2025. URL <https://arxiv.org/abs/2502.09056>.  
 724

725 Prime Intellect Team, Sami Jaghouar, Justus Mattern, Jack Min Ong, Jannik Straube, Manveer Basra,  
 726 Aaron Pazdera, Kushal Thaman, Matthew Di Ferrante, Felix Gabriel, Fares Obeid, Kemal Erdem,  
 727 Michael Keiblanger, and Johannes Hagemann. Intellect-2: A reasoning model trained through  
 728 globally decentralized reinforcement learning, 2025. URL <https://arxiv.org/abs/2505.07291>.  
 729

730 Qwen Team. Qwq-32b: Embracing the power of reinforcement learning, March 2025. URL  
 731 <https://qwenlm.github.io/blog/qwq-32b/>.  
 732

733 Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea  
 734 Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances  
 735 in neural information processing systems*, 36:53728–53741, 2023.  
 736

737 Vinay Venkatesh Ramasesh, Aitor Lewkowycz, and Ethan Dyer. Effect of scale on catastrophic  
 738 forgetting in neural networks. In *International Conference on Learning Representations*, 2022.  
 739 URL [https://openreview.net/forum?id=GhVS8\\_yPeEa](https://openreview.net/forum?id=GhVS8_yPeEa).  
 740

741 Roger Ratcliff. Connectionist models of recognition memory: Constraints imposed by learning and  
 742 forgetting functions. *Psychological Review*, 97(2):285–308, 1990. doi: 10.1037/0033-295X.97.2.285.  
 743

744 David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien  
 745 Dirani, Julian Michael, and Samuel R. Bowman. GPQA: A graduate-level google-proof q&a  
 746 benchmark. In *First Conference on Language Modeling*, 2024. URL <https://openreview.net/forum?id=Ti67584b98>.  
 747

748 Matthew Richardson, Christopher J.C. Burges, and Erin Renshaw. MCTest: A challenge dataset for  
 749 the open-domain machine comprehension of text. In David Yarowsky, Timothy Baldwin, Anna  
 750 Korhonen, Karen Livescu, and Steven Bethard (eds.), *Proceedings of the 2013 Conference on  
 751 Empirical Methods in Natural Language Processing*, pp. 193–203, Seattle, Washington, USA,  
 752 October 2013. Association for Computational Linguistics. URL <https://aclanthology.org/D13-1020/>.  
 753

754 Karsten Roth, Vishal Udandarao, Sebastian Dziadzio, Ameya Prabhu, Mehdi Cherti, Oriol Vinyals,  
 755 Olivier Hénaff, Samuel Albanie, Matthias Bethge, and Zeynep Akata. A practitioner’s guide to  
 continual multimodal pretraining. *arXiv preprint arXiv:2408.14471*, 2024.

756 Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. Social IQa: Common-  
 757 sense reasoning about social interactions. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan  
 758 (eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*  
 759 and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP),  
 760 pp. 4463–4473, Hong Kong, China, November 2019. Association for Computational Linguistics.  
 761 doi: 10.18653/v1/D19-1454. URL <https://aclanthology.org/D19-1454/>.

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

765

766 Zayne Rea Sprague, Xi Ye, Kaj Bostrom, Swarat Chaudhuri, and Greg Durrett. MuSR: Testing the  
 767 limits of chain-of-thought with multistep soft reasoning. In *The Twelfth International Conference*  
 768 on *Learning Representations*, 2024. URL <https://openreview.net/forum?id=jenyYQzuel>.

769

770 Mirac Suzgun, Nathan Scales, Nathanael Schärlí, Sebastian Gehrmann, Yi Tay, Hyung Won Chung,  
 771 Aakanksha Chowdhery, Quoc V Le, Ed H Chi, Denny Zhou, , and Jason Wei. Challenging  
 772 big-bench tasks and whether chain-of-thought can solve them. *arXiv preprint arXiv:2210.09261*,  
 773 2022.

774

775 Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. CommonsenseQA: A question  
 776 answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of*  
 777 *the North American Chapter of the Association for Computational Linguistics: Human Language*  
 778 *Technologies, Volume 1 (Long and Short Papers)*, pp. 4149–4158, Minneapolis, Minnesota, June  
 779 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1421. URL <https://aclanthology.org/N19-1421>.

780

781

782 Guiyao Tie, Zeli Zhao, Dingjie Song, Fuyang Wei, Rong Zhou, Yurou Dai, Wen Yin, Zhejian Yang,  
 783 Jiangyue Yan, Yao Su, Zhenhan Dai, Yifeng Xie, Yihan Cao, Lichao Sun, Pan Zhou, Lifang He,  
 784 Hechang Chen, Yu Zhang, Qingsong Wen, Tianming Liu, Neil Zhenqiang Gong, Jiliang Tang,  
 785 Caiming Xiong, Heng Ji, Philip S. Yu, and Jianfeng Gao. A survey on post-training of large  
 786 language models, 2025. URL <https://arxiv.org/abs/2503.06072>.

787

788 Mariya Toneva, Alessandro Sordoni, Remi Tachet des Combes, Adam Trischler, Yoshua Bengio, and  
 789 Geoffrey J. Gordon. An empirical study of example forgetting during deep neural network learning.  
 790 In *International Conference on Learning Representations*, 2019. URL <https://openreview.net/forum?id=BJ1xm30cKm>.

791

792 Liyuan Wang, Xingxing Zhang, Hang Su, and Jun Zhu. A comprehensive survey of continual learning:  
 793 Theory, method and application. *IEEE Transactions on Pattern Analysis and Machine Intelligence*,  
 794 46(8):5362–5383, 2024. doi: 10.1109/TPAMI.2024.3367329.

795

796 Zhichao Wang, Bin Bi, Zixu Zhu, Xiangbo Mao, Jun Wang, and Shiyu Wang. Uft: Unifying  
 797 fine-tuning of sft and rlhf/dpo/una through a generalized implicit reward function, 2025. URL  
 798 <https://arxiv.org/abs/2410.21438>.

799

800 Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H Chi, Quoc  
 801 Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models.  
 802 *arXiv preprint arXiv:2201.11903*, 2022.

803

804 Tongtong Wu, Massimo Caccia, Zhuang Li, Yuan-Fang Li, Guilin Qi, and Gholamreza Haffari.  
 805 Pretrained language model in continual learning: A comparative study. In *International Conference*  
 806 on *Learning Representations*, 2022. URL <https://openreview.net/forum?id=figzpGMrdD>.

807

808 Prateek Yadav, Derek Tam, Leshem Choshen, Colin Raffel, and Mohit Bansal. TIES-merging:  
 809 Resolving interference when merging models. In *Thirty-seventh Conference on Neural Information*  
 810 *Processing Systems*, 2023. URL <https://openreview.net/forum?id=xtaX3WyCj1>.

810 Prateek Yadav, Tu Vu, Jonathan Lai, Alexandra Chronopoulou, Manaal Faruqui, Mohit Bansal,  
 811 and Tsendsuren Munkhdalai. What matters for model merging at scale? *arXiv preprint*  
 812 *arXiv:2410.03617*, 2024.

813  
 814 An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li,  
 815 Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin  
 816 Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang,  
 817 Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tingyu Xia,  
 818 Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu  
 819 Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*,  
 820 2024a.

821 An Yang, Beichen Zhang, Binyuan Hui, Bofei Gao, Bowen Yu, Chengpeng Li, Dayiheng Liu,  
 822 Jianhong Tu, Jingren Zhou, Junyang Lin, Keming Lu, Mingfeng Xue, Runji Lin, Tianyu Liu,  
 823 Xingzhang Ren, and Zhenru Zhang. Qwen2.5-math technical report: Toward mathematical expert  
 824 model via self-improvement. *arXiv preprint arXiv:2409.12122*, 2024b.

825 Yixin Ye, Zhen Huang, Yang Xiao, Ethan Chern, Shijie Xia, and Pengfei Liu. Limo: Less is more for  
 826 reasoning, 2025. URL <https://arxiv.org/abs/2502.03387>.

827 Çağatay Yıldız, Nishaanth Kanna Ravichandran, Prishruit Punia, Matthias Bethge, and Beyza Ermis.  
 828 Investigating continual pretraining in large language models: Insights and implications. *arXiv*  
 829 *preprint arXiv:2402.17400*, 2024.

830 Çağatay Yıldız, Nishaanth Kanna Ravichandran, Nitin Sharma, Matthias Bethge, and Beyza  
 831 Ermis. Investigating continual pretraining in large language models: Insights and implica-  
 832 tions. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856. URL <https://openreview.net/forum?id=aKjJoEVKgO>.

833  
 834 Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. Hellaswag: Can a machine  
 835 really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for*  
 836 *Computational Linguistics*, 2019.

837  
 838  
 839  
 840  
 841  
 842  
 843  
 844  
 845  
 846  
 847  
 848  
 849  
 850  
 851  
 852  
 853  
 854  
 855  
 856  
 857  
 858  
 859  
 860  
 861  
 862  
 863

864

865

866

867

# Appendix

868

869

## CONTENTS

870

871

### A Related Works 19

872

873

### B Experimental Setup 20

874

#### B.1 Evaluation . . . . . 20

875

#### B.2 Datasets . . . . . 21

876

### C Evaluation Methodology 24

877

#### C.1 Prompting . . . . . 24

878

#### C.2 Sample vs. Aggregate Metric Comparison . . . . . 25

879

#### C.3 Metric Robustness under MCQA . . . . . 25

880

#### C.4 Extraction . . . . . 28

881

### D Model Merging 30

882

#### D.1 Weight Drift . . . . . 30

883

#### D.2 Failure Case: OpenThinker3 . . . . . 30

884

#### D.3 Failure Case: Coder Models . . . . . 32

885

#### D.4 Moderate Case: OpenThinker . . . . . 33

886

### E Expanded Comparisons 34

887

#### E.1 Data Diversity . . . . . 34

888

#### E.2 Objective Function (SFT vs. RL) . . . . . 35

889

#### E.3 Data Volume . . . . . 36

890

### F Additional Plots 38

891

### G Quantifying Forgetting Accurately (Tables for referencing plots)<sup>3</sup> 39

892

#### G.1 Instruction Tuning . . . . . 39

893

#### G.2 Domain-Continual Pretraining . . . . . 41

894

#### G.3 Trained from Base . . . . . 43

895

#### G.4 Trained from Instruct - High Data Scenario . . . . . 45

896

#### G.5 Trained from Instruct - Low Data Scenario . . . . . 47

897

#### G.6 Qwen2.5 Base and Coder Merge (Relative to Qwen2.5 Base) . . . . . 49

898

#### G.7 Qwen2.5 Base and Coder Merge (Relative to Qwen2.5 Coder) . . . . . 50

899

#### G.8 Qwen2.5 Instruct and OpenThinker 7B Merge (Relative to Qwen2.5 Instruct) . . . . . 51

900

#### G.9 Qwen2.5 Instruct and OpenThinker 7B Merge (Relative to OpenThinker) . . . . . 52

918	G.10 Qwen2.5 Instruct and OpenThinker3 7B Merge (Relative to Qwen2.5 Instruct) . . .	53
919		
920	G.11 Qwen2.5 Instruct and OpenThinker3 7B Merge (Relative to OpenThinker3) . . .	54
921		
922		
923		
924		
925		
926		
927		
928		
929		
930		
931		
932		
933		
934		
935		
936		
937		
938		
939		
940		
941		
942		
943		
944		
945		
946		
947		
948		
949		
950		
951		
952		
953		
954		
955		
956		
957		
958		
959		
960		
961		
962		
963		
964		
965		
966		
967		
968		
969		
970		
971		

972 A RELATED WORKS  
973

974 **Post-training techniques.** A broad set of post-training methods now underpins standard LLM  
975 pipelines. *Supervised fine-tuning (SFT)* (Ouyang et al., 2022) remains the core step, used for continued  
976 pre-training and instruction tuning. At later stages, *reinforcement learning from human feedback*  
977 (*RLHF*) (Ouyang et al., 2022) aligns model outputs with human preferences. To simplify preference  
978 learning, *direct preference optimization (DPO)* (Rafailov et al., 2023) provides a direct loss surrogate.  
979 With the rise of test-time scaling (e.g. sampling depth or compute at inference), *group relative policy*  
980 *optimization (GRPO)* (Shao et al., 2024) has been proposed to elicit stronger intrinsic reasoning.  
981 Taken together, these methods introduce distinct objectives and optimizers, increasing the complexity  
982 of the post-training stack (Wang et al., 2025).

983 **Measuring catastrophic forgetting.** Catastrophic forgetting is the loss of previously acquired  
984 knowledge when a network learns new information. Early studies examined the effect in small models  
985 and simplified settings (McCloskey & Cohen, 1989; Ratcliff, 1990; French, 1999). Lopez-Paz &  
986 Ranzato (2017) formalized forgetting via *backward transfer*, the effect of learning a new task on  
987 performance in earlier ones: positive values indicate improvement; negative values indicate forgetting.  
988 Recent work extends these analyses to deep networks trained on large-scale data, with growing  
989 attention to language models (Biesialska et al., 2020; Wu et al., 2022).

990 **Benchmark paradigm.** Task-incremental learning is the dominant paradigm for benchmarking  
991 forgetting (De Lange et al., 2022). Models learn a sequence of tasks with clear boundaries, and task  
992 labels are available at train and test time. Class-incremental learning removes test-time task identifiers,  
993 making evaluation stricter (Wang et al., 2024). Other views analyze continual learning through  
994 positive/negative transfer (Yildiz et al., 2025). At the sample level, Toneva et al. (2019) introduced  
995 forgetting metrics that identify “unforgettable” examples (stable once learned) and “catastrophically  
996 forgotten” examples (highly plastic), and showed these patterns are consistent across architectures  
997 and random seeds.

998 **Language-model forgetting.** Recent studies focus on forgetting induced by instruction tuning.  
999 Luo et al. (2025) trained models up to 7B parameters with SFT and evaluated multiple knowledge  
1000 categories. DeepSeek-AI (2024) reported instruction-tuning-related regressions on sentence com-  
1001 pletion even for 67B models. Fernando et al. (2025) examined forgetting across SFT followed by  
1002 RLHF and proposed joint-training strategies to mitigate it. Lin et al. (2024) framed instruction-tuning  
1003 degradation as an “alignment tax” (performance loss on pre-training skills due to alignment) and  
1004 found model merging to be the most Pareto-efficient mitigation among tested techniques. Li &  
1005 Lee (2024b) studied continual pre-training on aligned LMs and observed notable regressions in  
1006 alignment-related behavior.

1007 **Catastrophic forgetting in reasoning training pipelines.** Work on reasoning-oriented LMs high-  
1008 lights new failure modes. Li et al. (2025) defined *temporal forgetting*: models lose the ability to solve  
1009 problems they could solve at earlier training checkpoints. The effect appears in both RL-trained and  
1010 instruction-tuned models. They proposed *temporal sampling*—round-robin sampling from recent  
1011 checkpoints—as a mitigation. Pipatanakul et al. (2025) merged a language-fine-tuned model with  
1012 DeepSeek R1 Distill (70B; both derived from Llama 3.3 70B (Dubey et al., 2024)) to adapt reasoning  
1013 while preserving language competence. For multimodal models, Chen et al. (2025) found that later  
1014 layers primarily support reasoning, whereas early layers concentrate perception, suggesting layer-wise  
1015 interventions. We document forgetting extensively across post-training pipelines in our work.

1016 Each new method introduces its own objective and optimization procedure, adding to the complexity  
1017 of the post-training landscape (Wang et al., 2025).

1018 **Mitigation strategies.** Sequential SFT to RLHF/DPO can exacerbate forgetting. To counteract this,  
1019 researchers explore: (i) *model averaging*, interpolating between pre- and post-RLHF checkpoints to  
1020 trade off alignment and retention (Lin et al., 2024); (ii) *joint post-training*, optimizing supervised and  
1021 preference objectives simultaneously with convergence guarantees (Fernando et al., 2024); and (iii)  
1022 *unified fine-tuning (UFT)*, which folds instruction tuning and alignment into a single implicit-reward  
1023 objective (Wang et al., 2025). Additional techniques—including advantage models and selective  
1024 rehearsal—stabilize RLHF by shaping reward distributions and replaying curated data (Peng et al.,  
1025

1026 2023). *Online Merging Optimizers (OMO)* combine gradients from SFT and RLHF models during  
 1027 training to maximize reward while preserving pre-trained skills (Lu et al., 2024). Theory supports  
 1028 these interventions: up to permutation symmetries, weights of homologous models tend to lie in a  
 1029 shared low-loss basin (Ainsworth et al., 2023). Hence, we were quite surprised that model merging  
 1030 does not work for our simple case of mitigating forgetting during post-training with only two deep  
 1031 networks.

1032  
 1033 **Forgetting at scale.** Pre-training mitigates forgetting relative to training from scratch (Mehta et al.,  
 1034 2023; McRae & Hetherington, 1993). Ramasesh et al. (2022) further found that pretrained ResNets  
 1035 and Transformers (up to  $\sim 100M$  parameters) are robust to forgetting at scale; language experiments  
 1036 showed similar trends. However, Luo et al. (2025) reported increased forgetting with scale in the  
 1037 1–7B LM regime, suggesting modality- and regime-dependent behavior. In contrast to these works,  
 1038 we study forgetting during post-training of language models.

## 1040 B EXPERIMENTAL SETUP

1041  
 1042 We standardize settings across models for fair comparison. All experiments use the `LightEval`  
 1043 framework (Habib et al., 2023) and log per-sample accuracy. We apply a zero-shot chain-of-thought  
 1044 prompt to all models and require answers in a fixed MCQ format (see Appendix); base models receive  
 1045 a few-shot prompt solely to teach the format. When available<sup>6</sup>, we add chat-specific templates to be  
 1046 in line with best practices. We cap sequence length at 32K tokens, except for Qwen2.5-7B-Math and  
 1047 Qwen2.5-7B-Math-Instruct Yang et al. (2024a), which are limited to 4K<sup>7</sup>. Decoding uses temperature  
 1048 0.6 with nucleus sampling (`top_p`) of 0.95. We provide additional details in the Appendix. To  
 1049 facilitate reproducibility and further inquiry, we release per-sample logs for every sub-benchmark  
 1050 alongside code.

### 1051 B.1 EVALUATION

1052 To evaluate performance differences between models, we employ chain-of-thought (CoT) prompting  
 1053 Wei et al. (2022) on multiple-choice question answering (MCQA) datasets in addition to free-  
 1054 form/generative math questions. In this setup, the model auto-regressively generates a reasoning  
 1055 chain prior to producing its final answer. The predicted choice is then extracted from the generated text  
 1056 and compared against the ground-truth label. When available, chat-specific templates are incorporated  
 1057 into the prompt to ensure consistent formatting.

1058 Because some models, particularly base models, tend to continue generating responses for subsequent  
 1059 questions after completing the current one, we provide explicit stop sequences to terminate generation  
 1060 once a prediction has been produced.

1061 When applicable, to encourage answers in strict MCQA format (models sometimes output the option  
 1062 text instead of the letter), we prepend the following instruction prompt:

1063 {Instruction}

1064 On the very last line, write exactly "Answer: \$LETTER" (e.g.  
 1065 "Answer: B"), with no extra punctuation, no lowercase, no \*,  
 1066 and no trailing spaces.

1067 Think step by step, showing your reasoning.

1068 Question: "{Question}"

1069 We find adding the additional instructions to not use extra punctuation, asterisks, lowercase letters, or  
 1070 trailing spaces necessary, as we find `LightEval`'s letter extraction can fail in certain cases otherwise.  
 1071 We additionally tell the model to constrain its output to be of the form "Answer: \$LETTER" as  
 1072 otherwise models will often provide the corresponding answer to a given letter or provide the answer  
 1073 in another format, making extraction more prone to error.

1074  
 1075  
 1076  
 1077  
 1078 <sup>6</sup>This budget was sufficient in practice; we never required more tokens.

1079 <sup>7</sup>Because base models sometimes continue into subsequent questions, we set explicit stop sequences to end  
 1080 generation once a prediction is produced.

1080 For the case of base models, where few-shot prompting yields a more accurate elicitation of their  
 1081 knowledge, we use few-shot prompting:  
 1082

```
1083 {Instruction}
1084
1085 Question: "{Few-Shot Question 1}"
1086 Reasoning: {Few-shot Reasoning Trace 1}
1087 Answer: {Few-shot Answer 1}
1088 ...
1089 ... <--- more examples
1090
1091 Question: "{Question}"
1092 Reasoning:
```

1093 Datasets where CoT reasoning traces are provided for few-shot prompting, we use those. In the  
 1094 cases where this is not provided (PIQA, MCTest, Social-IQa, ARC, MCTest, and Hellaswag) CoT  
 1095 few-shot examples were generated and then confirmed these are not included in the benchmarks<sup>1</sup>.  
 1096 For free-form/generative math questions, we follow the prompt and extraction methods used in  
 1097 Hochlehnert et al. (2025).

1098 All experiments are conducted using the Hugging Face LightEval framework, with results logged  
 1099 at the sample level. For generation, we allow up to 32,768 tokens, which we found sufficient for  
 1100 models to complete their chain of thought and provide an answer. In cases where the maximum  
 1101 trained context length is smaller, then the generation is reduced to that number, as is the case with  
 1102 Qwen2.5-7B-Math and Qwen2.5-7B-Math-Instruct Yang et al. (2024a). The temperature is set to 0.6  
 1103 and nucleus sampling with  $p = 0.95$  is applied. All datasets are evaluated on at least 3 seeds and  
 1104 metrics reported with mean and standard deviation (c.f. Section G).

1105 **B.2 DATASETS**  
 1106

1107 To evaluate broad model knowledge and capabilities, we benchmark on eighteen public datasets:  
 1108 MMLU Hendrycks et al. (2021b;a), BBH Suzgun et al. (2022), GPQA Rein et al. (2024), MuSR  
 1109 Sprague et al. (2024), ARC Clark et al. (2018), TruthfulQA Lin et al. (2022), HellaSwag Zellers  
 1110 et al. (2019), Social IQa Sap et al. (2019), MCTest Richardson et al. (2013), PIQA Bisk et al.  
 1111 (2020), CommonsenseQA Talmor et al. (2019), SaladBench Li et al. (2024), AIME24, AIME25,  
 1112 AMC23, Math500, Minerva, and OlympiadBench. Several of these benchmarks, namely MMLU and  
 1113 BBH provide subject-level annotations, enabling fine-grained sub-benchmark analyses in addition to  
 1114 aggregate reporting. For the cases of MMLU and BBH, subcategory labels are provided which allow  
 1115 for splitting into further sub-benchmark evaluates by subjects. To enable easier understanding, we  
 1116 group these (sub-)benchmarks into high-level groups used to evaluate the capabilities of the models.  
 1117 They are grouped such that (sub-)benchmarks in the same group show similar trends in forgetting  
 1118 and improvement.

1119 They are grouped as follows:

1120 **Commonsense:**

1121

- 1122 • Commonsense QA
- 1123 • PIQA

1124 **Culture:**

1125

- 1126 • BBH (sports understanding and movie recommendation)

1127 **Logic**

1128

- 1129 • BBH (navigate, causal judgment, penguins in a table, web of lies, tracking shuffled objects  
 1130 three objects, tracking shuffled objects seven objects, tracking shuffled objects five ob-  
 1131 jects, temporal sequences, reasoning about colored objects, logical deduction three objects,  
 1132 logical deduction seven objects, logical deduction five objects, formal fallacies, and date  
 1133 understanding)

1134     • ARC (easy and challenge)  
 1135     • MuSR (murder mysteries, object placements, and team allocation)  
 1136     • MMLU (logical fallacies)

1137     **Knowledge**

1138     • BBH (object counting)  
 1139     • MMLU (miscellaneous and global facts)  
 1140     • MCTest

1141     **Language**

1142     • BBH (snarks, disambiguation qa, ruin names, and hyperbaton)  
 1143     • Social IQa  
 1144     • Hellaswag  
 1145     • BBH (salient translation error detection)

1146     **Liberal Arts**

1147     • MMLU (world religions, us foreign policy, sociology, security studies, public relations, professional psychology, professional law, prehistory, philosophy, management, international law, high school world history, high school us history, high school psychology, high school microeconomics, high school macroeconomics, high school government and politics, high school geography, and high school european history)

1148     **Math**

1149     • BBH (geometric shapes, and boolean expressions)  
 1150     • MMLU (high school statistics, high school mathematics, formal logic, elementary mathematics, econometrics, college mathematics, and abstract algebra)

1151     **Math (Generative)**

1152     • AIME24  
 1153     • AIME25  
 1154     • AMC23  
 1155     • Math500  
 1156     • Minerva  
 1157     • OlympiadBench

1158     **Safety**<sup>2</sup>

1159     • MMLU (moral scenarios, moral disputes, jurisprudence, and business ethics)  
 1160     • TruthfulQA (mc1)  
 1161     • SaladBench (mrq)

1162     **Science & Tech**

1163     • MMLU (marketing, virology, professional medicine, professional accounting, nutrition, medical genetics, machine learning, human sexuality, human aging, high school physics, high school computer science, high school chemistry, high school biology, electrical engineering, conceptual physics, computer security, college physics, college medicine, college computer science, college chemistry, college biology, clinical knowledge, astronomy, and anatomy)  
 1164     • GPQA (diamond)

1188 Unless otherwise noted, we follow the standard task formats and official evaluation splits; for Truth-  
1189 fulQA we report MC1, for GPQA the *Diamond* subset, and for SaladBench the MRQ configuration.  
1190 This taxonomy serves as the backbone for our analyses of capability acquisition and retention across  
1191 training and deployment.

1192

1193

1194

1195

1196

1197

1198

1199

1200

1201

1202

1203

1204

1205

1206

1207

1208

1209

1210

1211

1212

1213

1214

1215

1216

1217

1218

1219

1220

1221

1222

1223

1224

1225

1226

1227

1228

1229

1230

1231

1232

1233

1234

1235

1236

1237

1238

1239

<sup>1</sup>MMLU is evaluated with few-shot no CoT prompting for the base models

1240

1241

<sup>2</sup>These are only used in comparisons which do not include a base model because TruthfulQA and SaladBench are designed measure the default behavior of the model rather than knowledge, which few-shot prompting would bias.

1242 C EVALUATION METHODOLOGY  
12431244 C.1 PROMPTING  
1245

1246 In additional tests, we measure the ability of base models using the same prompting as instruction-  
1247 tuned models. Under these conditions we see ostensibly large forgetting in domain-continual  
1248 pretrained models (Figure 9). Our qualitative analysis suggests that this is largely due to the models  
1249 outputting code, wherein the location of the answer can be obscured. When this is contrasted with  
1250 the few-shot prompting, where there is much less forgetting, we conclude that forgetting metrics can  
1251 vary significantly depending on the way knowledge is elicited, especially when training on narrow  
1252 tasks, which few-shot prompting alleviates.

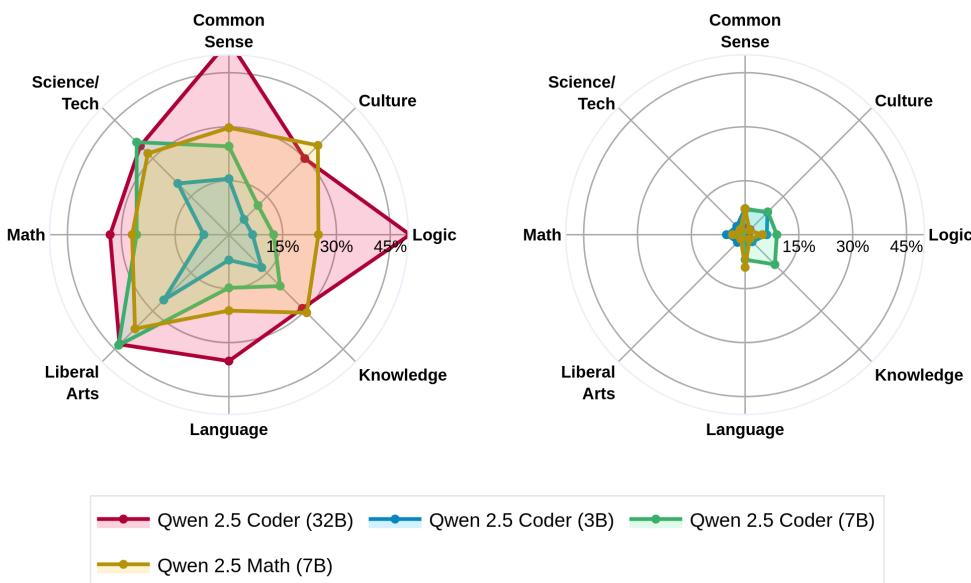
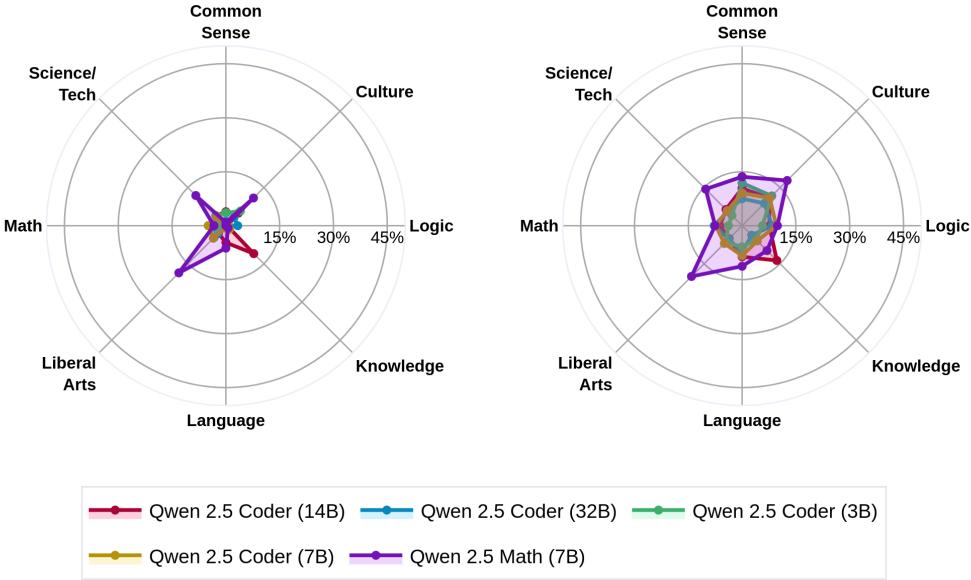


Figure 8: Domain-Adaptive Pretraining models using chat template prompting

For these reasons, measuring the performance of base models on behavioral evaluations can become nontrivial. While benchmarks measuring knowledge or capabilities may be elicited through few-shot prompting, others, such as truthfulness or safety become more difficult as prompting them with examples would bias their behavior. Further works should consider exploring the effect of providing no-knowledge few-shot prompting, where the format of the question and answer is provided without leaking examples to avoid biasing the base model’s output.

1296 C.2 SAMPLE VS. AGGREGATE METRIC COMPARISON  
12971318 **Figure 9: Coder model comparing conventional forgetting (left) against our sample-wise forget-  
1319 ting (right).** More forgetting is uncovered when using the sample-wise forgetting metric.  
13201321 The sample-wise nature of our introduced metric uncovers more forgetting than the standard metric,  
1322 defined as  
1323

$$F_{\text{standard}} = \max(\bar{a}^{\text{pre}} - \bar{a}^{\text{post}}, 0)$$

1325 Figure 9 illustrates this effect: our sample-based metric reveals substantially more forgetting relative  
1326 to the standard formulation, in some cases finding what was originally low forgetting is actually  
1327 moderate. This highlights sample-level degradation that is otherwise hidden when averaging over  
1328 tasks.1330 C.3 METRIC ROBUSTNESS UNDER MCQA  
13311332 We review the robustness of the chance adjusted forgetting in measuring true knowledge loss, which is  
1333 particularly relevant when evaluating under MCQA benchmarks as models can often guess the answer  
1334 correctly. While our metric accounts for this by subtracting out an estimate of this probability, we  
1335 compare this to another sample-level metric which filters out noisy samples to empirically demonstrate  
1336 this. Namely, we consider samples where there is agreement among two out of three seeds on average,  
1337 and consider the cases where there is forgetting relative to the other cases.

$$F = \frac{(1 \rightarrow 0)_2}{(0 \rightarrow 0)_2 + (0 \rightarrow 1)_2 + (1 \rightarrow 0)_2 + (1 \rightarrow 1)_2}.$$

1341 where we formally define the two-seed sample-agreement metric as follows:  
1342

$$1343 (x \rightarrow y)_2 := \frac{1}{\binom{|\mathcal{S}|}{2}} \sum_{\{s,t\} \in \binom{\mathcal{S}}{2}} \sum_{i=1}^N \mathbf{1}\{a_{i,s}^{\text{pre}} = x = a_{i,t}^{\text{pre}} \wedge a_{i,s}^{\text{post}} = y = a_{i,t}^{\text{post}}\}.$$

1347 where  $\mathcal{S}$  is the set of seeds and  $N$  the number of samples. Intuitively, this measures robust knowledge  
1348 loss relative to stable knowledge. We find that this metric agrees with our results across the post-  
1349 training pipeline, thereby indicating the chance adjusted forgetting metric captures robust knowledge  
loss, rather than random forgetting. We show the correlation between these metrics in Figure 10.

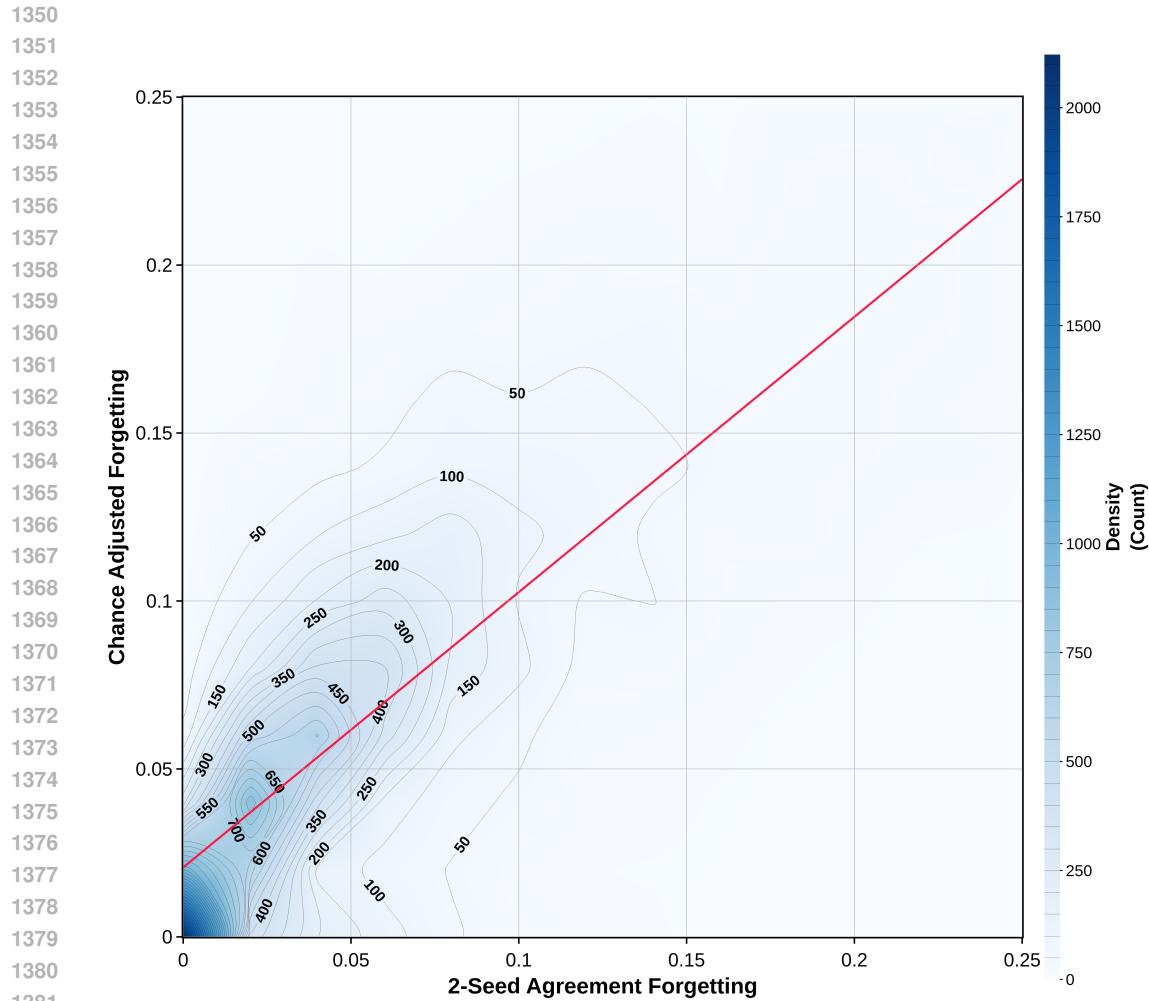


Figure 10: **Sub-benchmark count across 2-Seed Agreement Forgetting (x-axis) and Chance Adjusted Forgetting (y-axis).** The line of best fit (red) of the subbenchmark forgetting values shows both metrics are highly correlated.

We also provide a specific comparison of the metrics in Figure 11.

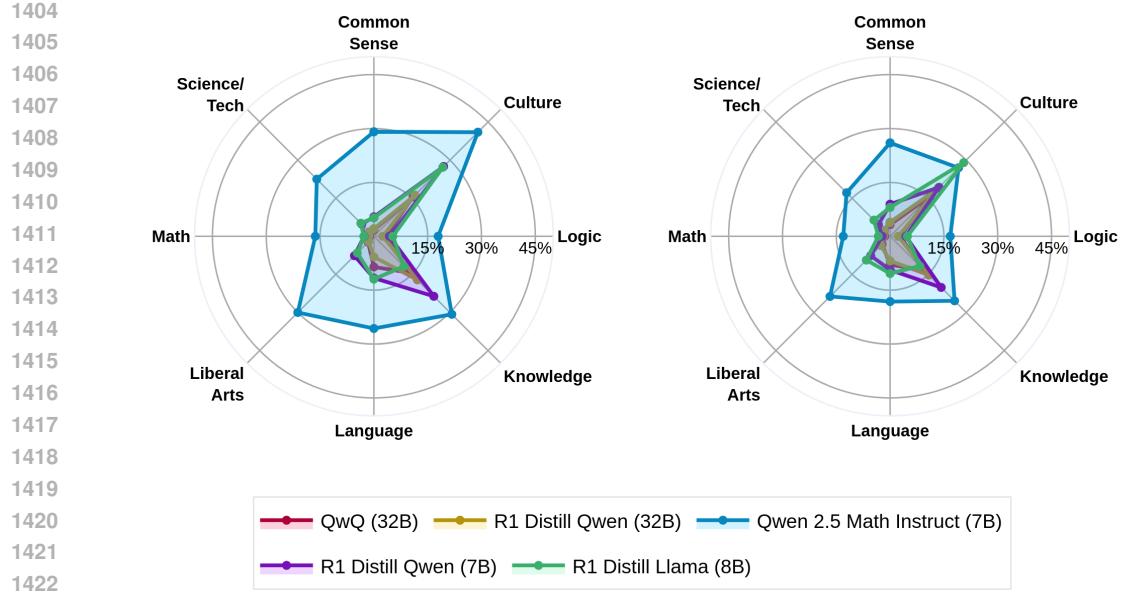


Figure 11: **2 Seed Forgetting (left) and Chance Adjusted Forgetting (right) of Models Trained from Base.** The overall trends remain the same, with only minor differences such as a slight increase in forgetting for Qwen2.5 Math Instruct (7B) in the 2 seed forgetting case.

1427  
1428  
1429  
1430  
1431  
1432  
1433  
1434  
1435  
1436  
1437  
1438  
1439  
1440  
1441  
1442  
1443  
1444  
1445  
1446  
1447  
1448  
1449  
1450  
1451  
1452  
1453  
1454  
1455  
1456  
1457

1458 C.4 EXTRACTION  
1459

1460 Sample-level inspection occasionally shows answers that are correct, but unable to be extracted  
1461 correctly through the regex extraction described in Section B.1. This occurs particularly in models  
1462 trained for specific tasks. For example, we find that coding models will assign the correct answer to  
1463 a variable in code and then in a print statement provide the answer variable. We control for these  
1464 extraction related errors by using LLM extractor, specifically Qwen2.5-14B-Instruct, which we find  
1465 sufficient to correct for errors. We do this by providing the question, response, and ground truth  
1466 answers using the below prompt without a chat template, in order to encourage immediate json  
1467 output:

1468

1469 You are a strict extractor.

1470 Given the FULL\_PROMPT (the original prompt to the model), the model  
1471 output SNIPPET (last part, quoted), and the gold extraction (quoted),  
1472 return ONLY valid JSON with exactly two keys:

- 1473 - "extraction\": the final answer token as a string  
1474 (e.g. "A", "C", "42") or null if unknown
- 1475 - "correct": true if the extraction matches the gold, false  
1476 if it does not, or null if unknown

1477 Do NOT output anything else (no explanation, no code fences).

1478

1479 FULL\_PROMPT: {q\_full\_prompt}

1480 SNIPPET: {q\_snippet}

1481 GOLD: {q\_gold}

1482

1483 Return JSON now:

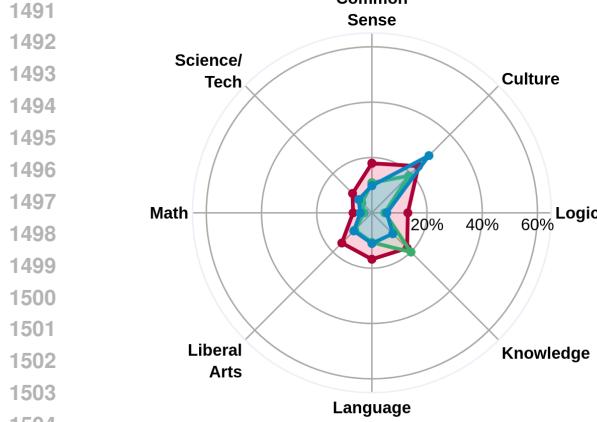
1484

1485 We find that this primarily corrects for outliers while all trends remain the same, which we find to  
1486 be true across all knowledge and post-training categories. By comparing Figures 12, where LLM  
1487 extraction is used, and 13, where regex extraction is used, we see the outlier Qwen2.5 Math Instruct  
(7B) is reduced in its outlier effect. Additionally, the overall trends remain the same.

1488

1489

1490



1491

1492

1493

1494

1495

1496

1497

1498

1499

1500

1501

1502

1503

1504

1505

1506

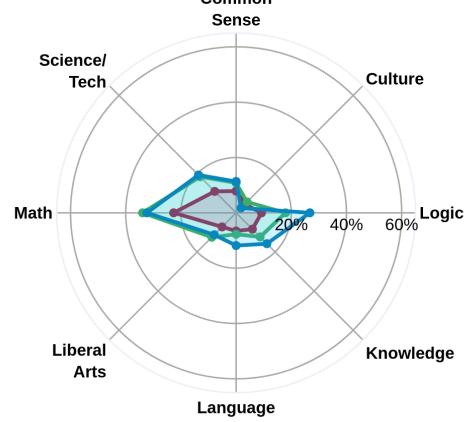
1507

1508

1509

1510

1511



1512 Figure 12: **Forgetting (left) and Backward Transfer (right) of Models Trained from Base**  
1513 **using LLM Extraction.** Trends are the same as in Figure 13, but Qwen2.5-Math-Instruct's outlier  
1514 tendencies are reduced.

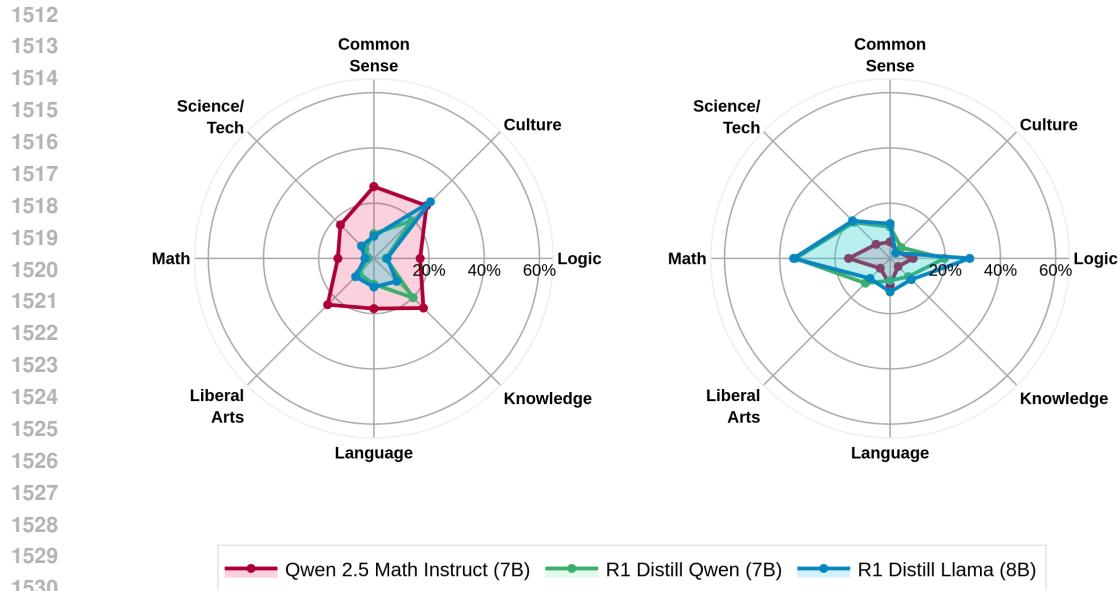


Figure 13: **Forgetting (left) and Backward Transfer (right) of Models Trained from Base using Regex Extraction.** Trends are the same as in Figure 12, but Qwen2.5-Math-Instruct's outlier tendencies are increased.

1535  
1536  
1537  
1538  
1539  
1540  
1541  
1542  
1543  
1544  
1545  
1546  
1547  
1548  
1549  
1550  
1551  
1552  
1553  
1554  
1555  
1556  
1557  
1558  
1559  
1560  
1561  
1562  
1563  
1564  
1565

1566  
1567

## D MODEL MERGING

1568  
1569  
1570  
1571

Recent work shows that offline model merging can combine capabilities from multiple models (Dzidzio et al., 2025). Unlike classical continual learning (De Lange et al., 2022), it requires neither the original training data nor the ability to resume training, which is practical in resource-constrained settings.

1572  
1573  
1574  
1575

**Setup.** We evaluate Exponential Moving Average (EMA) merging; in the two-checkpoint case this is linear interpolation,

$$\theta_{\text{EMA}}(\alpha) = \alpha \theta_{\text{pre}} + (1 - \alpha) \theta_{\text{post}}.$$

1576  
1577

## D.1 WEIGHT DRIFT

1578  
1579  
1580  
1581  
1582

We observe large weight drift among models in which merging fails. Specifically in the case of model trained from instruction tuned bases, we compute the ratio of the  $L_2$  norm of the task vector (from the model to the instruct model) to the  $L_2$  norm of base. In the case of OpenThinker3 this is just above 20%. Likewise Qwen2.5 Coder (7B) has a value of 87%. OpenThinker and s1.1, which we find are both mergeable, have values of only 1.8% and 0.6% resp.

1583  
1584

## D.2 FAILURE CASE: OPENTHINKER3

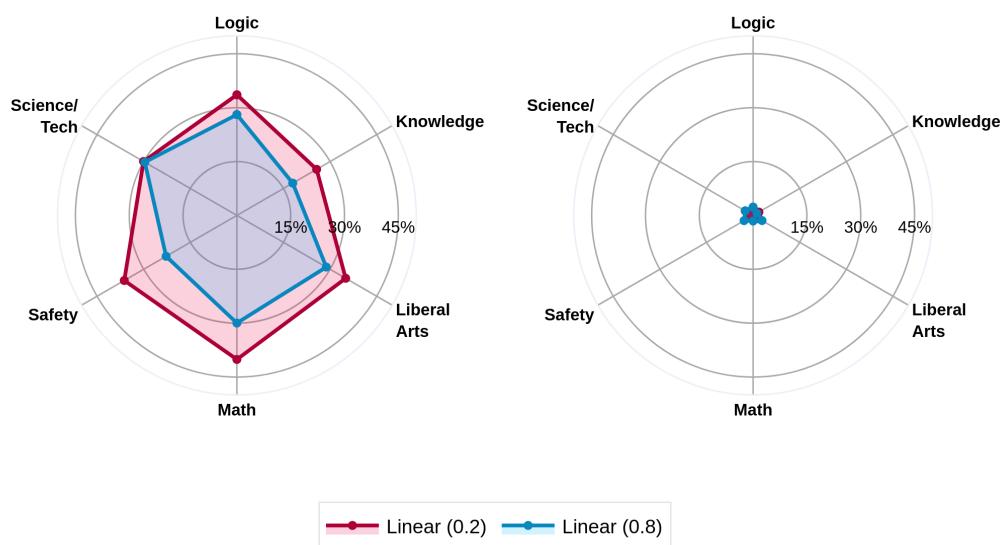
1585  
15861605  
1606  
1607  
1608  
1609

Figure 14: **Forgetting** (left) and **Backward Transfer** (right) of Qwen 2.5 Instruct merged with OpenThinker3 (7B) relative to Qwen 2.5 Instruct on MMLU. Large forgetting occurs. Sample-level analysis shows the model output degeneration, with the model often repeating words or phrases, typically without providing a final answer.

1610  
1611  
1612  
1613  
1614  
1615  
1616  
1617  
1618  
1619

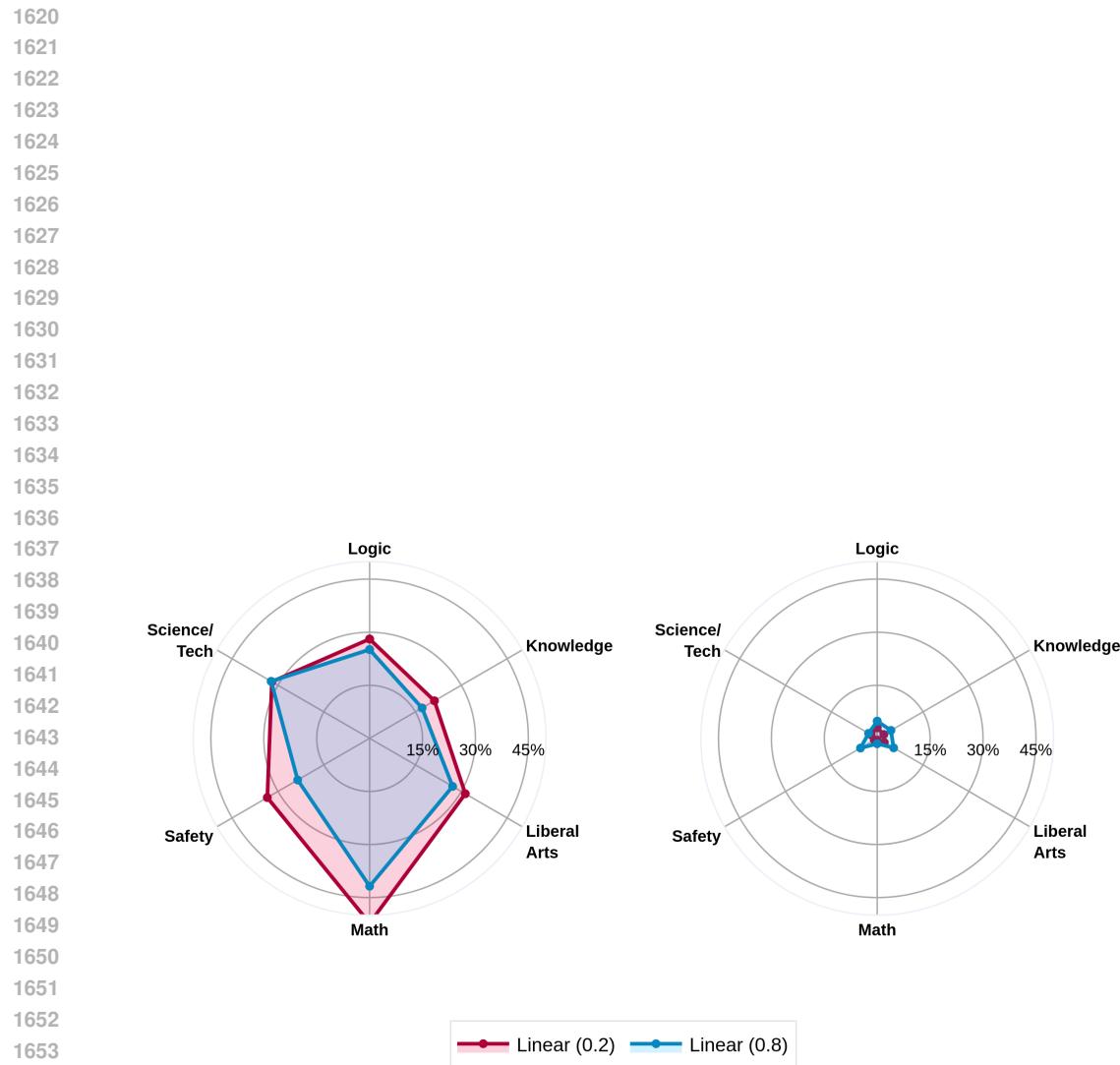


Figure 15: **Forgetting (left) and Backward Transfer (right) of Qwen 2.5 Instruct merged with OpenThinker3 (7B) relative to OpenThinker3 on MMLU.** Large forgetting occurs. Sample-level analysis shows the model output degeneration, with the model often repeating words or phrases, typically without providing a final answer.

1674 D.3 FAILURE CASE: CODER MODELS  
1675

1676

1677

1678

1679

1680

1681

1682

1683

1684

1685

1686

1687

1688

1689

1690

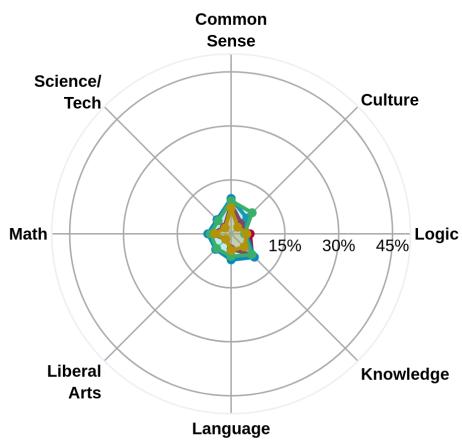
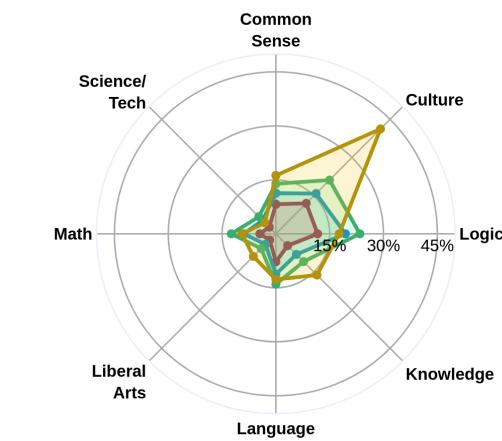
1691

1692

1693

1694

1695



—●— Linear (0.2) —●— Linear (0.8) —●— Slerp (0.2) —●— Slerp (0.8)

1696 **Figure 16: Forgetting (left) and Backward Transfer (right) of Qwen 2.5 Base merged with Qwen  
1697 2.5 Coder (7B) relative to Qwen 2.5 Base on all Benchmarks.** Moderate-to-large forgetting occurs  
1698 with low backward transfer.

1699

1700

1701

1702

1703

1704

1705

1706

1707

1708

1709

1710

1711

1712

1713

1714

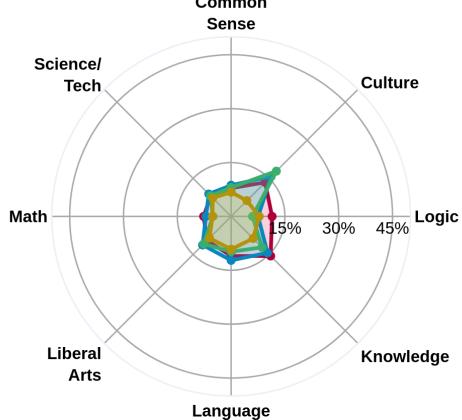
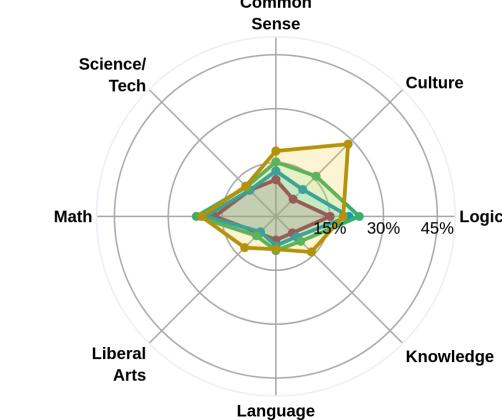
1715

1716

1717

1718

1719



—●— Linear (0.2) —●— Linear (0.8) —●— Slerp (0.2) —●— Slerp (0.8)

1720 **Figure 17: Forgetting (left) and Backward Transfer (right) of Qwen 2.5 Base merged with Qwen  
1721 2.5 Coder (7B) relative to Qwen 2.5 Coder on all Benchmarks.** Moderate-to-large forgetting occurs  
1722 with low-to-moderate backward transfer.

1723

1724

1725

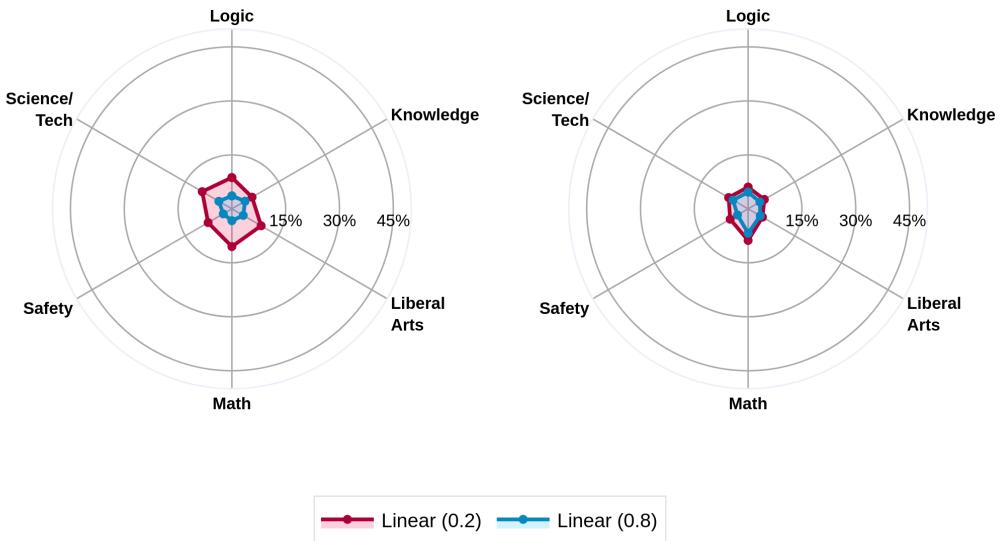
1726

1727

1728

1729

#### D.4 MODERATE CASE: OPENTHINKER

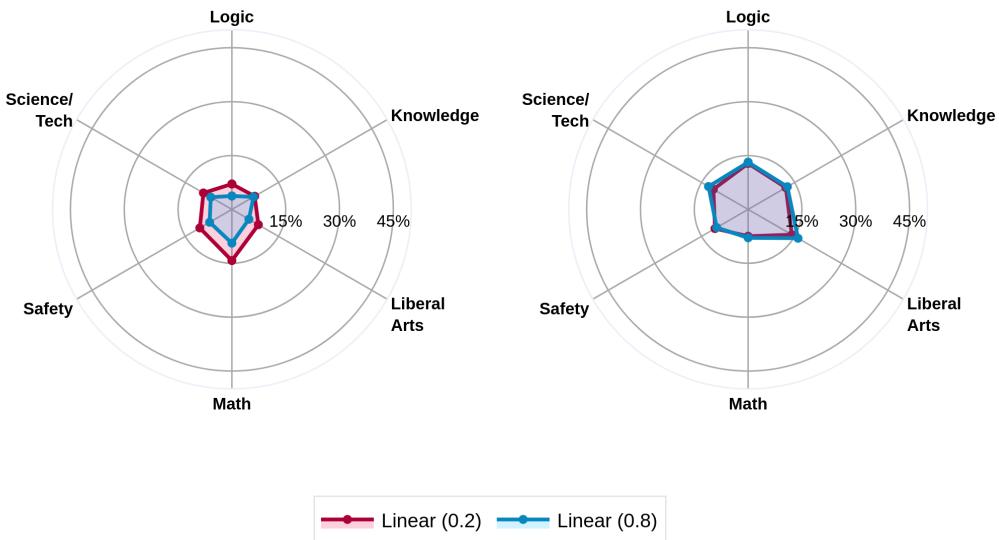


1750

1751

Figure 18: Forgetting (left) and Backward Transfer (right) of Qwen 2.5 Instruct merged with OpenThinker Merge (7B) relative to Qwen 2.5 Instruct on MMLU. We see a marginal overall performance improvement in the case of Linear (0.8).

1753



1774

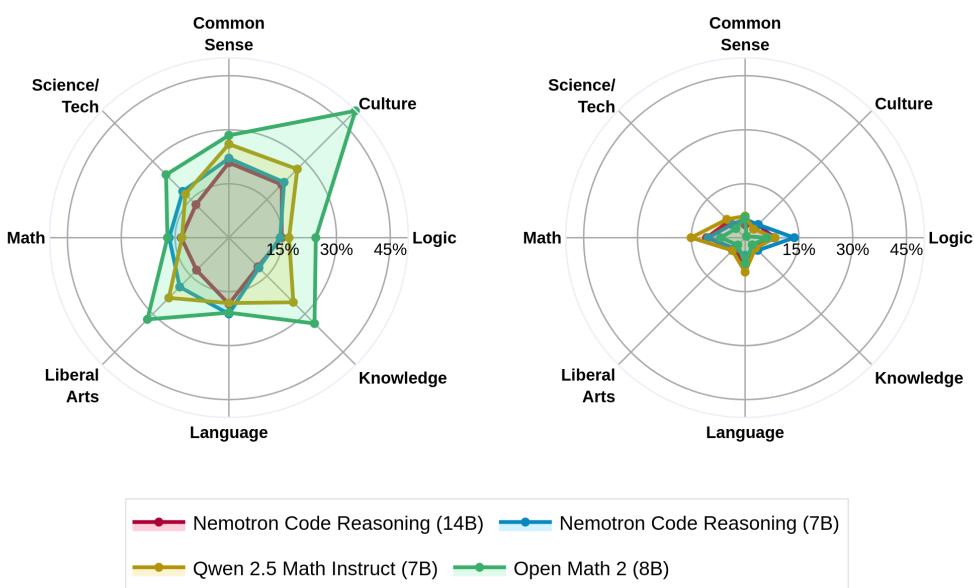
**Figure 19: Forgetting (left) and Backward Transfer (right) of Qwen 2.5 Instruct merged with OpenThinker (7B) relative to OpenThinker on MMLU.** We see a marginal overall performance improvement in both cases.

1777

1778

1779

1780

1782 E EXPANDED COMPARISONS  
17831784  
1785  
1786  
1787  
1788  
1789  
1790  
1791 E.1 DATA DIVERSITY  
17921793  
1794  
1795  
1796  
1797  
1798 We split reasoning models into the cases of being trained on narrow domains, where they are trained  
1799 on one or two benchmark categories (e.g. math or code), and mixed data, where they are jointly  
1800 trained on many tasks or on general data. As indicated in Section 3.3.3, increased data diversity  
1801 generally mitigates forgetting and helps with backward transfer (Figure 21) whereas decreased data  
1802 diversity shows the opposite trend (Figure 20).  
1803  
1804  
1805  
1806  
1807  
1808  
1809  
1810  
1811  
1812  
18131829  
1830  
1831  
1832  
1833 Figure 20: **Forgetting (left) and Backward Transfer (Right) of Reasoning Models Trained with  
1834 Narrow Data.** Backward transfer is generally low or moderate and forgetting is larger than training  
1835 on mixed data.

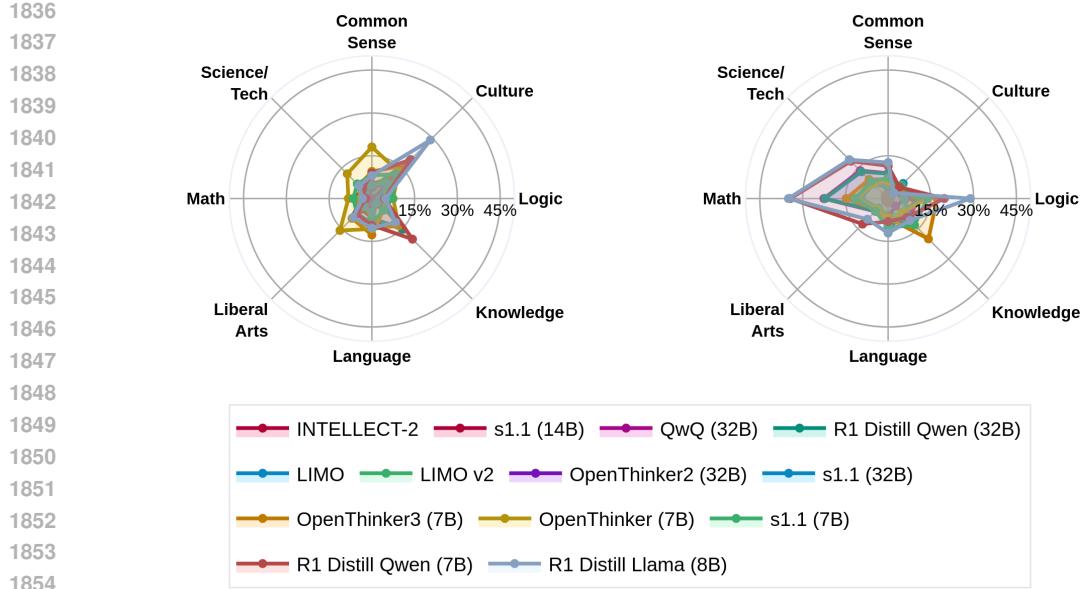


Figure 21: **Forgetting (left) and Backward Transfer (Right) of Reasoning Models Trained with Mixed Data.** Backward transfer is generally moderate to high.

## E.2 OBJECTIVE FUNCTION (SFT vs. RL)

Reasoning models tend to be trained using SFT or RL or both. Based on the tested models, we do not find evidence that these have differing behaviors. (Figures 22,23).

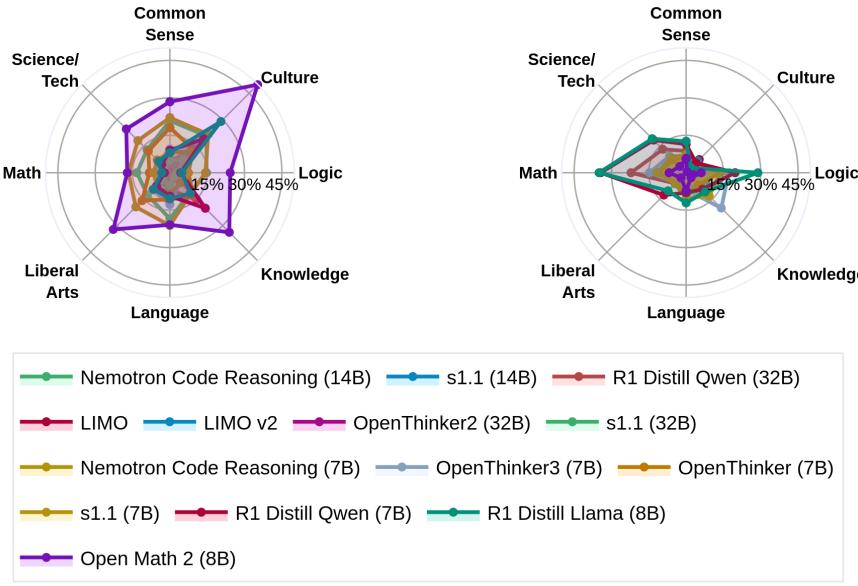


Figure 22: **Forgetting (left) and Backward Transfer (Right) of Reasoning Models Trained with SFT Data.**

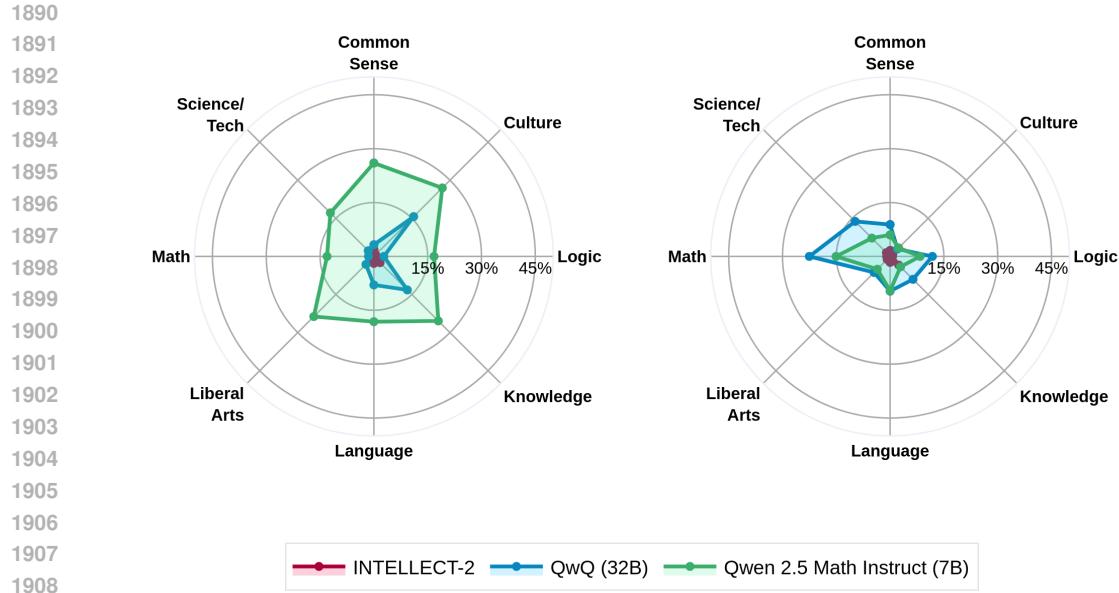


Figure 23: **Forgetting (left) and Backward Transfer (Right) of Reasoning Models Trained with RL Data.**

### E.3 DATA VOLUME

Data volume is another factor we consider in the forgetting and backward transfer dynamics. While low data volume is indicative of low forgetting and backward transfer (Figure 24), no trend is apparent in the case of high data volume (Figure 25).

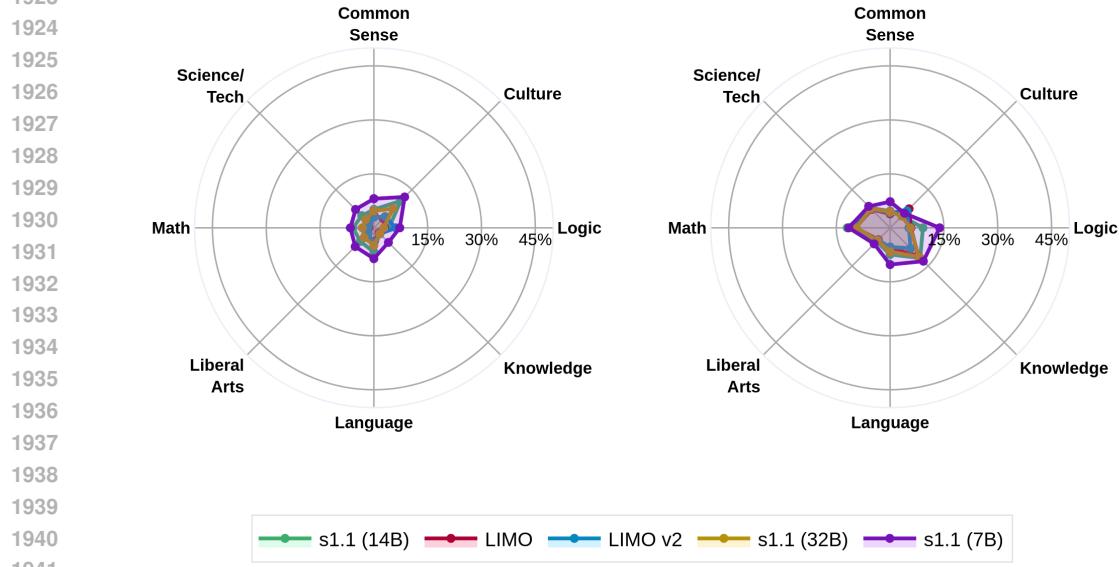


Figure 24: **Forgetting (left) and Backward Transfer (Right) of Reasoning Models Trained with Low Data Volume.** Both metrics are generally low across categories.

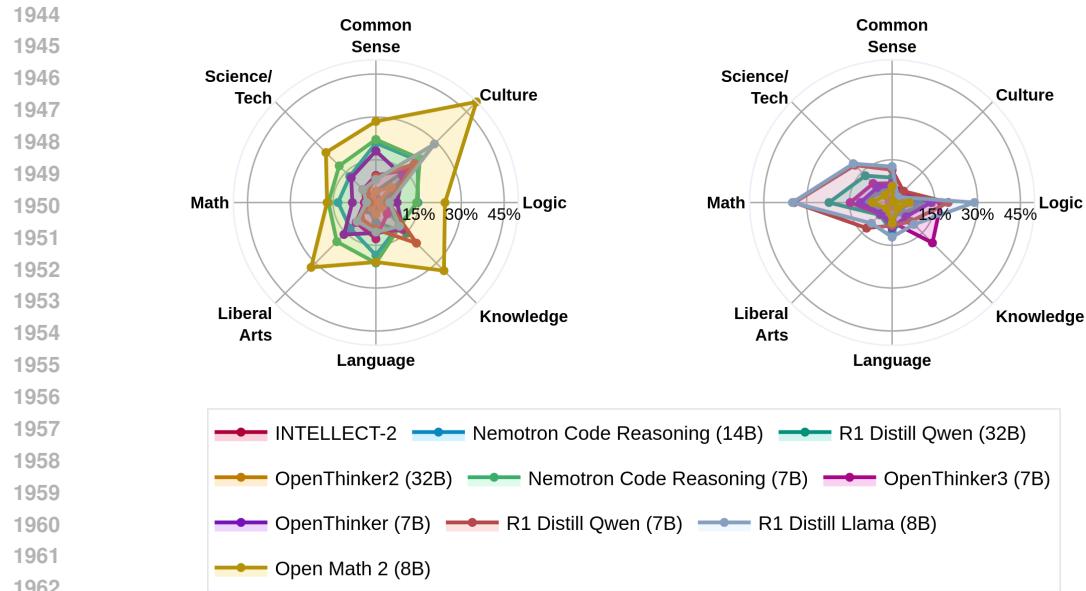


Figure 25: **Forgetting (left) and Backward Transfer (Right) of Reasoning Models Trained with High Data Volume.** Forgetting and backward transfer vary significantly between models.

## F ADDITIONAL PLOTS

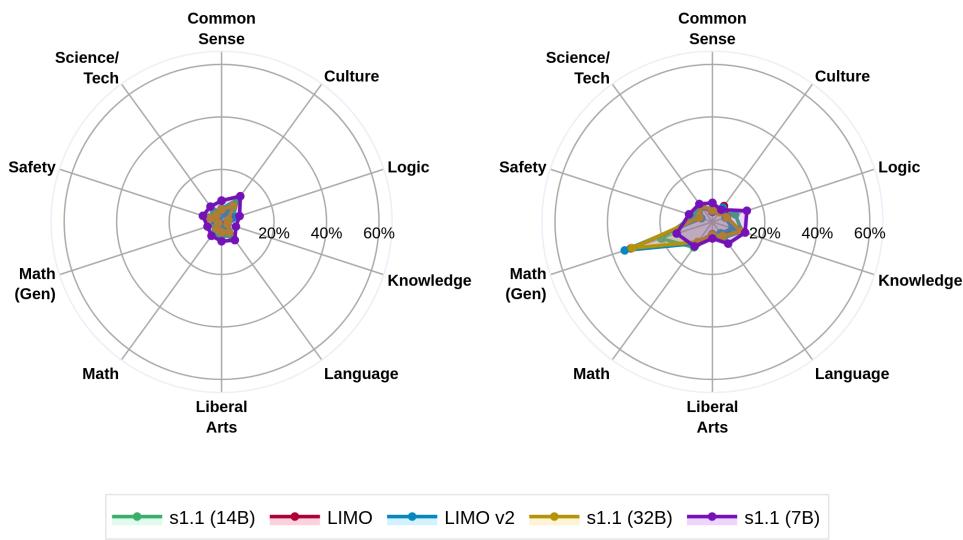


Figure 26: **Forgetting (left) and Backward Transfer (right) after reasoning training from instruct: low data scenario.** Yields little forgetting and backward transfer. Forgetting decreases with model scale.

1998  
1999  
2000  
2001  
2002  
2003  
2004  
2005  
2006  
2007  
2008  
2009  
2010  
2011  
2012  
2013  
2014  
2015  
2016  
2017  
2018  
2019  
2020  
2021  
2022  
2023  
2024  
2025  
2026  
2027  
2028  
2029  
2030  
2031  
2032  
2033  
2034  
2035  
2036  
2037  
2038  
2039  
2040  
2041  
2042  
2043  
2044  
2045  
2046  
2047  
2048  
2049  
2050  
2051

2052 **G QUANTIFYING FORGETTING ACCURATELY (TABLES FOR REFERENCING**  
 2053 **PLOTS)<sup>3</sup>**  
 2054

2055 Forgetting/backward transfer tables are listed with forgetting as the first number in each entry, standard  
 2056 deviation after the "±", and maximum possible forgetting/backward transfer respectively in brackets.  
 2057

2058 **G.1 INSTRUCTION TUNING**  
 2059

2060 **Table 2: Instruction Tuning: Forgetting (Part 1 of 3)**  
 2061

Category	Q2.5 Inst. (3B)	Q2.5 Inst. (7B)	Q2.5 Inst. (14B)	Q2.5 Inst. (32B)
Common Sense	7.0 ±0.3 (64.2)	5.4 ±0.1 (60.8)	3.9 ±0.6 (75.8)	3.7 ±0.5 (79.5)
Culture	11.7 ±0.9 (55.8)	16.5 ±2.8 (64.8)	14.0 ±0.1 (76.4)	15.2 ±0.6 (78.9)
Logic	10.9 ±0.5 (36.2)	8.4 ±0.2 (52.5)	5.5 ±0.4 (65.3)	4.6 ±0.6 (74.4)
Knowledge/QA	6.8 ±1.3 (47.3)	15.0 ±1.0 (58.4)	23.4 ±0.3 (76.9)	15.3 ±1.9 (69.4)
Language	8.7 ±0.8 (31.0)	9.2 ±0.6 (45.1)	9.5 ±0.8 (59.2)	8.6 ±1.0 (60.2)
Liberal Arts	8.7 ±0.7 (65.3)	6.6 ±0.7 (74.2)	5.3 ±0.4 (78.6)	5.3 ±0.3 (81.9)
Math	7.7 ±0.4 (35.4)	4.2 ±0.4 (47.3)	6.2 ±0.7 (57.9)	4.5 ±1.5 (64.4)
Science/Tech	6.5 ±0.3 (45.6)	5.4 ±0.4 (56.5)	4.5 ±0.5 (65.2)	4.4 ±0.3 (69.7)
<b>Total</b>	<b>8.5 ±0.1 (50.2)</b>	<b>8.2 ±0.3 (58.6)</b>	<b>8.0 ±0.3 (69.7)</b>	<b>6.7 ±0.5 (72.3)</b>

2072 **Table 3: Instruction Tuning: Forgetting (Part 2 of 3)**  
 2073

Category	Q2.5 Coder Inst. (3B)	Q2.5 Coder Inst. (7B)	Q2.5 Coder Inst. (14B)	Q2.5 Coder Inst. (32B)
Common Sense	11.7 ±1.0 (59.5)	8.1 ±0.3 (67.2)	6.1 ±0.2 (70.7)	4.6 ±0.4 (77.9)
Culture	15.0 ±2.6 (45.7)	19.0 ±1.9 (60.8)	16.6 ±1.6 (66.9)	19.1 ±1.1 (73.8)
Logic	17.2 ±0.6 (41.2)	14.1 ±0.2 (51.9)	6.9 ±0.2 (64.0)	5.8 ±0.3 (69.8)
Knowledge/QA	12.6 ±0.1 (48.0)	14.4 ±0.3 (56.4)	14.3 ±0.3 (64.3)	17.6 ±1.0 (77.7)
Language	15.1 ±0.7 (36.4)	13.7 ±1.0 (43.2)	10.3 ±0.8 (52.7)	8.6 ±0.5 (59.2)
Liberal Arts	13.9 ±0.6 (59.2)	9.6 ±0.0 (67.6)	7.8 ±0.4 (74.6)	6.7 ±0.2 (77.7)
Math	8.9 ±0.9 (32.8)	6.8 ±0.2 (40.8)	6.2 ±0.4 (54.4)	4.9 ±0.6 (58.6)
Science/Tech	9.4 ±0.7 (40.5)	8.4 ±0.5 (52.0)	7.0 ±0.2 (59.9)	5.6 ±0.5 (65.1)
<b>Total</b>	<b>13.0 ±0.2 (48.2)</b>	<b>10.9 ±0.3 (57.6)</b>	<b>8.9 ±0.3 (66.2)</b>	<b>8.4 ±0.1 (72.2)</b>

2085 **Table 4: Instruction Tuning: Forgetting (Part 3 of 3)**  
 2086

Category	Llama 3.1 Inst. (8B)
Common Sense	6.9 ±0.4 (64.5)
Culture	25.3 ±2.2 (79.1)
Logic	10.9 ±0.5 (42.5)
Knowledge/QA	20.6 ±0.8 (60.8)
Language	10.4 ±0.7 (39.9)
Liberal Arts	7.6 ±0.6 (64.6)
Math	7.3 ±0.9 (30.5)
Science/Tech	5.9 ±0.9 (45.5)
<b>Total</b>	<b>10.8 ±0.3 (54.5)</b>

2099 <sup>3</sup>For brevity we shorten Qwen 2.5 to Q2.5 as well as the associated models (e.g. Qwen 2.5 Instruct to Q2.5 Inst.)

Table 5: Instruction Tuning: Backward Transfer (Part 1 of 3)

Category	Q2.5 Inst. (3B)	Q2.5 Inst. (7B)	Q2.5 Inst. (14B)	Q2.5 Inst. (32B)
Common Sense	10.8 $\pm$ 0.4 (69.2)	17.5 $\pm$ 0.3 (76.9)	8.7 $\pm$ 0.1 (82.1)	7.5 $\pm$ 0.3 (84.6)
Culture	7.9 $\pm$ 2.4 (49.2)	5.2 $\pm$ 1.9 (45.3)	5.3 $\pm$ 0.3 (63.7)	3.0 $\pm$ 1.1 (62.1)
Logic	10.2 $\pm$ 0.7 (33.6)	14.6 $\pm$ 0.5 (60.4)	13.3 $\pm$ 0.5 (75.6)	9.0 $\pm$ 0.3 (80.4)
Knowledge/QA	13.3 $\pm$ 2.8 (55.8)	7.5 $\pm$ 2.0 (52.5)	5.1 $\pm$ 1.0 (57.9)	7.7 $\pm$ 1.7 (61.3)
Language	7.7 $\pm$ 0.5 (29.8)	8.3 $\pm$ 0.4 (41.4)	7.2 $\pm$ 0.0 (54.0)	8.3 $\pm$ 2.4 (57.0)
Liberal Arts	7.2 $\pm$ 1.4 (63.3)	5.7 $\pm$ 0.6 (73.0)	5.6 $\pm$ 0.9 (78.9)	4.6 $\pm$ 1.1 (80.9)
Math	18.9 $\pm$ 0.9 (51.0)	19.0 $\pm$ 0.8 (67.3)	17.8 $\pm$ 1.4 (73.8)	15.9 $\pm$ 1.9 (80.0)
Science/Tech	11.4 $\pm$ 0.8 (52.1)	11.9 $\pm$ 0.9 (65.2)	10.4 $\pm$ 1.0 (73.0)	9.5 $\pm$ 0.8 (76.5)
<b>Total</b>	10.5 $\pm$ 0.5 (52.7)	11.0 $\pm$ 0.5 (62.1)	8.9 $\pm$ 0.4 (71.3)	8.2 $\pm$ 0.7 (74.3)

Table 6: Instruction Tuning: Backward Transfer (Part 2 of 3)

Category	Q2.5 Coder Inst. (3B)	Q2.5 Coder Inst. (7B)	Q2.5 Coder Inst. (14B)	Q2.5 Coder Inst. (32B)
Common Sense	10.9 $\pm$ 0.1 (58.3)	10.6 $\pm$ 1.0 (70.5)	10.2 $\pm$ 0.8 (76.2)	6.9 $\pm$ 0.1 (80.9)
Culture	4.8 $\pm$ 0.9 (31.0)	4.2 $\pm$ 0.7 (37.8)	2.8 $\pm$ 0.5 (47.0)	2.7 $\pm$ 0.7 (50.0)
Logic	6.6 $\pm$ 0.3 (25.3)	11.7 $\pm$ 0.7 (47.2)	11.5 $\pm$ 0.3 (70.6)	10.8 $\pm$ 0.1 (76.4)
Knowledge/QA	8.3 $\pm$ 1.2 (43.4)	9.4 $\pm$ 1.9 (52.5)	10.2 $\pm$ 1.3 (60.6)	4.8 $\pm$ 1.5 (64.2)
Language	4.7 $\pm$ 0.4 (19.2)	5.6 $\pm$ 1.2 (29.4)	7.9 $\pm$ 0.2 (48.0)	6.5 $\pm$ 0.7 (55.2)
Liberal Arts	5.8 $\pm$ 1.1 (48.4)	6.4 $\pm$ 0.7 (63.3)	5.5 $\pm$ 1.1 (71.6)	5.1 $\pm$ 0.8 (75.6)
Math	14.4 $\pm$ 1.8 (41.0)	21.7 $\pm$ 1.4 (61.2)	17.5 $\pm$ 1.9 (69.6)	19.1 $\pm$ 1.5 (77.8)
Science/Tech	8.7 $\pm$ 1.6 (39.5)	10.2 $\pm$ 0.5 (54.3)	10.4 $\pm$ 1.0 (64.4)	9.8 $\pm$ 0.4 (70.6)
<b>Total</b>	7.4 $\pm$ 0.2 (40.4)	9.3 $\pm$ 0.6 (54.9)	8.6 $\pm$ 0.3 (65.8)	7.6 $\pm$ 0.4 (71.2)

Table 7: Instruction Tuning: Backward Transfer (Part 3 of 3)

Category	Llama 3.1 Inst. (8B)
Common Sense	11.2 $\pm$ 0.3 (70.3)
Culture	3.9 $\pm$ 0.2 (46.2)
Logic	17.0 $\pm$ 0.9 (49.6)
Knowledge/QA	10.8 $\pm$ 2.3 (52.9)
Language	8.9 $\pm$ 1.4 (36.3)
Liberal Arts	9.8 $\pm$ 1.7 (67.5)
Math	19.3 $\pm$ 0.7 (46.9)
Science/Tech	15.1 $\pm$ 1.3 (57.8)
<b>Total</b>	11.4 $\pm$ 1.0 (55.1)

2160 G.2 DOMAIN-CONTINUAL PRETRAINING  
21612162 Table 8: Domain-Continual Pretraining: Forgetting (Part 1 of 2)  
2163

Category	Q2.5 Coder (3B)	Q2.5 Coder (7B)	Q2.5 Coder (14B)	Q2.5 Coder (32B)
Common Sense	11.9 $\pm$ 0.5 (64.2)	9.0 $\pm$ 0.4 (60.8)	10.4 $\pm$ 0.6 (75.8)	7.5 $\pm$ 0.6 (79.5)
Culture	11.7 $\pm$ 0.7 (55.8)	10.9 $\pm$ 0.4 (64.8)	10.8 $\pm$ 0.7 (76.4)	8.8 $\pm$ 0.8 (78.9)
Logic	5.7 $\pm$ 0.2 (36.2)	9.5 $\pm$ 0.2 (52.5)	7.6 $\pm$ 0.3 (65.3)	8.4 $\pm$ 0.1 (74.4)
Knowledge/QA	5.6 $\pm$ 0.5 (47.3)	6.0 $\pm$ 0.8 (58.4)	13.7 $\pm$ 0.2 (76.9)	3.8 $\pm$ 0.5 (69.4)
Language	5.9 $\pm$ 0.6 (31.0)	8.4 $\pm$ 1.5 (45.1)	8.6 $\pm$ 0.9 (59.2)	7.4 $\pm$ 1.3 (60.2)
Liberal Arts	6.4 $\pm$ 0.5 (65.3)	7.0 $\pm$ 0.3 (74.2)	5.1 $\pm$ 0.4 (78.6)	5.2 $\pm$ 0.3 (81.9)
Math	3.8 $\pm$ 0.9 (35.4)	7.4 $\pm$ 1.0 (47.3)	6.2 $\pm$ 0.2 (57.9)	7.8 $\pm$ 0.5 (64.4)
Science/Tech	4.0 $\pm$ 0.5 (45.6)	5.9 $\pm$ 0.3 (56.5)	6.4 $\pm$ 0.5 (65.2)	5.7 $\pm$ 0.5 (69.7)
<b>Total</b>	<b>6.8 <math>\pm</math> 0.0 (50.9)</b>	<b>7.6 <math>\pm</math> 0.2 (60.4)</b>	<b>8.0 <math>\pm</math> 0.1 (71.8)</b>	<b>6.4 <math>\pm</math> 0.2 (74.6)</b>

2174 Table 9: Domain-Continual Pretraining: Forgetting (Part 2 of 2)  
2175

Category	Q2.5 Math (7B)
Common Sense	13.6 $\pm$ 0.8 (60.8)
Culture	17.8 $\pm$ 0.3 (64.8)
Logic	9.8 $\pm$ 0.4 (52.5)
Knowledge/QA	9.8 $\pm$ 0.5 (58.4)
Language	11.3 $\pm$ 0.9 (45.1)
Liberal Arts	20.0 $\pm$ 1.3 (74.2)
Math	7.5 $\pm$ 1.2 (47.3)
Science/Tech	14.4 $\pm$ 0.5 (56.5)
<b>Total</b>	<b>12.9 <math>\pm</math> 0.4 (60.4)</b>

2188 Table 10: Domain-Continual Pretraining: Backward Transfer (Part 1 of 2)  
2189

Category	Q2.5 Coder (3B)	Q2.5 Coder (7B)	Q2.5 Coder (14B)	Q2.5 Coder (32B)
Common Sense	8.3 $\pm$ 0.4 (59.5)	13.8 $\pm$ 0.3 (67.2)	6.6 $\pm$ 0.3 (70.7)	6.3 $\pm$ 0.5 (77.9)
Culture	6.3 $\pm$ 2.1 (45.7)	10.0 $\pm$ 1.3 (60.8)	5.8 $\pm$ 0.4 (66.9)	5.6 $\pm$ 0.2 (73.8)
Logic	9.4 $\pm$ 0.5 (41.2)	9.7 $\pm$ 0.2 (51.9)	7.7 $\pm$ 0.8 (64.0)	5.3 $\pm$ 0.4 (69.8)
Knowledge/QA	6.6 $\pm$ 1.1 (48.0)	5.4 $\pm$ 0.5 (56.4)	2.7 $\pm$ 0.3 (64.3)	12.4 $\pm$ 0.7 (77.7)
Language	8.9 $\pm$ 0.9 (36.4)	7.6 $\pm$ 0.8 (43.2)	5.3 $\pm$ 0.5 (52.7)	7.8 $\pm$ 2.2 (59.2)
Liberal Arts	2.0 $\pm$ 0.2 (59.2)	2.2 $\pm$ 0.3 (67.6)	2.3 $\pm$ 0.3 (74.6)	2.1 $\pm$ 0.1 (77.7)
Math	2.4 $\pm$ 0.2 (32.8)	3.2 $\pm$ 0.3 (40.8)	3.4 $\pm$ 1.2 (54.4)	3.4 $\pm$ 1.0 (58.6)
Science/Tech	1.0 $\pm$ 0.3 (40.5)	2.8 $\pm$ 0.0 (52.0)	2.6 $\pm$ 0.1 (59.9)	2.4 $\pm$ 0.2 (65.1)
<b>Total</b>	<b>5.1 <math>\pm</math> 0.2 (48.2)</b>	<b>6.2 <math>\pm</math> 0.1 (57.6)</b>	<b>4.2 <math>\pm</math> 0.2 (66.2)</b>	<b>5.1 <math>\pm</math> 0.2 (72.2)</b>

2200  
2201  
2202  
2203  
2204  
2205  
2206  
2207  
2208  
2209  
2210  
2211  
2212  
2213

2214  
 2215  
 2216  
 2217  
 2218  
 2219  
 2220  
 2221  
 2222  
 2223  
 2224  
 2225  
 2226  
 2227  
 2228  
 2229  
 2230  
 2231  
 2232  
 2233  
 2234  
 2235  
 2236

Table 11: Domain-Continual Pretraining: Backward Transfer (Part 2 of 2)

Category	Q2.5 Math (7B)
Common Sense	12.7 $\pm$ 0.5 (59.6)
Culture	6.9 $\pm$ 0.8 (46.3)
Logic	11.9 $\pm$ 0.2 (53.9)
Knowledge/QA	9.9 $\pm$ 1.8 (55.3)
Language	6.1 $\pm$ 0.2 (36.3)
Liberal Arts	1.6 $\pm$ 0.3 (49.5)
Math	4.3 $\pm$ 0.5 (43.1)
Science/Tech	3.2 $\pm$ 0.7 (40.7)
<b>Total</b>	<b>6.3 <math>\pm</math>0.2 (50.1)</b>

2247  
 2248  
 2249  
 2250  
 2251  
 2252  
 2253  
 2254  
 2255  
 2256  
 2257  
 2258  
 2259  
 2260  
 2261  
 2262  
 2263  
 2264  
 2265  
 2266  
 2267

2268 G.3 TRAINED FROM BASE  
22692270 Table 12: Trained from Base: Forgetting (Part 1 of 2)  
2271

Category	Q2.5 Math Inst. (7B)	QwQ (32B)	R1 Distill Qwen (7B)	R1 Distill Llama (8B)
Common Sense	26.0 $\pm$ 0.3 (59.6)	3.2 $\pm$ 0.4 (79.5)	8.9 $\pm$ 0.3 (59.6)	8.1 $\pm$ 0.2 (64.5)
Culture	27.0 $\pm$ 2.4 (46.3)	15.7 $\pm$ 0.6 (78.9)	19.3 $\pm$ 2.1 (46.3)	29.0 $\pm$ 1.0 (79.1)
Logic	16.8 $\pm$ 0.4 (53.9)	2.8 $\pm$ 0.2 (74.4)	4.4 $\pm$ 0.4 (53.9)	4.8 $\pm$ 0.2 (42.5)
Knowledge/QA	25.4 $\pm$ 2.0 (55.3)	13.2 $\pm$ 2.1 (69.4)	20.1 $\pm$ 0.7 (55.3)	11.7 $\pm$ 0.7 (60.8)
Language	18.2 $\pm$ 0.9 (36.3)	7.9 $\pm$ 1.2 (60.2)	9.4 $\pm$ 0.8 (36.3)	10.4 $\pm$ 0.4 (39.9)
Liberal Arts	23.7 $\pm$ 0.7 (49.5)	3.1 $\pm$ 0.2 (81.9)	7.6 $\pm$ 0.4 (49.5)	9.4 $\pm$ 0.9 (64.6)
Math	13.1 $\pm$ 0.3 (43.1)	1.5 $\pm$ 0.3 (64.4)	2.2 $\pm$ 0.4 (43.1)	3.3 $\pm$ 0.3 (30.5)
Science/Tech	17.1 $\pm$ 0.9 (40.7)	2.2 $\pm$ 0.2 (69.7)	4.6 $\pm$ 0.2 (40.7)	6.4 $\pm$ 0.4 (45.5)
<b>Total</b>	21.4 $\pm$ 0.5 (50.1)	5.4 $\pm$ 0.3 (72.3)	8.7 $\pm$ 0.1 (50.1)	9.3 $\pm$ 0.3 (54.5)

2282 Table 13: Trained from Base: Forgetting (Part 2 of 2)  
2283

Category	R1 Distill Qwen (32B)
Common Sense	3.9 $\pm$ 0.3 (79.5)
Culture	18.8 $\pm$ 0.7 (78.9)
Logic	2.3 $\pm$ 0.3 (74.4)
Knowledge/QA	15.3 $\pm$ 1.3 (69.4)
Language	6.9 $\pm$ 1.4 (60.2)
Liberal Arts	3.6 $\pm$ 0.3 (81.9)
Math	1.8 $\pm$ 0.5 (64.4)
Science/Tech	2.4 $\pm$ 0.3 (69.7)
<b>Total</b>	6.0 $\pm$ 0.2 (72.3)

2295 Table 14: Trained from Base: Backward Transfer (Part 1 of 2)  
2296

Category	Q2.5 Math Inst. (7B)	QwQ (32B)	R1 Distill Qwen (7B)	R1 Distill Llama (8B)
Common Sense	6.0 $\pm$ 0.6 (32.9)	8.9 $\pm$ 0.5 (87.0)	11.4 $\pm$ 0.1 (63.0)	12.6 $\pm$ 0.8 (70.5)
Culture	3.4 $\pm$ 1.3 (9.0)	2.9 $\pm$ 1.0 (59.8)	5.6 $\pm$ 1.4 (23.1)	2.8 $\pm$ 0.4 (36.5)
Logic	8.4 $\pm$ 0.4 (42.8)	11.8 $\pm$ 0.5 (86.7)	19.7 $\pm$ 0.1 (74.9)	28.9 $\pm$ 0.1 (74.6)
Knowledge/QA	4.1 $\pm$ 1.4 (31.2)	9.0 $\pm$ 1.6 (65.2)	9.3 $\pm$ 2.4 (45.6)	10.8 $\pm$ 1.8 (61.8)
Language	9.6 $\pm$ 0.6 (20.9)	9.7 $\pm$ 2.8 (60.6)	8.1 $\pm$ 0.8 (33.7)	12.1 $\pm$ 1.0 (42.2)
Liberal Arts	5.0 $\pm$ 0.4 (24.5)	6.3 $\pm$ 1.0 (86.1)	12.7 $\pm$ 1.6 (56.2)	10.3 $\pm$ 1.5 (65.8)
Math	15.1 $\pm$ 1.6 (46.5)	22.5 $\pm$ 2.2 (92.2)	34.3 $\pm$ 2.4 (85.8)	35.0 $\pm$ 0.5 (72.8)
Science/Tech	7.2 $\pm$ 0.2 (27.4)	13.8 $\pm$ 1.0 (85.0)	18.5 $\pm$ 1.5 (59.3)	19.3 $\pm$ 1.2 (62.7)
<b>Total</b>	6.7 $\pm$ 0.2 (30.1)	10.3 $\pm$ 0.6 (78.6)	14.3 $\pm$ 1.0 (57.4)	15.4 $\pm$ 0.6 (62.1)

2322  
 2323  
 2324  
 2325  
 2326  
 2327  
 2328  
 2329  
 2330  
 2331  
 2332  
 2333  
 2334  
 2335  
 2336  
 2337  
 2338  
 2339  
 2340  
 2341  
 2342  
 2343  
 2344

Table 15: Trained from Base: Backward Transfer (Part 2 of 2)

Category	R1 Distill Qwen (32B)
Common Sense	8.7 $\pm$ 0.3 (85.9)
Culture	2.5 $\pm$ 0.6 (55.4)
Logic	11.7 $\pm$ 0.4 (87.2)
Knowledge/QA	6.8 $\pm$ 1.3 (60.6)
Language	10.2 $\pm$ 2.7 (62.9)
Liberal Arts	6.1 $\pm$ 1.0 (85.2)
Math	22.1 $\pm$ 2.5 (91.4)
Science/Tech	13.3 $\pm$ 1.0 (84.2)
<b>Total</b>	10.0 $\pm$ 0.6 (77.5)

2355  
 2356  
 2357  
 2358  
 2359  
 2360  
 2361  
 2362  
 2363  
 2364  
 2365  
 2366  
 2367  
 2368  
 2369  
 2370  
 2371  
 2372  
 2373  
 2374  
 2375

2376 G.4 TRAINED FROM INSTRUCT - HIGH DATA SCENARIO  
23772378 Table 16: Trained from Instruct - High Data Scenario: Forgetting (Part 1 of 2)  
2379

Category	INTELLECT-2 (32B)	Open Math 2 (8B)	OpenThinker (7B)	OpenThinker2 (32B)
Common Sense	1.9 $\pm$ 0.3 (87.0)	28.5 $\pm$ 0.9 (64.5)	18.1 $\pm$ 1.3 (76.9)	3.9 $\pm$ 0.3 (84.6)
Culture	0.7 $\pm$ 0.6 (60.2)	49.9 $\pm$ 1.9 (79.1)	13.3 $\pm$ 4.6 (45.3)	7.6 $\pm$ 2.1 (62.1)
Logic	0.6 $\pm$ 0.1 (87.4)	24.2 $\pm$ 0.5 (42.5)	7.4 $\pm$ 0.2 (60.4)	1.0 $\pm$ 0.2 (80.4)
Knowledge/QA	2.5 $\pm$ 0.6 (65.9)	33.8 $\pm$ 1.4 (60.8)	13.4 $\pm$ 1.1 (52.5)	2.8 $\pm$ 0.5 (61.3)
Language	1.9 $\pm$ 0.3 (62.7)	20.9 $\pm$ 1.2 (39.9)	10.7 $\pm$ 1.2 (41.4)	4.2 $\pm$ 0.4 (57.0)
Liberal Arts	1.2 $\pm$ 0.2 (86.1)	32.1 $\pm$ 2.3 (64.6)	15.8 $\pm$ 1.0 (73.0)	2.3 $\pm$ 0.1 (80.9)
Math	0.9 $\pm$ 0.1 (91.9)	17.1 $\pm$ 1.4 (30.5)	8.3 $\pm$ 0.8 (67.3)	1.0 $\pm$ 0.3 (80.0)
Safety/Truth	0.9 $\pm$ 0.2 (66.8)	19.5 $\pm$ 0.7 (36.3)	8.6 $\pm$ 0.5 (50.0)	3.3 $\pm$ 0.2 (64.5)
Science/Tech	1.6 $\pm$ 0.1 (85.0)	24.8 $\pm$ 1.8 (45.5)	12.2 $\pm$ 0.5 (65.2)	2.1 $\pm$ 0.2 (76.5)
<b>Total</b>	1.3 $\pm$ 0.1 (79.0)	28.8 $\pm$ 0.8 (54.5)	12.6 $\pm$ 1.5 (62.0)	2.9 $\pm$ 0.2 (74.3)

2391 Table 17: Trained from Instruct - High Data Scenario: Forgetting (Part 2 of 2)  
2392

Category	OpenThinker3 (7B)	Nemotron Code Reasoning (7B)	Nemotron Code Reasoning (14B)
Common Sense	9.5 $\pm$ 0.4 (76.9)	22.1 $\pm$ 0.1 (76.9)	20.9 $\pm$ 2.4 (79.0)
Culture	16.4 $\pm$ 4.5 (45.3)	21.8 $\pm$ 3.1 (45.3)	20.9 $\pm$ 0.7 (63.7)
Logic	5.2 $\pm$ 0.2 (60.4)	14.3 $\pm$ 0.4 (60.4)	14.7 $\pm$ 0.3 (75.6)
Knowledge/QA	5.6 $\pm$ 0.9 (52.5)	11.9 $\pm$ 1.6 (52.5)	11.5 $\pm$ 2.1 (57.9)
Language	12.7 $\pm$ 0.8 (41.4)	21.2 $\pm$ 1.4 (48.3)	18.4 $\pm$ 1.2 (62.4)
Liberal Arts	9.8 $\pm$ 0.6 (73.0)	19.4 $\pm$ 0.7 (73.0)	12.8 $\pm$ 1.0 (79.6)
Math	4.1 $\pm$ 0.1 (67.3)	16.6 $\pm$ 0.5 (67.3)	13.3 $\pm$ 0.4 (73.8)
Safety/Truth	8.7 $\pm$ 0.6 (50.0)	14.2 $\pm$ 0.1 (50.0)	12.3 $\pm$ 1.2 (63.0)
Science/Tech	6.4 $\pm$ 0.3 (65.2)	18.2 $\pm$ 0.4 (65.2)	13.0 $\pm$ 0.2 (73.0)
<b>Total</b>	8.4 $\pm$ 0.5 (62.1)	17.7 $\pm$ 0.3 (62.4)	14.9 $\pm$ 0.0 (72.1)

2405 Table 18: Trained from Instruct - High Data Scenario: Backward Transfer (Part 1 of 2)  
2406

Category	INTELLECT-2 (32B)	Open Math 2 (8B)	OpenThinker (7B)	OpenThinker2 (32B)
Common Sense	1.7 $\pm$ 0.1 (86.7)	5.7 $\pm$ 0.4 (34.1)	6.1 $\pm$ 0.8 (60.9)	4.2 $\pm$ 0.5 (85.0)
Culture	1.1 $\pm$ 0.3 (60.8)	0.5 $\pm$ 0.4 (6.5)	5.7 $\pm$ 1.6 (33.6)	2.9 $\pm$ 0.3 (54.3)
Logic	0.7 $\pm$ 0.1 (87.4)	6.0 $\pm$ 0.4 (18.3)	13.7 $\pm$ 0.6 (67.4)	6.2 $\pm$ 0.4 (87.5)
Knowledge/QA	3.4 $\pm$ 0.9 (67.5)	2.8 $\pm$ 1.2 (25.3)	7.0 $\pm$ 0.7 (43.0)	3.8 $\pm$ 0.1 (62.3)
Language	1.7 $\pm$ 0.2 (62.7)	7.2 $\pm$ 0.8 (18.1)	8.8 $\pm$ 1.6 (39.1)	6.4 $\pm$ 0.7 (61.0)
Liberal Arts	1.2 $\pm$ 0.1 (86.0)	2.9 $\pm$ 0.2 (25.5)	5.5 $\pm$ 0.5 (59.2)	5.6 $\pm$ 0.3 (85.4)
Math	0.9 $\pm$ 0.2 (92.1)	6.9 $\pm$ 1.1 (15.0)	11.1 $\pm$ 1.0 (71.3)	10.1 $\pm$ 1.1 (91.7)
Safety/Truth	1.3 $\pm$ 0.5 (67.2)	4.0 $\pm$ 0.6 (15.9)	7.2 $\pm$ 1.2 (48.1)	4.1 $\pm$ 0.8 (65.5)
Science/Tech	1.5 $\pm$ 0.1 (85.0)	3.6 $\pm$ 0.9 (17.0)	7.4 $\pm$ 0.4 (58.7)	7.6 $\pm$ 0.3 (83.8)
<b>Total</b>	1.4 $\pm$ 0.1 (79.2)	4.2 $\pm$ 0.2 (21.1)	7.7 $\pm$ 0.5 (55.1)	5.3 $\pm$ 0.2 (77.3)

2430  
 2431  
 2432  
 2433  
 2434  
 2435  
 2436  
 2437  
 2438  
 2439  
 2440  
 2441  
 2442  
 2443  
 2444  
 2445  
 2446  
 2447  
 2448  
 2449  
 2450

2451 Table 19: Trained from Instruct - High Data Scenario: Backward Transfer (Part 2 of 2)

2452  
 2453  
 2454  
 2455  
 2456  
 2457  
 2458  
 2459  
 2460  
 2461  
 2462  
 2463  
 2464  
 2465  
 2466  
 2467  
 2468  
 2469  
 2470  
 2471  
 2472  
 2473  
 2474  
 2475  
 2476  
 2477  
 2478  
 2479  
 2480  
 2481  
 2482  
 2483

Category	OpenThinker3 (7B)	Nemotron Code Reasoning (7B)	Nemotron Code Reasoning (14B)
Common Sense	6.7 $\pm$ 0.2 (73.2)	5.2 $\pm$ 0.4 (54.3)	3.6 $\pm$ 1.2 (56.0)
Culture	4.0 $\pm$ 1.0 (23.3)	5.1 $\pm$ 2.3 (19.2)	5.2 $\pm$ 0.5 (37.8)
Logic	17.2 $\pm$ 0.9 (76.3)	13.7 $\pm$ 0.7 (57.4)	8.1 $\pm$ 0.3 (65.4)
Knowledge/QA	20.0 $\pm$ 0.3 (66.6)	5.0 $\pm$ 0.5 (42.6)	3.2 $\pm$ 0.2 (46.6)
Language	6.8 $\pm$ 0.7 (33.6)	4.9 $\pm$ 0.8 (23.3)	7.6 $\pm$ 0.8 (46.3)
Liberal Arts	5.7 $\pm$ 0.2 (67.5)	4.3 $\pm$ 0.1 (52.8)	4.2 $\pm$ 0.5 (68.3)
Math	14.7 $\pm$ 0.3 (81.2)	10.4 $\pm$ 0.0 (56.4)	10.7 $\pm$ 0.8 (69.4)
Safety/Truth	7.5 $\pm$ 0.9 (48.5)	8.6 $\pm$ 1.2 (42.5)	6.8 $\pm$ 0.6 (55.6)
Science/Tech	9.4 $\pm$ 0.5 (69.2)	5.2 $\pm$ 0.0 (47.7)	6.4 $\pm$ 0.1 (64.3)
<b>Total</b>	9.7 $\pm$ 0.2 (62.8)	6.7 $\pm$ 0.4 (46.5)	5.8 $\pm$ 0.1 (59.0)

2484 G.5 TRAINED FROM INSTRUCT - LOW DATA SCENARIO  
24852486 Table 20: Trained from Instruct - Low Data Scenario: Forgetting (Part 1 of 2)  
2487

Category	s1.1 (7B)	s1.1 (14B)	s1.1 (32B)	LIMO (32B)
Common Sense	8.1 $\pm$ 0.4 (76.9)	5.1 $\pm$ 1.0 (82.1)	4.7 $\pm$ 0.5 (84.6)	3.5 $\pm$ 0.3 (84.6)
Culture	12.1 $\pm$ 1.2 (45.3)	10.6 $\pm$ 0.4 (63.7)	7.5 $\pm$ 0.9 (62.1)	3.7 $\pm$ 1.9 (62.1)
Logic	7.2 $\pm$ 0.1 (60.4)	4.2 $\pm$ 0.4 (75.6)	2.8 $\pm$ 0.2 (80.4)	3.0 $\pm$ 0.3 (80.4)
Knowledge/QA	5.7 $\pm$ 1.1 (52.5)	2.3 $\pm$ 0.1 (57.9)	2.3 $\pm$ 1.2 (61.3)	2.0 $\pm$ 0.4 (61.3)
Language	8.6 $\pm$ 0.8 (41.4)	6.7 $\pm$ 0.4 (54.0)	4.9 $\pm$ 0.1 (57.0)	3.8 $\pm$ 0.6 (57.0)
Liberal Arts	7.4 $\pm$ 0.5 (73.0)	5.4 $\pm$ 0.7 (78.9)	3.9 $\pm$ 0.1 (80.9)	2.6 $\pm$ 0.0 (80.9)
Math	6.6 $\pm$ 0.8 (67.3)	5.8 $\pm$ 0.9 (73.8)	3.2 $\pm$ 0.4 (80.0)	1.3 $\pm$ 0.1 (80.0)
Safety/Truth	7.5 $\pm$ 0.9 (50.0)	5.3 $\pm$ 0.3 (59.4)	4.5 $\pm$ 0.4 (64.5)	2.9 $\pm$ 0.3 (64.5)
Science/Tech	7.2 $\pm$ 0.4 (65.2)	4.8 $\pm$ 0.6 (73.0)	3.3 $\pm$ 0.4 (76.5)	2.3 $\pm$ 0.1 (76.5)
<b>Total</b>	7.7 $\pm$ 0.2 (62.1)	5.3 $\pm$ 0.2 (71.3)	3.9 $\pm$ 0.1 (74.3)	2.6 $\pm$ 0.1 (74.3)

2500 Table 21: Trained from Instruct - Low Data Scenario: Forgetting (Part 2 of 2)  
2501

Category	LIMO v2 (32B)
Common Sense	3.0 $\pm$ 0.3 (84.6)
Culture	4.4 $\pm$ 0.9 (62.1)
Logic	6.1 $\pm$ 0.4 (80.4)
Knowledge/QA	2.4 $\pm$ 0.3 (61.3)
Language	4.2 $\pm$ 0.6 (57.0)
Liberal Arts	2.3 $\pm$ 0.3 (80.9)
Math	1.7 $\pm$ 0.1 (80.0)
Safety/Truth	2.8 $\pm$ 0.2 (64.5)
Science/Tech	1.9 $\pm$ 0.2 (76.5)
<b>Total</b>	3.0 $\pm$ 0.2 (74.3)

2513 Table 22: Trained from Instruct - Low Data Scenario: Backward Transfer (Part 1 of 2)  
2514

Category	s1.1 (7B)	s1.1 (14B)	s1.1 (32B)	LIMO (32B)
Common Sense	7.3 $\pm$ 0.5 (75.8)	4.6 $\pm$ 0.3 (81.5)	4.5 $\pm$ 0.2 (84.3)	3.9 $\pm$ 0.4 (85.1)
Culture	5.7 $\pm$ 0.5 (35.0)	4.6 $\pm$ 0.4 (54.2)	4.6 $\pm$ 0.4 (56.7)	7.5 $\pm$ 0.8 (66.1)
Logic	13.8 $\pm$ 0.6 (68.8)	9.2 $\pm$ 0.5 (81.1)	5.7 $\pm$ 0.4 (83.9)	5.8 $\pm$ 0.3 (84.6)
Knowledge/QA	13.1 $\pm$ 0.3 (59.3)	11.9 $\pm$ 1.3 (69.0)	11.3 $\pm$ 0.5 (71.2)	11.1 $\pm$ 1.3 (71.2)
Language	10.2 $\pm$ 0.7 (44.7)	7.4 $\pm$ 1.2 (55.7)	6.6 $\pm$ 0.5 (60.4)	5.6 $\pm$ 0.7 (60.3)
Liberal Arts	6.3 $\pm$ 0.7 (71.5)	5.3 $\pm$ 0.6 (78.9)	5.0 $\pm$ 0.1 (82.4)	4.7 $\pm$ 0.2 (83.7)
Math	11.5 $\pm$ 0.3 (74.2)	12.0 $\pm$ 0.7 (81.9)	9.4 $\pm$ 0.9 (87.8)	9.6 $\pm$ 0.7 (90.7)
Safety/Truth	9.4 $\pm$ 0.1 (52.5)	7.4 $\pm$ 0.9 (62.2)	5.2 $\pm$ 0.4 (65.4)	5.0 $\pm$ 0.4 (67.1)
Science/Tech	8.5 $\pm$ 0.1 (66.8)	7.9 $\pm$ 0.3 (77.3)	7.6 $\pm$ 0.5 (82.2)	7.2 $\pm$ 0.3 (83.2)
<b>Total</b>	9.1 $\pm$ 0.2 (63.5)	7.2 $\pm$ 0.3 (73.5)	6.2 $\pm$ 0.2 (76.9)	6.2 $\pm$ 0.2 (78.8)

2538  
 2539  
 2540  
 2541  
 2542  
 2543  
 2544  
 2545  
 2546  
 2547  
 2548  
 2549  
 2550  
 2551  
 2552  
 2553  
 2554  
 2555  
 2556  
 2557  
 2558

2559 Table 23: Trained from Instruct - Low Data Scenario: Backward Transfer (Part 2 of 2)  
 2560

Category	LIMO v2 (32B)
Common Sense	4.2 $\pm$ 0.2 (86.3)
Culture	6.7 $\pm$ 0.8 (64.6)
Logic	5.3 $\pm$ 0.6 (79.6)
Knowledge/QA	8.0 $\pm$ 0.7 (67.7)
Language	5.3 $\pm$ 0.3 (59.3)
Liberal Arts	4.9 $\pm$ 0.2 (84.4)
Math	10.1 $\pm$ 0.8 (90.8)
Safety/Truth	4.5 $\pm$ 0.7 (66.7)
Science/Tech	7.5 $\pm$ 0.2 (84.0)
<b>Total</b>	<b>5.8 <math>\pm</math>0.3 (77.9)</b>

2571  
 2572  
 2573  
 2574  
 2575  
 2576  
 2577  
 2578  
 2579  
 2580  
 2581  
 2582  
 2583  
 2584  
 2585  
 2586  
 2587  
 2588  
 2589  
 2590  
 2591

2592 G.6 QWEN2.5 BASE AND CODER MERGE (RELATIVE TO QWEN2.5 BASE)  
25932594 Table 24: Qwen2.5 Base and Coder Merge: Forgetting  
2595

Category	Linear (0.2)	Linear (0.8)	Slerp (0.2)	Slerp (0.8)
Common Sense	8.2 $\pm$ 0.6 (67.2)	11.3 $\pm$ 0.7 (67.2)	13.9 $\pm$ 2.2 (67.2)	16.2 $\pm$ 6.4 (67.2)
Culture	12.0 $\pm$ 2.0 (60.8)	15.8 $\pm$ 1.8 (60.8)	21.1 $\pm$ 3.4 (60.8)	41.3 $\pm$ 3.3 (60.8)
Logic	11.6 $\pm$ 0.4 (51.9)	19.4 $\pm$ 1.0 (51.9)	23.4 $\pm$ 1.1 (51.9)	17.7 $\pm$ 0.9 (51.9)
Knowledge/QA	4.6 $\pm$ 0.4 (56.4)	8.1 $\pm$ 0.3 (56.4)	10.9 $\pm$ 1.0 (56.4)	16.2 $\pm$ 1.0 (56.4)
Language	7.8 $\pm$ 0.4 (43.2)	11.2 $\pm$ 0.2 (43.2)	14.0 $\pm$ 0.5 (43.2)	12.7 $\pm$ 0.4 (43.2)
Liberal Arts	2.5 $\pm$ 0.5 (67.6)	4.1 $\pm$ 0.9 (67.6)	5.8 $\pm$ 1.6 (67.6)	8.9 $\pm$ 0.5 (67.6)
Math	4.5 $\pm$ 1.2 (44.9)	8.9 $\pm$ 2.1 (44.9)	12.4 $\pm$ 1.9 (44.9)	9.4 $\pm$ 0.7 (44.9)
Safety/Truth	1.9 $\pm$ 0.7 (49.9)	3.7 $\pm$ 0.6 (49.9)	6.0 $\pm$ 0.3 (49.9)	4.7 $\pm$ 1.0 (60.4)
Science/Tech	2.6 $\pm$ 0.6 (52.0)	5.2 $\pm$ 0.7 (52.0)	6.7 $\pm$ 1.3 (52.0)	4.4 $\pm$ 0.9 (52.0)
<b>Total</b>	6.1 $\pm$ 0.2 (58.0)	9.3 $\pm$ 0.4 (58.0)	12.3 $\pm$ 0.2 (58.0)	13.9 $\pm$ 0.4 (59.0)

2608 Table 25: Qwen2.5 Base and Coder Merge: Backward Transfer  
2609

Category	Linear (0.2)	Linear (0.8)	Slerp (0.2)	Slerp (0.8)
Common Sense	8.3 $\pm$ 0.4 (67.4)	9.8 $\pm$ 0.5 (65.2)	9.2 $\pm$ 0.2 (61.0)	7.2 $\pm$ 0.9 (55.3)
Culture	4.1 $\pm$ 0.8 (46.8)	6.2 $\pm$ 1.8 (46.1)	8.3 $\pm$ 2.0 (44.0)	2.6 $\pm$ 0.5 (15.1)
Logic	5.3 $\pm$ 0.6 (43.7)	4.2 $\pm$ 0.1 (31.0)	3.7 $\pm$ 0.1 (24.6)	4.4 $\pm$ 0.2 (33.3)
Knowledge/QA	8.3 $\pm$ 0.4 (61.6)	9.2 $\pm$ 0.4 (58.2)	8.2 $\pm$ 0.8 (53.4)	5.1 $\pm$ 0.6 (44.1)
Language	4.2 $\pm$ 0.8 (39.1)	7.2 $\pm$ 0.4 (37.4)	6.3 $\pm$ 0.2 (32.1)	4.6 $\pm$ 0.5 (30.7)
Liberal Arts	2.1 $\pm$ 0.4 (67.3)	6.1 $\pm$ 1.6 (70.2)	5.7 $\pm$ 2.0 (67.6)	2.1 $\pm$ 0.3 (58.7)
Math	5.5 $\pm$ 1.0 (45.0)	6.5 $\pm$ 1.6 (39.9)	5.8 $\pm$ 2.9 (34.0)	4.6 $\pm$ 1.4 (36.3)
Safety/Truth	2.4 $\pm$ 0.7 (51.8)	5.0 $\pm$ 1.9 (52.7)	6.8 $\pm$ 3.0 (51.5)	3.0 $\pm$ 1.0 (57.9)
Science/Tech	2.7 $\pm$ 0.3 (51.4)	5.6 $\pm$ 1.5 (52.3)	5.2 $\pm$ 1.6 (49.7)	2.3 $\pm$ 0.3 (48.1)
<b>Total</b>	4.8 $\pm$ 0.1 (56.0)	6.5 $\pm$ 0.7 (53.9)	6.4 $\pm$ 0.9 (49.9)	4.0 $\pm$ 0.1 (46.1)

2646 G.7 QWEN2.5 BASE AND CODER MERGE (RELATIVE TO QWEN2.5 CODER)  
26472648 Table 26: Qwen2.5 Base and Coder Merge: Forgetting  
2649

Category	Linear (0.2)	Linear (0.8)	Slerp (0.2)	Slerp (0.8)
Common Sense	10.2 $\pm$ 0.7 (70.5)	12.7 $\pm$ 0.8 (70.5)	15.2 $\pm$ 2.5 (70.5)	18.2 $\pm$ 7.0 (70.5)
Culture	6.8 $\pm$ 1.9 (37.8)	10.6 $\pm$ 1.4 (37.8)	15.8 $\pm$ 2.3 (37.8)	28.5 $\pm$ 3.2 (37.8)
Logic	15.1 $\pm$ 0.9 (47.2)	20.4 $\pm$ 0.7 (47.2)	23.2 $\pm$ 0.5 (47.2)	18.8 $\pm$ 0.3 (47.2)
Knowledge/QA	6.5 $\pm$ 0.8 (52.5)	8.0 $\pm$ 0.6 (52.5)	9.8 $\pm$ 0.3 (52.5)	14.0 $\pm$ 1.9 (52.5)
Language	6.6 $\pm$ 0.6 (29.4)	8.1 $\pm$ 0.5 (29.4)	9.5 $\pm$ 1.1 (29.4)	9.2 $\pm$ 0.9 (29.4)
Liberal Arts	6.6 $\pm$ 0.1 (63.3)	6.1 $\pm$ 1.0 (63.3)	7.6 $\pm$ 1.7 (63.3)	12.3 $\pm$ 0.3 (63.3)
Math	17.1 $\pm$ 0.5 (58.9)	19.6 $\pm$ 1.2 (58.9)	22.2 $\pm$ 2.1 (58.9)	20.8 $\pm$ 0.3 (58.9)
Safety/Truth	10.9 $\pm$ 1.7 (42.0)	10.0 $\pm$ 1.4 (42.0)	10.6 $\pm$ 1.3 (42.0)	11.2 $\pm$ 1.2 (43.4)
Science/Tech	10.3 $\pm$ 0.4 (54.3)	10.3 $\pm$ 0.9 (54.3)	11.6 $\pm$ 1.9 (54.3)	11.9 $\pm$ 0.3 (54.3)
<b>Total</b>	9.5 $\pm$ 0.3 (54.0)	11.1 $\pm$ 0.4 (54.0)	13.4 $\pm$ 0.3 (54.0)	15.3 $\pm$ 0.9 (54.1)

2661 Table 27: Qwen2.5 Base and Coder Merge: Backward Transfer  
2662

Category	Linear (0.2)	Linear (0.8)	Slerp (0.2)	Slerp (0.8)
Common Sense	7.9 $\pm$ 0.4 (67.4)	8.7 $\pm$ 0.5 (65.2)	8.0 $\pm$ 1.3 (61.0)	6.8 $\pm$ 1.6 (55.3)
Culture	13.4 $\pm$ 1.7 (46.8)	15.8 $\pm$ 3.2 (46.1)	17.8 $\pm$ 3.6 (44.0)	6.2 $\pm$ 1.9 (15.1)
Logic	11.4 $\pm$ 0.2 (43.7)	7.5 $\pm$ 0.3 (31.0)	6.0 $\pm$ 0.2 (24.6)	7.8 $\pm$ 0.0 (33.3)
Knowledge/QA	15.6 $\pm$ 1.0 (61.6)	14.3 $\pm$ 0.2 (58.2)	12.2 $\pm$ 1.2 (53.4)	8.7 $\pm$ 1.5 (44.1)
Language	11.0 $\pm$ 0.5 (39.1)	12.2 $\pm$ 0.3 (37.4)	9.9 $\pm$ 0.6 (32.1)	9.1 $\pm$ 0.2 (30.7)
Liberal Arts	9.6 $\pm$ 0.4 (67.3)	11.2 $\pm$ 0.9 (70.2)	10.8 $\pm$ 1.3 (67.6)	8.5 $\pm$ 0.2 (58.7)
Math	7.7 $\pm$ 1.0 (45.0)	7.0 $\pm$ 0.9 (39.9)	5.1 $\pm$ 1.5 (34.0)	5.2 $\pm$ 0.8 (36.3)
Safety/Truth	8.7 $\pm$ 1.7 (39.1)	10.3 $\pm$ 1.3 (42.4)	9.1 $\pm$ 1.7 (39.9)	9.2 $\pm$ 0.4 (41.0)
Science/Tech	8.1 $\pm$ 0.2 (51.4)	8.8 $\pm$ 0.6 (52.3)	8.1 $\pm$ 0.6 (49.7)	7.2 $\pm$ 0.2 (48.1)
<b>Total</b>	10.0 $\pm$ 0.3 (54.8)	10.2 $\pm$ 0.6 (52.9)	9.3 $\pm$ 0.7 (48.8)	7.5 $\pm$ 0.1 (44.4)

2675  
2676  
2677  
2678  
2679  
2680  
2681  
2682  
2683  
2684  
2685  
2686  
2687  
2688  
2689  
2690  
2691  
2692  
2693  
2694  
2695  
2696  
2697  
2698  
2699

2700 G.8 QWEN2.5 INSTRUCT AND OPENTHINKER 7B MERGE (RELATIVE TO QWEN2.5  
2701 INSTRUCT)  
27022703 Table 28: Qwen2.5 Instruct and OpenThinker 7B Merge: Forgetting  
2704

Category	Linear (0.2)	Linear (0.8)
Logic	8.7 $\pm$ 1.5 (75.2)	3.6 $\pm$ 0.4 (75.2)
Knowledge/QA	6.5 $\pm$ 0.4 (60.3)	4.2 $\pm$ 2.8 (60.3)
Liberal Arts	9.4 $\pm$ 0.8 (73.0)	3.7 $\pm$ 0.1 (73.0)
Math	10.5 $\pm$ 0.6 (65.3)	3.3 $\pm$ 0.5 (65.3)
Safety/Truth	7.7 $\pm$ 1.3 (55.1)	2.8 $\pm$ 0.4 (55.1)
Science/Tech	9.6 $\pm$ 0.3 (67.4)	4.2 $\pm$ 0.2 (67.4)
<b>Total</b>	<b>8.7 <math>\pm</math>0.2 (66.1)</b>	<b>3.6 <math>\pm</math>0.5 (66.1)</b>

2714 Table 29: Qwen2.5 Instruct and OpenThinker 7B Merge: Backward Transfer  
2715

Category	Linear (0.2)	Linear (0.8)
Logic	6.0 $\pm$ 0.7 (71.6)	4.6 $\pm$ 1.3 (76.6)
Knowledge/QA	5.2 $\pm$ 1.8 (58.6)	3.8 $\pm$ 0.9 (59.8)
Liberal Arts	4.6 $\pm$ 0.5 (66.6)	3.9 $\pm$ 0.6 (73.3)
Math	8.8 $\pm$ 0.3 (63.1)	6.8 $\pm$ 0.5 (70.1)
Safety/Truth	5.8 $\pm$ 0.3 (52.7)	3.4 $\pm$ 0.6 (56.1)
Science/Tech	6.3 $\pm$ 0.3 (62.9)	4.9 $\pm$ 0.3 (68.1)
<b>Total</b>	<b>6.1 <math>\pm</math>0.1 (62.6)</b>	<b>4.6 <math>\pm</math>0.2 (67.3)</b>

2754 G.9 QWEN2.5 INSTRUCT AND OPENTHINKER 7B MERGE (RELATIVE TO OPENTHINKER)  
27552756 Table 30: Qwen2.5 Instruct and OpenThinker 7B Merge: Forgetting  
2757

Category	Linear (0.2)	Linear (0.8)
Logic	7.1 $\pm$ 0.8 (64.0)	3.7 $\pm$ 1.7 (64.0)
Knowledge/QA	7.3 $\pm$ 1.6 (52.2)	7.0 $\pm$ 1.2 (52.2)
Liberal Arts	8.5 $\pm$ 0.2 (59.2)	5.5 $\pm$ 0.4 (59.2)
Math	14.3 $\pm$ 1.6 (72.1)	9.4 $\pm$ 1.5 (72.1)
Safety/Truth	10.3 $\pm$ 2.3 (52.2)	7.2 $\pm$ 1.3 (52.2)
Science/Tech	9.2 $\pm$ 0.2 (60.2)	6.8 $\pm$ 0.6 (60.2)
<b>Total</b>	9.4 $\pm$ 0.6 (60.0)	6.6 $\pm$ 0.6 (60.0)

2767 Table 31: Qwen2.5 Instruct and OpenThinker 7B Merge: Backward Transfer  
2768

Category	Linear (0.2)	Linear (0.8)
Logic	12.8 $\pm$ 1.6 (71.6)	13.1 $\pm$ 3.7 (76.6)
Knowledge/QA	12.1 $\pm$ 2.5 (58.6)	12.6 $\pm$ 2.0 (59.8)
Liberal Arts	14.1 $\pm$ 0.2 (66.6)	16.1 $\pm$ 0.6 (73.3)
Math	7.5 $\pm$ 1.5 (63.1)	7.9 $\pm$ 1.1 (70.1)
Safety/Truth	10.7 $\pm$ 1.0 (52.7)	10.1 $\pm$ 1.1 (56.1)
Science/Tech	11.2 $\pm$ 0.8 (62.9)	12.7 $\pm$ 0.4 (68.1)
<b>Total</b>	11.4 $\pm$ 0.7 (62.6)	12.1 $\pm$ 1.0 (67.3)

2808    **G.10 QWEN2.5 INSTRUCT AND OPENTHINKER3 7B MERGE (RELATIVE TO QWEN2.5**  
 2809    **INSTRUCT)**

2811    Table 32: Qwen2.5 Instruct and OpenThinker3 7B Merge: Forgetting

Category	Linear (0.2)	Linear (0.8)
Logic	$33.6 \pm 2.7$ (75.2)	$28.1 \pm 1.8$ (74.2)
Knowledge/QA	$25.7 \pm 8.0$ (60.3)	$18.0 \pm 5.2$ (45.0)
Liberal Arts	$35.0 \pm 2.9$ (74.4)	$28.8 \pm 1.7$ (73.8)
Math	$40.1 \pm 6.1$ (65.3)	$30.0 \pm 0.7$ (65.3)
Safety/Truth	$36.2 \pm 2.8$ (64.1)	$22.8 \pm 3.7$ (61.0)
Science/Tech	$30.1 \pm 3.2$ (67.4)	$29.6 \pm 0.9$ (67.4)
<b>Total</b>	$33.5 \pm 3.7$ (67.8)	$26.0 \pm 0.8$ (63.8)

2822    Table 33: Qwen2.5 Instruct and OpenThinker3 7B Merge: Backward Transfer

Category	Linear (0.2)	Linear (0.8)
Logic	$1.3 \pm 2.0$ (31.3)	$2.3 \pm 1.0$ (39.9)
Knowledge/QA	$1.9 \pm 1.0$ (28.8)	$1.2 \pm 1.4$ (21.7)
Liberal Arts	$1.9 \pm 0.8$ (30.2)	$2.9 \pm 0.1$ (39.2)
Math	$0.7 \pm 0.8$ (12.6)	$1.6 \pm 0.8$ (27.0)
Safety/Truth	$0.9 \pm 1.1$ (15.8)	$2.9 \pm 0.6$ (34.6)
Science/Tech	$2.5 \pm 1.0$ (30.4)	$2.5 \pm 0.2$ (30.9)
<b>Total</b>	$1.5 \pm 1.1$ (24.9)	$2.2 \pm 0.3$ (31.7)

2862 G.11 QWEN2.5 INSTRUCT AND OPENTHINKER3 7B MERGE (RELATIVE TO OPENTHINKER3)  
28632864 Table 34: Qwen2.5 Instruct and OpenThinker3 7B Merge: Forgetting  
2865

2866 Category	2867 Linear (0.2)	2868 Linear (0.8)
2869 Logic	28.1 $\pm$ 3.2 (65.4)	25.1 $\pm$ 2.6 (66.9)
2870 Knowledge/QA	21.2 $\pm$ 6.4 (54.3)	17.2 $\pm$ 7.3 (38.6)
2871 Liberal Arts	31.3 $\pm$ 2.9 (68.9)	27.2 $\pm$ 1.7 (68.3)
2872 Math	52.3 $\pm$ 6.9 (81.1)	41.8 $\pm$ 1.8 (81.1)
2873 Safety/Truth	33.5 $\pm$ 2.1 (59.8)	23.5 $\pm$ 1.7 (58.6)
2874 Science/Tech	32.0 $\pm$ 3.5 (70.7)	32.2 $\pm$ 1.4 (70.5)
2875 <b>Total</b>	2876 33.1 $\pm$ 3.8 (66.7)	2877 28.0 $\pm$ 1.2 (63.8)

2878 Table 35: Qwen2.5 Instruct and OpenThinker3 7B Merge: Backward Transfer  
2879

2880 Category	2881 Linear (0.2)	2882 Linear (0.8)
2883 Logic	3.0 $\pm$ 5.0 (31.3)	4.9 $\pm$ 0.9 (39.9)
2884 Knowledge/QA	2.2 $\pm$ 0.7 (28.8)	4.5 $\pm$ 2.3 (21.7)
2885 Liberal Arts	2.3 $\pm$ 0.8 (30.2)	5.4 $\pm$ 0.7 (39.2)
2886 Math	0.5 $\pm$ 0.6 (12.6)	1.5 $\pm$ 0.2 (27.0)
2887 Safety/Truth	1.0 $\pm$ 0.8 (15.8)	5.5 $\pm$ 0.4 (34.6)
2888 Science/Tech	2.0 $\pm$ 0.7 (30.4)	2.8 $\pm$ 0.3 (30.9)
2889 <b>Total</b>	2890 1.8 $\pm$ 1.3 (24.9)	2891 4.0 $\pm$ 0.4 (31.7)

2916 **DISCLAIMER FOR USE OF LLMs**  
29172918 We primarily used LLMs in coding co-pilot applications to facilitate experimentation and help with  
2919 plotting code for result presentation. LLMs were also used as writing tools to assist in refining the  
2920 paper. However, the final version was carefully reviewed and finalized by the authors. No LLMs  
2921 were used in ideation and experimental design.  
2922  
2923  
2924  
2925  
2926  
2927  
2928  
2929  
2930  
2931  
2932  
2933  
2934  
2935  
2936  
2937  
2938  
2939  
2940  
2941  
2942  
2943  
2944  
2945  
2946  
2947  
2948  
2949  
2950  
2951  
2952  
2953  
2954  
2955  
2956  
2957  
2958  
2959  
2960  
2961  
2962  
2963  
2964  
2965  
2966  
2967  
2968  
2969