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 shows that: (1) Domain-continual pretraining induces moderate forgetting with low backward backward transfer; (2) RL/SFT post-training applied to base models and Instruction tuning yield substantial backward transfer with minimal 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, its considered common knowledge in continual learning that this sequential training would lead to catastrophic forgetting (see tab:forgetlit). 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 inflates accuracy and can create illusory transitions: an apparent "$1 \rightarrow 0$" may simply be a lucky guess that later becomes an incorrect answer, even when the underlying knowledge did not change;

likewise for $0 \rightarrow 1$ transitions. When the answer is only among few options (e.g., 4), performance by random guessing can account for a substantial share of observed transitions, distorting both level and trend estimates of forgetting. Thus a principled metric should (i) resolve outcomes at the *item* level and (ii) explicitly correct for chance.

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

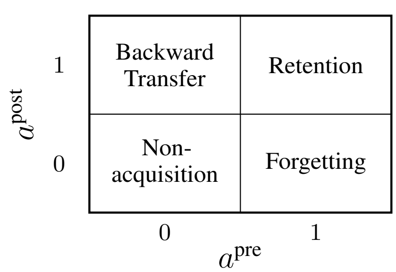
Our primary contribution is a large-scale study measuring forgetting caused by post-training across post-training pipelines. By evaluating the models on the same set of samples before and after each stage, we obtain a map of what was retained, what was forgotten, and where losses concentrate. We seek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

Key Findings

- **Domain-Continual Pretraining** induces moderate forgetting across most categories; backward transfer is limited. Both effects marginally improve with increasing model scale.
- **Instruction-Tuning and SFT/RL from Base model** yield low forgetting but large backward-transfer gains across categories and model families; both effects improve with increasing model scale. Reasoning training yields lower forgetting and larger backward transfer than instruction tuning.
- **SFT/RL Reasoning Post-Training from Instruct model** has data-scale dependent behaviour: For low-data regime, it yields low forgetting and backward transfer. For high-data regime, no dominant factor robustly described theforgetting and backward transfer dynamics.
- **Model Merging** does not reliably mitigate forgetting across post-training pipelines (yet).

2 Measuring Samplewise Forgetting and Backward Transfer

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. 2, each sample falls into one of four quadrants based on effect by training on new task:



(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 ($0 \rightarrow 0$).

We define sample-wise *forgetting* and *backward transfer* as the proportions of $1 \rightarrow 0$ and $0 \rightarrow 1$ flips, respectively:

$$\mathrm{F} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}\{a_i^{\text{pre}} = 1 \wedge a_i^{\text{post}} = 0\} \tag{1}$$

$$\mathrm{BT} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}\{a_i^{\text{pre}} = 0 \wedge a_i^{\text{post}} = 1\} \tag{2}$$

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 $\mathrm{F} = 0.25$ because $0.5 \times 0.5 = 0.25$.

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

Stage	Name	Level	Summary
CPT (§3.1)	Investigating Continual Pretraining in LLMs: Insights and Implications (Yıldız et al., 2024)	Med	Most models show continual improvement; only Llama-2 models degrade.
	Examining Forgetting in Continual Pre-training of Aligned LLMs (Li & Lee, 2024a)	High	Continual pre-training degrades capabilities, alignment and alters output behavior.
SFT/DPO (§3.2)	Mitigating Forgetting in LLM Supervised Fine-Tuning and Preference Learning (Fernando et al., 2024)	Low	Combining SFT and DPO sequentially leads to forgetting and a poor balance between goals ($\sim 2\%$ on MMLU).
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)	Med	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).

A chance baseline for flips. 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 $(k-1)/k$,

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

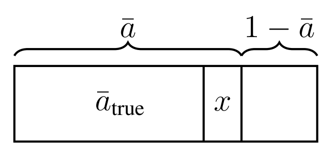


Figure 1: Accuracy $\bar{a}$ decomposes into true knowledge $\bar{a}_{\text{true}}$ and lucky guesses x.

A $1 \to 0$ 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,

$$\mathrm{F}_{\text{chance}} = \underbrace{\frac{1-\bar{a}^{\text{pre}}}{k-1}}_{\text{correct by chance (pre)}} \cdot \underbrace{\left(1-\bar{a}^{\text{post}}\right)}_{\text{incorrect (post)}}, \qquad \mathrm{BT}_{\text{chance}} = \underbrace{\left(1-\bar{a}^{\text{pre}}\right)}_{\text{incorrect (pre)}} \cdot \underbrace{\frac{1-\bar{a}^{\text{post}}}{k-1}}_{\text{correct by chance (post)}}.$$

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

Chance-adjusted forgetting and backward transfer. To isolate knowledge change beyond chance, subtract the baselines and clip at zero:

$$\mathrm{F}_{\text{true}} = \max\!\left(\mathrm{F} - \mathrm{F}_{\text{chance}},\, 0\right), \qquad \mathrm{BT}_{\text{true}} = \max\!\left(\mathrm{BT} - \mathrm{BT}_{\text{chance}},\, 0\right).$$

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.

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:

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

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

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

By construction $\mathrm{F}_{\text{true}} \leq \mathrm{F}_{\max}$ and $\mathrm{BT}_{\text{true}} \leq \mathrm{BT}_{\max}$. 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_{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 $\mathrm{BT}_{\text{true}}$ often reflect improved elicitation rather than newly acquired knowledge.

3 When, What & How Much is Pretraining Knowledge Forgotten?

In this section, we ask three questions: when, what, and how much pretraining knowledge do post-trained LLMs forget?

1. When is pretraining knowledge forgotten? Our analysis spans four widely used continual-training regimes: (i) domain-continual training (§3.1), (ii) instruction tuning (§3.2), (iii) light SFT/RL on reasoning traces, and (iv) large-scale SFT/RL for reasoning (§3.3). In total, we evaluate almost 30 model–training combinations chosen to reflect common practice, providing broad coverage of how contemporary LLMs are post-trained in the wild. Each post-trained model is compared with its initial checkpoint (details in the Appendix).

2. What pretraining knowledge is forgotten? We evaluate each model on 12 public benchmarks, collectively subdivided into close to a 100 total subdomains. To summarize systematic patterns, we cluster sub-benchmarks into nine semantically coherent groups that exhibit similar forgetting trends (e.g., common sense, culture, deduction, language/communication, liberal arts, science/tech). These clusters provide an better map of which pretraining knowledge areas are most affected by a given post-training recipe.

3. How much pretraining knowledge is forgotten? Unless stated otherwise, chance-adjusted metrics for forgetting (F_{true}) and backward transfer ($\mathrm{BT}_{\text{true}}$) are used to quantify the severity.

Experimental setup. We standardize settings across models for fair comparison. All experiments use the `LightEval` framework (Habib et al., 2023) and log per-sample accuracy. We apply a zero-shot chain-of-thought prompt to all models and require answers in a fixed MCQ format (see Appendix); base models receive a few-shot prompt solely to teach the format. When available[1], we add chat-specific templates to be in line with best practices. We cap sequence length at 32K tokens, except for Qwen2.5-7B-Math and Qwen2.5-7B-Math-Instruct Yang et al. (2024a), which are limited to 4K[2]. Decoding uses temperature 0.6 with nucleus sampling ($p = 0.95$). We provide additional details in the Appendix. To facilitate reproducibility and further inquiry, we will release per-sample logs for every sub-benchmark alongside code.

We now showcase our results in the subsections below.

3.1 Subarea 1: Domain-Continual Pretraining

Motivation. A popular class of continual learning works adapt general LLMs at the application layer for domains such as coding, mathematics, search, and tool use. As generalist LLMs are increasingly

[1]This budget was sufficient in practice, we never required more tokens.

[2]Because base models sometimes continue into subsequent questions, we set explicit stop sequences to end generation once a prediction is produced.

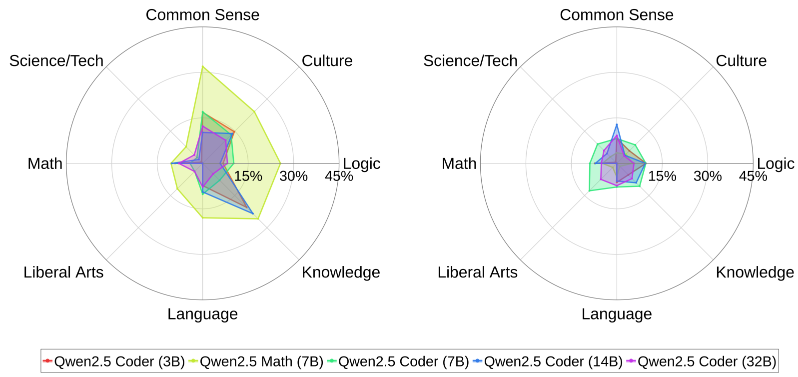


Figure 2: **Forgetting (left) and Backward Transfer (right) after domain-continual pretraining.** Forgetting is moderate and consistent across categories; backward transfer is low. Scaling model size reduces forgetting and marginally improves backward transfer.

wrapped with tools and domain-specific interfaces, specialization must not erode broad pretraining knowledge. Models still need to contextualize domain outputs, communicate with diverse users, respect cultural norms, and uphold safety and ethical standards. These needs motivate our study of forgetting and backward transfer under domain-continual pretraining.

Setup. We study continual pretraining that converts a general base model into a specialized one, exemplified by Qwen2.5-Coder (Hui et al., 2024) and Qwen2.5-Math (Yang et al., 2024b).[3] Unlike general instruction tuning or reasoning post-training, domain-continual pretraining shifts the underlying representation using large, relatively uncurated, web-scale domain corpora.

Main results. Figure 2 summarizes our findings. Domain-continual pretraining induces moderate-large amount of forgetting among all post-training methods we evaluate. Backward transfer to general abilities is weak: gains in the specialized domain rarely improve non-target tasks. The effect spans categories of pretraining knowledge, no single category drives it, although math-specialized models show significantly more forgetting. Lastly, larger models forget less and have marginally better backward transfer.

Qualitative analysis. We performed manual errors analyses to check causes of forgetting. Our analysis indicates reduced instruction-following fidelity (e.g., weaker adherence to constraints, formats, and role-specific directives). Domain-continual pretraining appears to overfit to domain-specific behaviors. For example, a coder model may answer "Who was the president of the US?" by burying the answer within code. This is the primary driver of performance degradation, with worse answer accuracy an important factor.

Takeaway

Domain-continual pretraining yields moderate forgetting across categories; backward transfer is limited. Scaling model size reduces forgetting and marginally improves backward transfer.

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?

[3]We treat domain-continual reasoning via SFT/RL separately in §3.3 and focus on domain-continual training here.

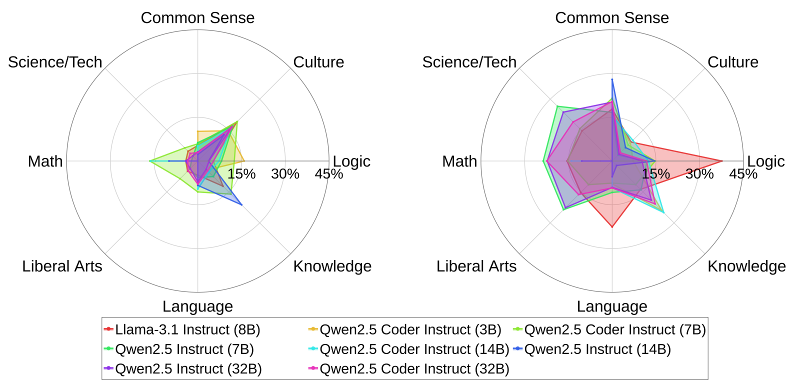


Figure 3: **Forgetting (left) and Backward Transfer (right) after instruction-tuning.** yields moderate forgetting categories-wise with large backward transfer. Scaling model size reduces forgetting and marginally improves backward transfer.

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)[4].

Results. As shown in Figure 3, there is minimal forgetting across models, with small spikes in the Culture cluster. However, there is large backward transfer in all categories except culture and language. Furthermore, scaling model size reduces forgetting and increases backward transfer. This effect is consistent across domain-general and domain-specific base models. These findings suggest that much of the continual-learning literature aims to reduce forgetting in a regime where forgetting is already minimal. Shifting focus to other subareas of post-training might spur interesting research directions.

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 forgetting but large backward-transfer gains across categories and model families; both effects improve with increasing model scale.

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 4, we see that across scales, model families, and training types, we observe large gains in backward transfer with minimal forgetting similar in nature to instruction tuning. Similarly,

[4]Qwen2.5-Math Instruct is surprisingly tuned with GRPO which leads to it being classified under Reasoning

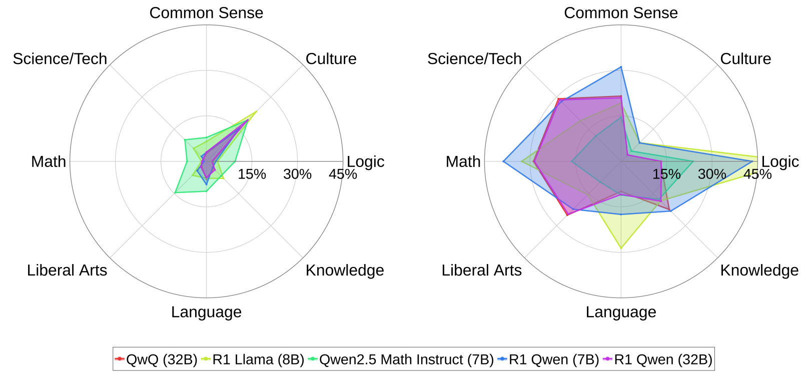


Figure 4: **Forgetting (left) and Backward Transfer (right) after reasoning training (SFT/RL) from base model. Yields minimal forgetting and has high backward-transfer gains.**

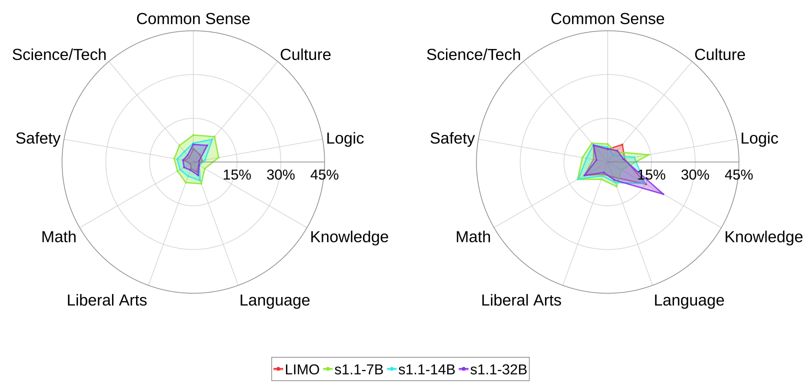


Figure 5: **Forgetting (left) and Backward Transfer (right) after reasoning training from instruct: low data scenario.** Yields little forgetting and backward transfer. Forgetting decreases with model scale.

forgetting is moderate for the lone category of Culture and correspondingly backward transfer there is low. However, when compared to instruction tuning on the same base model (Figure 3)[5], we see lower forgetting and larger back-transfer.

Takeaway

Training with SFT/RL for reasoning yields minimal forgetting and large backward-transfer gains, similar in nature to instruction tuning. However, compared to instruction tuning, we see lower forgetting and larger back-transfer.

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 (low data) behave differently from those trained at scale (high data). We therefore present these two cases separately.

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 (Ye et al., 2025) all tuned from corresponding sized Qwen instruct models.

[5] Corresponding tables are available in Appendix E.4 for detailed comparison

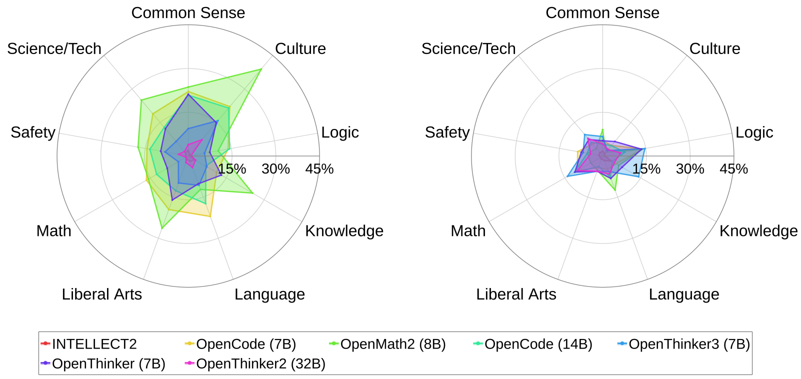


Figure 6: **Forgetting (left) and Backward Transfer (right) after reasoning training from instruct: high data scenario.** No single factor robustly explains the dynamics of forgetting and backward transfer

Results. Figure 5 summarizes our findings. Across categories, models show minimal forgetting and low backward transfer. This makes sense, training for a few passes on little data leaves pretraining knowledge largely intact: 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.

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 fall with a narrowing student-teacher gap.

3.3.3 Reasoning Training from Instruction-Tuned Models: High-Data Scenario

Models. We evaluate OpenCodeReasoner and OpenMath2 (Bercovich et al., 2025), OpenThinker-7B, Openthinker2-32B, and OpenThinker3-7B (Guha et al., 2025), and Intellect-2-32B (Prime Intellect Team et al., 2025). This spans SFT (former) and RL (Intellect-2).

Results. Results vary by domain mix and model quality. The OpenThinker models shows low–to–moderate forgetting and moderate backward transfer, perhaps due to the breadth of the training datamix, whereas OpenCodeReasoner models show consistently high forgetting with low backward transfer gains due to the narrower training data. Scaling model size, if compared in Openthinker models, signals improvements in both forgetting and backward transfer – as seen in most previous sections. Decentralized training, as in Intellect-2, in contrast showed minimal forgetting or backward transfer. We conjecture that the model largely remain unchanged compared to the original model as it shows negligible gains on the optimized math benchmarks Hochlehnert et al. (2025). However, the results here remain preliminary. We do not find a single dominant factor—initialization, data regime, or scale that sufficiently explains forgetting and backward-transfer dynamics. We believe controlling the finer details which determine the quality of the trained model might lead to better conclusions.

Takeaway

No single factor robustly explains the dynamics of forgetting and backward transfer – training on a mix of domains seems to improve both forgetting and backward transfer.

4 Does Model Merging Reduce Forgetting?

textbfMotivation. Recent work shows that offline model merging can combine capabilities from multiple models (Dziadzio et al., 2025). Unlike classical continual learning (De Lange et al., 2022),

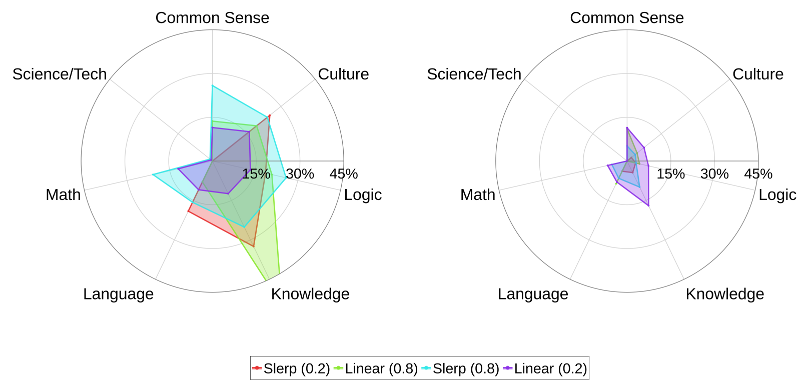


Figure 7: **Qwen2.5 Base with Coder Merge (7B) Forgetting and Backward-Transfer relative to Qwen2.5 Base (without MMLU)**. Induces large forgetting and little backward transfer.

it requires neither the original training data nor the ability to resume training, which is practical in resource-constrained settings.

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

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

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

Results. We compare merged checkpoints (LERP/SLERP) to the post-trained model θ_{post}; results for θ_{pre} appear in the Appendix. For Qwen2.5-Coder-7B and OpenThinker3-7B, even small mixes with the base checkpoint severely degrade performance (Figures 7, 10). In contrast, OpenThinker-7B shows small overall gains, accompanied by moderate forgetting (Figure 11). 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

Model merging does not yet 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 increased scale can offset these difficulties.

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→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.

REFERENCES

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

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

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

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

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

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

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

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

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

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

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

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

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

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. URL `https://www.sciencedirect.com/science/article/pii/S1364661399012942`.

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

Nathan Habib, Clémentine Fourrier, Hynek Kydlíček, Thomas Wolf, and Lewis Tunstall. Lighteval: A lightweight framework for llm evaluation, 2023. URL `https://github.com/huggingface/lighteval`.

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

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

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

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

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

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

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

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

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

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

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

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

Yuetai Li, Zhangchen Xu, Fengqing Jiang, Bhaskar Ramasubramanian, Luyao Niu, Bill Yuchen Lin, Xiang Yue, and Radha Poovendran. Temporal sampling for forgotten reasoning in LLMs. In *Second Workshop on Test-Time Adaptation: Putting Updates to the Test! at ICML 2025*, 2025. URL https://openreview.net/forum?id=J0HWRSSpSJ.

Stephanie Lin, Jacob Hilton, and Owain Evans. TruthfulQA: Measuring how models mimic human falsehoods. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 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.229/.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Vinay Venkatesh Ramasesh, Aitor Lewkowycz, and Ethan Dyer. Effect of scale on catastrophic forgetting in neural networks. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=GhVS8_yPeEa.

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

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

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

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

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

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

Zayne Rea Sprague, Xi Ye, Kaj Bostrom, Swarat Chaudhuri, and Greg Durrett. MuSR: Testing the limits of chain-of-thought with multistep soft reasoning. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=jenyYQzue1.

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

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

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

Mariya Toneva, Alessandro Sordoni, Remi Tachet des Combes, Adam Trischler, Yoshua Bengio, and Geoffrey J. Gordon. An empirical study of example forgetting during deep neural network learning. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=BJlxm30cKm.

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

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

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

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

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

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

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

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

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

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

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

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

Appendix

CONTENTS

A Extended Related Works

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

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

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

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

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

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

Mitigation strategies. Sequential SFT to RLHF/DPO can exacerbate forgetting. To counteract this, researchers explore: (i) *model averaging*, interpolating between pre- and post-RLHF checkpoints to trade off alignment and retention (Lin et al., 2024); (ii) *joint post-training*, optimizing supervised and preference objectives simultaneously with convergence guarantees (Fernando et al., 2024); and (iii) *unified fine-tuning (UFT)*, which folds instruction tuning and alignment into a single implicit-reward objective (Wang et al., 2025). Additional techniques—including advantage models and selective rehearsal—stabilize RLHF by shaping reward distributions and replaying curated data (Peng et al., 2023). *Online Merging Optimizers (OMO)* combine gradients from SFT and RLHF models during training to maximize reward while preserving pre-trained skills (Lu et al., 2024). Theory supports these interventions: up to permutation symmetries, weights of homologous models tend to lie in a shared low-loss basin (Ainsworth et al., 2023). Hence, we were quite surprised that model merging does not work for our simple case of mitigating forgetting during post-training with only two deep networks.

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

B Experimental Setup

B.1 Evaluation

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

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

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

```
{Instruction}
On the very last line, write exactly "Answer: $LETTER" (e.g. "Answer: B"),
with no extra punctuation, no lowercase, no *, and no trailing spaces.
Think step by step, showing your reasoning.
Question: "{Question}"
```

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

```
{Instruction}
Question: "{Few-Shot Question 1}"
{Example answer 1}

... <--- more examples

Question: "{Question}"
```

Datasets where CoT reasoning traces are provided for few-shot prompting, we use those. In the cases where this is not provided (PIQA, MCTest, Social-IQa, ARC, MCTest, and Hellaswag) CoT few-shot examples were generated and then confirmed these are not included in the benchmarks[1].

All experiments are conducted using the Hugging Face `LightEval` framework, with results logged at the sample level. For generation, we allow up to 32,768 tokens, which we found sufficient for models to complete their chain of thought and provide an answer. In cases where the maximum trained context length is smaller, then the generation is reduced to that number, as is the case with Qwen2.5-7B-Math and Qwen2.5-7B-Math-Instruct Yang et al. (2024a). The temperature is set to 0.6 and nucleus sampling with $p = 0.95$ is applied.

B.2 Datasets

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

They are grouped as follows:

Commonsense:

- Commonsense QA
- PIQA

Culture:

- BBH (sports understanding and movie recommendation)

Logic

- BBH (navigate, causal judgment, penguins in a table, web of lies, tracking shuffled objects three objects, tracking shuffled objects seven objects, tracking shuffled objects five objects, temporal sequences, reasoning about colored objects, logical deduction three objects, logical deduction seven objects, logical deduction five objects, formal fallacies, and date understanding)
- ARC (easy and challenge)
- MuSR (murder mysteries, object placements, and team allocation)
- MMLU (logical fallacies)

Knowledge

- BBH (object counting)
- MMLU (miscellaneous and global facts)
- MCTest

Language

- BBH (snarks, disambiguation qa, ruin names, and hyperbaton)
- Social IQa
- Hellaswag
- BBH (salient translation error detection)

Liberal Arts

- 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)

Math

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

Safety [2]

- MMLU (moral scenarios, moral disputes, jurisprudence, and business ethics)
- TruthfulQA (mc1)
- SaladBench (mrq)

Science & Tech

- 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)
- GPQA (diamond)

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

[1] MMLU is evaluated with few-shot no CoT prompting for the base models

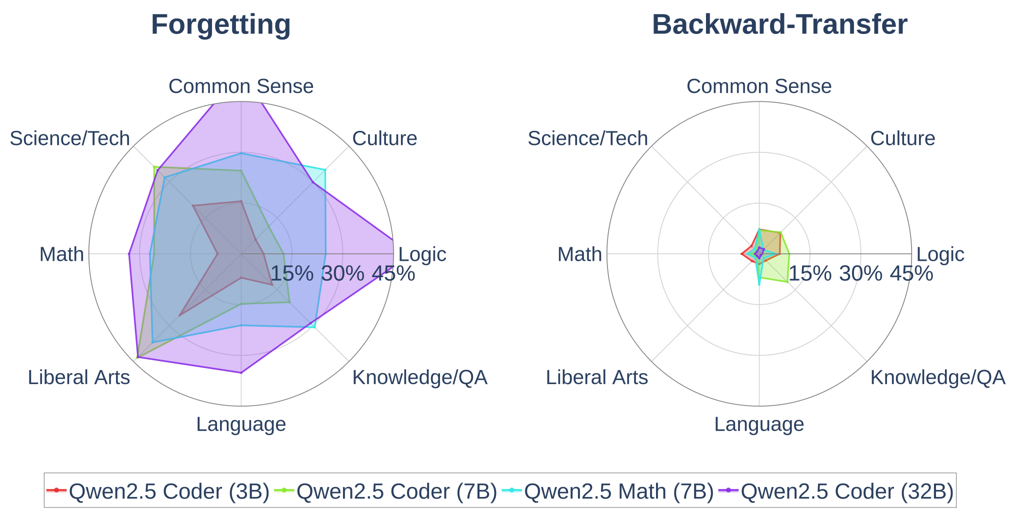
[2] These are only used in comparisons which do not include a base model because TruthfulQA and SaladBench are designed measure the default behavior of the model rather than knowledge, which few-shot prompting would bias.

C Evaluation Comparison

C.1 Prompting Method

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

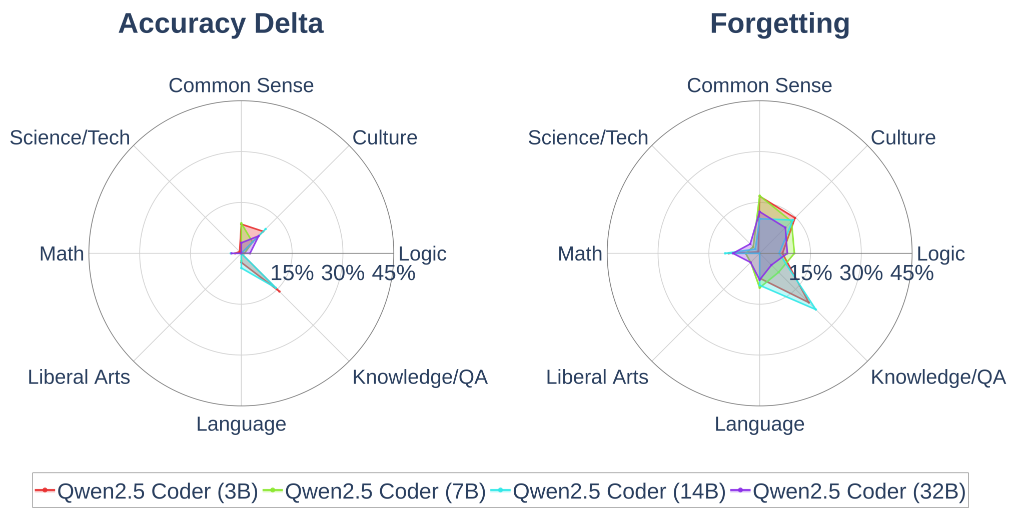
Figure 8: Coder model using chat template prompting



Due to this, measuring performance of base models on certain datasets can become nontrivial. While benchmarks measuring knowledge or capabilities may be elicted through few-shot prompting, others, such as truthfulness or safety become more difficult as prompting them with examples would bias their behavior, invalidating the purpose of the dataset. Further works consider exploring the effect of providing no-knowledge few-shot prompting, where the format of the question and answer is provided without leaking examples biasing the models output.

C.2 Metric

Figure 9: Coder model comparing conventional forgetting (left) against sample-wise forgetting (right)



Additionally, the sample-wise nature of the introduced metric allows more forgetting to be uncovered than is possible with the standard metric, the difference between the accuracy after training from the accuracy before training, clipped at 0. We demonstrate this on the example of the forgetting when undergoing domain continual pretraining for Qwen2.5 to Qwen2.5 Coder (7B) when using the standard metric

From the Figure 9 there appears to be little forgetting when using the conventional forgetting. However, using sample-wise data, moderate forgetting can be seen.

D MODEL MERGING: RESULTS

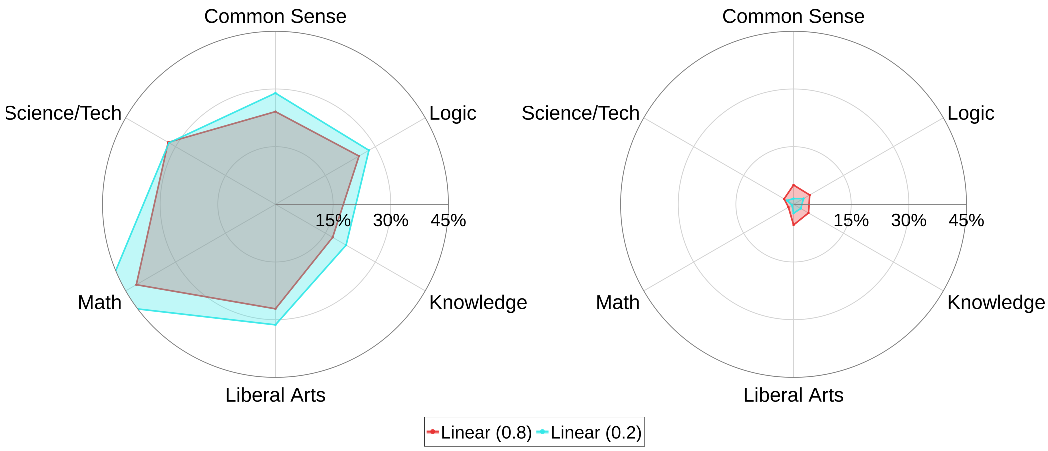


Figure 10: Qwen2.5 Instruct with OpenThinker3 Merge (7B) Forgetting and Backward-Transfer relative to OpenThinker3 (on MMLU)

MODERATE SUCCESS CASE

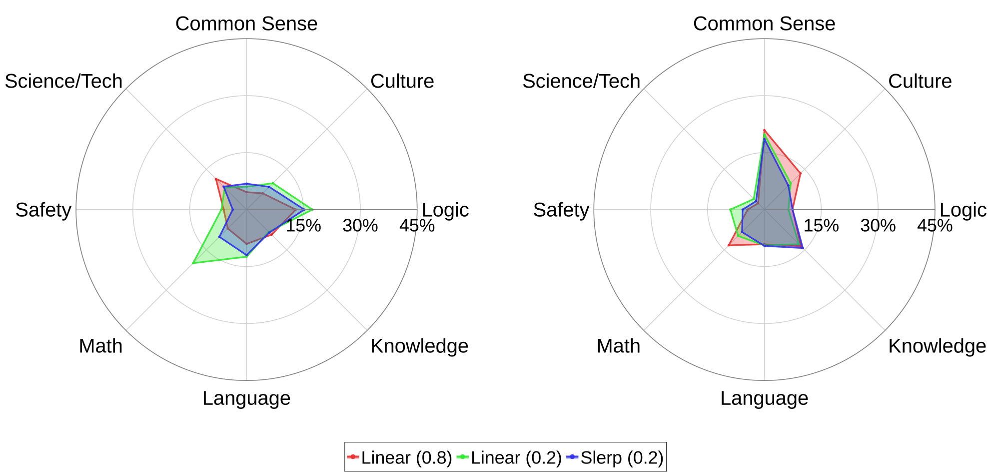


Figure 11: Qwen2.5 Instruct with OpenThinker Merge (7B) Forgetting and Backward-Transfer relative to OpenThinker (all benchmarks)

E QUANTIFYING FORGETTING ACCURATELY (TABLES FOR REFERENCING PLOTS)

E.1 INSTRUCTION TUNING (GENERAL)

Table 2: Instruction Tuning (General): Forgetting (Part 1 of 2)

Category	Qwen2.5-14B-Instruct	Qwen2.5-32B-Instruct	Qwen2.5-7B-Instruct
Common Sense	2.2 ±0.7 (42.6)	3.1 ±0.3 (61.8)	6.3 ±0.8 (63.0)
Culture	12.7 ±0.9 (74.7)	16.2 ±1.3 (79.5)	17.0 ±3.8 (70.2)
Logic	3.7 ±0.5 (67.1)	3.9 ±0.3 (71.8)	8.1 ±0.6 (51.0)
Knowledge/QA	21.2 ±6.3 (21.9)	4.1 ±5.1 (39.7)	7.5 ±6.0 (43.4)
Language	8.3 ±1.5 (57.2)	7.1 ±0.5 (54.9)	5.7 ±2.8 (33.1)
Liberal Arts	-	3.7 ±0.3 (55.4)	3.8 ±0.5 (46.6)
Math	9.9 ±2.1 (70.9)	4.2 ±0.6 (53.7)	2.2 ±0.6 (37.1)
Safety/Truth	2.4 ±2.6 (7.4)	2.0 ±0.3 (29.4)	4.3 ±1.7 (24.3)
Science/Tech	-	2.5 ±0.3 (47.7)	2.4 ±0.2 (33.2)
Total	7.8 ±0.4 (53.1)	4.8 ±0.5 (56.2)	6.1 ±1.0 (45.7)

Table 3: Instruction Tuning (General): Forgetting (Part 2 of 2)

Category	Llama-3.1-8B-Instruct
Common Sense	5.3 ±1.3 (53.6)
Culture	18.9 ±5.9 (65.0)
Logic	5.3 ±0.4 (18.3)
Knowledge/QA	12.2 ±2.8 (48.2)
Language	4.9 ±2.8 (22.2)
Liberal Arts	5.0 ±0.3 (53.7)
Math	4.1 ±0.5 (32.8)
Safety/Truth	6.8 ±0.8 (32.6)
Science/Tech	4.7 ±0.3 (46.8)
Total	6.9 ±1.1 (40.3)

Table 4: Instruction Tuning (General): Backward Transfer (Part 1 of 2)

Category	Qwen2.5-14B-Instruct	Qwen2.5-32B-Instruct	Qwen2.5-7B-Instruct
Common Sense	28.0 ±3.9 (76.9)	20.2 ±3.3 (84.6)	16.7 ±2.8 (76.8)
Culture	6.5 ±0.6 (65.0)	3.3 ±1.1 (61.4)	2.8 ±2.1 (43.8)
Logic	14.6 ±1.3 (81.7)	10.2 ±0.5 (80.2)	11.9 ±4.3 (55.5)
Knowledge/QA	2.2 ±1.0 (1.8)	19.0 ±1.1 (60.4)	14.0 ±3.1 (54.8)
Language	5.4 ±0.6 (51.6)	9.1 ±0.6 (57.1)	10.8 ±1.2 (38.7)
Liberal Arts	-	22.6 ±1.6 (80.8)	23.6 ±0.6 (73.3)
Math	10.4 ±2.1 (73.1)	22.3 ±1.3 (78.8)	23.6 ±1.5 (66.8)
Safety/Truth	27.3 ±5.5 (42.1)	27.9 ±2.5 (64.3)	23.1 ±2.6 (50.1)
Science/Tech	-	23.7 ±2.1 (76.0)	26.5 ±1.2 (65.2)
Total	13.0 ±0.7 (60.8)	18.0 ±1.4 (73.9)	16.6 ±0.7 (59.2)

Table 5: Instruction Tuning (General): Backward Transfer (Part 2 of 2)

Category	Llama-3.1-8B-Instruct
Common Sense	17.8 ±5.8 (70.3)
Culture	9.2 ±5.6 (46.2)
Logic	37.6 ±1.3 (49.6)
Knowledge/QA	13.7 ±1.4 (52.9)
Language	22.6 ±6.5 (42.2)
Liberal Arts	15.3 ±2.2 (67.5)
Math	15.5 ±0.8 (46.9)
Safety/Truth	11.4 ±1.8 (38.5)
Science/Tech	14.6 ±0.9 (60.1)
Total	20.0 ±1.7 (55.9)

Table 6: Instruction Tuning (Task): Forgetting (Part 1 of 2)

Category	Qwen2.5-Coder-14B-Instruct	Qwen2.5-Coder-32B-Instruct	Qwen2.5-Coder-3B-Instruct
Common Sense	5.1 ±0.3 (62.3)	2.9 ±0.5 (57.7)	10.2 ±1.6 (49.8)
Culture	15.9 ±2.0 (54.1)	16.6 ±0.9 (68.2)	14.7 ±3.6 (38.7)
Logic	7.3 ±0.6 (75.8)	5.4 ±0.3 (66.8)	15.9 ±0.6 (38.0)
Knowledge/QA	5.4 ±1.7 (35.6)	4.7 ±0.7 (43.9)	4.4 ±4.1 (25.2)
Language	9.6 ±1.0 (49.9)	8.0 ±0.5 (55.3)	8.0 ±0.8 (20.7)
Liberal Arts	-	4.7 ±0.5 (60.6)	-
Math	16.2 ±2.4 (45.2)	3.4 ±1.0 (50.6)	14.4 ±1.2 (39.6)
Safety/Truth	2.7 ±1.7 (19.0)	4.4 ±0.6 (35.1)	5.1 ±5.5 (10.4)
Science/Tech	-	3.7 ±0.5 (50.3)	-
Total	8.0 ±0.8 (50.7)	5.5 ±0.3 (56.4)	9.5 ±1.0 (31.4)

Table 7: Instruction Tuning (Task): Forgetting (Part 2 of 2)

Category	Qwen2.5-Coder-7B-Instruct
Common Sense	5.9 ±0.7 (50.1)
Culture	19.1 ±5.0 (57.1)
Logic	12.5 ±1.1 (53.4)
Knowledge/QA	16.1 ±2.7 (55.0)
Language	10.6 ±1.9 (37.1)
Liberal Arts	8.6 ±1.0 (59.4)
Math	16.4 ±11.4 (47.8)
Safety/Truth	5.6 ±3.3 (26.0)
Science/Tech	6.3 ±1.2 (43.8)
Total	10.9 ±1.4 (48.6)

Table 8: Instruction Tuning (Task): Backward Transfer (Part 1 of 2)

Category	Qwen2.5-Coder-14B-Instruct	Qwen2.5-Coder-32B-Instruct	Qwen2.5-Coder-3B-Instruct
Common Sense	15.5 ±1.6 (76.2)	20.4 ±2.3 (81.0)	16.4 ±4.4 (58.1)
Culture	10.0 ±1.5 (47.1)	4.0 ±0.8 (50.2)	8.8 ±4.5 (30.3)
Logic	12.4 ±0.6 (82.0)	11.8 ±0.8 (75.0)	7.2 ±1.3 (25.8)
Knowledge/QA	25.0 ±3.3 (60.2)	20.8 ±2.4 (64.7)	24.0 ±2.8 (50.4)
Language	9.2 ±1.1 (47.9)	9.2 ±1.0 (55.6)	7.8 ±0.5 (20.4)
Liberal Arts	-	16.3 ±3.8 (76.1)	-
Math	14.0 ±2.8 (42.8)	22.4 ±2.1 (76.1)	4.0 ±1.1 (27.7)
Safety/Truth	13.0 ±2.2 (32.9)	21.8 ±1.4 (57.5)	18.6 ±4.0 (29.0)
Science/Tech	-	18.9 ±3.8 (70.8)	-
Total	15.2 ±1.2 (60.0)	16.1 ±1.3 (70.2)	15.2 ±1.3 (39.1)

Table 9: Instruction Tuning (Task): Backward Transfer (Part 2 of 2)

Category	Qwen2.5-Coder-7B-Instruct
Common Sense	21.5 ±2.1 (70.9)
Culture	7.8 ±7.0 (39.9)
Logic	14.8 ±2.8 (55.1)
Knowledge/QA	11.4 ±0.7 (51.6)
Language	7.6 ±1.5 (30.5)
Liberal Arts	11.5 ±3.1 (63.3)
Math	15.3 ±4.2 (48.3)
Safety/Truth	14.4 ±3.2 (38.4)
Science/Tech	15.7 ±3.1 (56.2)
Total	14.3 ±1.2 (53.1)

Table 10: Domain-continual pretraining: Forgetting (Part 1 of 2)

Category	Qwen2.5-Coder-14B	Qwen2.5-Coder-32B	Qwen2.5-Coder-3B
Common Sense	10.2 ±2.2 (40.4)	12.3 ±1.6 (61.8)	16.8 ±8.4 (51.8)
Culture	13.9 ±1.4 (74.0)	10.7 ±1.4 (79.5)	14.8 ±4.8 (52.8)
Logic	5.8 ±0.4 (66.7)	8.2 ±0.6 (71.8)	6.7 ±0.7 (34.3)
Knowledge/QA	23.4 ±12.7 (29.2)	4.9 ±4.2 (39.7)	20.5 ±20.9 (23.7)
Language	9.4 ±1.5 (56.1)	7.8 ±0.6 (54.9)	7.3 ±0.8 (23.3)
Liberal Arts	-	3.8 ±1.3 (55.4)	-
Math	10.2 ±1.9 (71.8)	7.9 ±1.5 (53.7)	9.1 ±2.1 (52.0)
Safety/Truth	1.1 ±1.8 (7.3)	2.7 ±1.3 (29.4)	4.9 ±2.5 (16.6)
Science/Tech	1.8 ±1.8 (7.1)	3.9 ±1.6 (47.7)	0.6 ±1.4 (3.7)
Total	10.6 ±1.0 (48.2)	7.2 ±0.8 (56.2)	12.1 ±2.3 (35.4)

Table 11: Domain-continual pretraining: Forgetting (Part 2 of 2)

Category	Qwen2.5-Coder-7B	Qwen2.5-Math-7B
Common Sense	17.0 ±3.7 (61.9)	32.0 ±4.2 (61.9)
Culture	13.3 ±4.3 (67.9)	24.1 ±3.6 (67.9)
Logic	10.2 ±0.9 (51.3)	25.7 ±3.8 (51.3)
Knowledge/QA	7.9 ±2.9 (48.4)	25.8 ±5.1 (48.4)
Language	10.2 ±1.4 (40.0)	17.9 ±3.0 (40.0)
Liberal Arts	3.7 ±0.3 (46.6)	11.8 ±1.2 (46.6)
Math	4.2 ±0.6 (37.8)	10.5 ±1.9 (37.8)
Safety/Truth	2.9 ±0.7 (24.3)	8.1 ±2.6 (24.3)
Science/Tech	2.7 ±0.4 (33.2)	7.7 ±0.9 (33.2)
Total	9.2 ±1.0 (49.0)	20.1 ±1.4 (49.0)

Table 12: Domain-continual pretraining: Backward Transfer (Part 1 of 2)

Category	Qwen2.5-Coder-14B	Qwen2.5-Coder-32B	Qwen2.5-Coder-3B
Common Sense	12.8 ±1.3 (43.8)	9.2 ±3.4 (57.7)	8.2 ±3.2 (40.3)
Culture	3.8 ±1.1 (58.1)	3.5 ±0.4 (68.2)	5.9 ±2.0 (38.8)
Logic	9.4 ±0.8 (70.4)	5.7 ±0.9 (68.2)	9.6 ±1.0 (38.1)
Knowledge/QA	9.0 ±5.4 (14.0)	7.1 ±3.6 (43.9)	5.4 ±6.9 (7.5)
Language	5.8 ±0.4 (49.9)	7.4 ±1.0 (55.3)	6.2 ±1.1 (20.7)
Liberal Arts	-	7.5 ±3.5 (60.6)	-
Math	7.1 ±2.2 (69.4)	5.1 ±1.5 (50.6)	7.3 ±2.0 (49.7)
Safety/Truth	9.4 ±3.4 (20.5)	6.5 ±0.9 (35.1)	2.4 ±2.4 (14.5)
Science/Tech	4.6 ±0.9 (11.0)	6.0 ±3.1 (50.3)	0.5 ±0.9 (2.8)
Total	7.6 ±1.2 (44.5)	7.3 ±1.0 (56.6)	5.5 ±0.8 (26.8)

Table 13: Domain-continual pretraining: Backward Transfer (Part 2 of 2)

Category	Qwen2.5-Coder-7B	Qwen2.5-Math-7B
Common Sense	8.2 ±1.5 (50.1)	3.5 ±2.3 (23.9)
Culture	8.6 ±1.3 (59.1)	4.3 ±1.7 (37.1)
Logic	9.6 ±0.4 (49.9)	5.1 ±2.1 (28.4)
Knowledge/QA	10.7 ±2.6 (53.7)	1.4 ±1.4 (18.6)
Language	7.8 ±1.2 (37.1)	4.5 ±1.4 (21.1)
Liberal Arts	12.8 ±4.3 (59.9)	1.9 ±2.4 (31.1)
Math	8.9 ±1.6 (45.8)	4.7 ±1.6 (33.1)
Safety/Truth	9.7 ±3.4 (33.9)	1.9 ±2.1 (13.3)
Science/Tech	9.0 ±3.7 (42.6)	3.0 ±2.2 (24.6)
Total	9.2 ±1.2 (49.3)	3.2 ±1.6 (26.3)

Table 14: Trained from Base: Forgetting (Part 1 of 2)

Category	QwQ-32B	DeepSeek-R1-Distill-Qwen-32B	Qwen2.5-Math-7B-Instruct
Common Sense	2.6 ±0.4 (61.8)	3.0 ±0.4 (61.8)	7.9 ±4.0 (23.9)
Culture	15.9 ±1.3 (79.5)	19.2 ±1.3 (79.5)	19.3 ±5.4 (37.1)
Logic	2.2 ±0.2 (71.6)	1.8 ±0.3 (71.8)	9.4 ±2.1 (28.4)
Knowledge/QA	3.4 ±4.4 (39.7)	4.0 ±5.2 (39.7)	6.8 ±1.2 (18.6)
Language	5.9 ±0.8 (54.9)	5.4 ±0.6 (54.9)	9.9 ±4.4 (21.1)
Liberal Arts	2.2 ±0.2 (55.4)	2.6 ±0.4 (55.4)	14.6 ±1.6 (31.1)
Math	1.5 ±0.4 (53.7)	1.5 ±0.2 (53.7)	6.4 ±0.8 (33.1)
Safety/Truth	1.9 ±0.6 (29.4)	2.1 ±0.6 (29.4)	5.9 ±2.6 (13.3)
Science/Tech	1.3 ±0.2 (47.7)	1.4 ±0.1 (47.7)	10.0 ±1.6 (24.6)
Total	3.7 ±0.5 (56.1)	4.2 ±0.5 (56.2)	9.9 ±2.2 (26.3)

Table 15: Trained from Base: Forgetting (Part 2 of 2)

Category	DeepSeek-R1-Distill-Qwen-7B	DeepSeek-R1-Distill-Llama-8B
Common Sense	3.0 ±1.5 (25.5)	6.6 ±1.6 (53.6)
Culture	15.4 ±3.9 (37.4)	23.2 ±4.5 (65.0)
Logic	3.1 ±0.2 (28.2)	3.7 ±0.4 (18.3)
Knowledge/QA	2.9 ±2.1 (19.1)	7.9 ±2.2 (48.2)
Language	7.6 ±3.3 (26.0)	5.6 ±2.2 (22.2)
Liberal Arts	4.3 ±0.7 (31.8)	6.5 ±0.9 (53.7)
Math	0.8 ±0.1 (34.3)	1.7 ±0.4 (32.8)
Safety/Truth	1.6 ±0.9 (13.4)	6.5 ±0.2 (32.6)
Science/Tech	2.3 ±0.7 (25.8)	6.0 ±0.1 (46.8)
Total	4.2 ±1.2 (27.4)	6.9 ±0.9 (40.3)

Table 16: Trained from Base: Backward Transfer (Part 1 of 2)

Category	QwQ-32B	DeepSeek-R1-Distill-Qwen-32B	Qwen2.5-Math-7B-Instruct
Common Sense	21.5 ±3.7 (87.0)	21.0 ±3.4 (85.7)	14.5 ±2.3 (32.8)
Culture	3.0 ±0.9 (60.2)	3.0 ±0.9 (56.2)	4.8 ±0.7 (10.9)
Logic	13.0 ±0.7 (85.8)	13.2 ±0.6 (87.2)	23.7 ±3.4 (43.0)
Knowledge/QA	22.5 ±2.0 (65.9)	18.6 ±0.7 (60.8)	17.7 ±1.0 (32.0)
Language	9.9 ±0.9 (60.7)	10.9 ±0.5 (62.9)	11.1 ±1.9 (20.0)
Liberal Arts	25.1 ±1.9 (86.1)	24.6 ±1.8 (84.9)	9.7 ±0.7 (24.6)
Math	28.9 ±1.4 (90.2)	28.6 ±1.7 (90.1)	16.3 ±2.1 (45.9)
Safety/Truth	29.5 ±2.4 (66.8)	29.3 ±2.7 (66.1)	8.9 ±0.6 (17.6)
Science/Tech	29.2 ±2.3 (85.0)	28.6 ±2.2 (84.1)	11.8 ±0.8 (27.0)
Total	20.5 ±1.5 (78.5)	20.0 ±1.6 (77.5)	13.9 ±1.0 (30.0)

Table 17: Trained from Base: Backward Transfer (Part 2 of 2)

Category	DeepSeek-R1-Distill-Qwen-7B	DeepSeek-R1-Distill-Llama-8B
Common Sense	31.1 ±4.8 (63.0)	19.2 ±4.9 (70.5)
Culture	8.6 ±3.5 (23.1)	8.7 ±4.5 (36.5)
Logic	43.2 ±7.8 (74.9)	57.1 ±1.6 (74.6)
Knowledge/QA	23.2 ±1.0 (45.6)	18.4 ±1.8 (61.8)
Language	17.5 ±4.7 (39.2)	28.6 ±9.2 (49.2)
Liberal Arts	22.3 ±2.3 (56.2)	15.5 ±1.8 (65.8)
Math	38.8 ±2.8 (83.0)	32.8 ±0.7 (72.8)
Safety/Truth	21.3 ±2.8 (39.4)	17.2 ±0.7 (46.5)
Science/Tech	27.9 ±2.7 (60.4)	18.9 ±0.5 (64.0)
Total	27.4 ±3.1 (57.0)	26.1 ±1.8 (63.0)

Table 18: Trained from Instruct - High Data Scenario: Forgetting (Part 1 of 3)

Category	INTELLECT-2	OpenCodeReasoning-Nemotron-1.1-14B	OpenThinker2-32B
Common Sense	1.9 ±0.3 (87.0)	20.9 (76.2)	3.9 ±0.2 (84.6)
Culture	0.7 ±0.6 (60.2)	21.5 ±0.6 (65.0)	7.3 ±1.9 (61.4)
Logic	0.6 ±0.1 (87.0)	14.4 ±0.3 (76.5)	1.0 ±0.1 (81.2)
Knowledge/QA	2.5 ±0.6 (65.9)	6.9 ±2.6 (23.1)	2.9 ±0.5 (60.4)
Language	1.9 ±0.3 (62.7)	17.4 ±0.9 (61.7)	4.2 ±0.3 (57.1)
Liberal Arts	1.2 ±0.2 (86.1)	12.7 ±0.9 (81.5)	2.4 ±0.1 (80.8)
Math	0.9 ±0.1 (91.9)	12.5 ±0.6 (70.6)	0.9 ±0.3 (80.2)
Safety/Truth	0.9 ±0.2 (66.8)	13.4 ±1.1 (70.1)	3.4 ±0.1 (64.3)
Science/Tech	1.6 ±0.1 (85.0)	12.9 ±0.5 (75.1)	2.0 ±0.2 (76.0)
Total	1.3 ±0.1 (79.0)	14.1 ±0.6 (66.7)	2.9 ±0.3 (74.2)

Table 19: Trained from Instruct - High Data Scenario: Forgetting (Part 2 of 3)

Category	OpenCodeReasoning-Nemotron-1.1-7B	OpenThinker3-7B	OpenThinker-7B
Common Sense	22.1 ±0.1 (76.8)	9.4 ±0.5 (76.8)	21.2 ±4.0 (74.7)
Culture	22.1 ±4.3 (46.4)	15.8 ±3.9 (43.8)	14.9 ±4.4 (46.4)
Logic	13.8 ±0.9 (59.8)	5.6 ±0.8 (59.1)	7.5 ±0.3 (60.7)
Knowledge/QA	11.2 ±1.3 (51.2)	7.3 ±3.1 (54.8)	13.1 ±1.0 (51.2)
Language	22.1 ±1.7 (49.7)	10.7 ±4.3 (38.7)	10.1 ±1.2 (40.1)
Liberal Arts	19.5 ±0.7 (73.1)	9.8 ±0.4 (73.3)	16.1 ±1.1 (73.1)
Math	16.6 ±0.4 (67.6)	3.9 ±0.4 (68.4)	8.4 ±1.1 (67.6)
Safety/Truth	14.4 ±0.6 (50.1)	8.1 ±0.6 (50.1)	9.6 ±0.8 (53.6)
Science/Tech	18.9 ±0.3 (67.4)	6.5 ±0.3 (65.2)	12.2 ±0.4 (65.2)
Total	17.8 ±0.4 (62.8)	8.3 ±0.4 (59.8)	13.2 ±1.6 (62.1)

Table 20: Trained from Instruct - High Data Scenario: Forgetting (Part 3 of 3)

Category	OpenMath2-Llama3.1-8B
Common Sense	23.6 ±2.8 (53.6)
Culture	38.9 ±6.8 (65.0)
Logic	10.4 ±0.7 (18.3)
Knowledge/QA	25.4 ±3.5 (48.2)
Language	12.2 ±6.3 (22.2)
Liberal Arts	26.3 ±2.1 (53.7)
Math	15.8 ±1.0 (32.8)
Safety/Truth	17.5 ±0.8 (32.6)
Science/Tech	25.0 ±0.9 (46.8)
Total	20.8 ±1.5 (40.3)

Table 21: Trained from Instruct - High Data Scenario: Backward Transfer (Part 1 of 3)

Category	INTELLECT-2	OpenCodeReasoning-Nemotron-1.1-14B	OpenThinker2-32B
Common Sense	1.7 ±0.1 (86.7)	4.1 (53.8)	4.2 ±0.3 (85.0)
Culture	1.1 ±0.3 (60.8)	4.6 ±0.4 (37.8)	2.9 ±0.4 (54.0)
Logic	0.7 ±0.1 (87.1)	7.7 ±0.5 (66.6)	6.2 ±0.1 (88.4)
Knowledge/QA	3.4 ±0.9 (67.5)	3.8 ±1.0 (18.5)	4.6 ±0.7 (62.5)
Language	1.7 ±0.2 (62.7)	7.7 ±0.8 (47.1)	6.6 ±0.5 (61.3)
Liberal Arts	1.2 ±0.1 (86.0)	3.9 ±0.1 (69.7)	5.6 ±0.3 (85.2)
Math	0.9 ±0.2 (92.1)	11.2 ±0.9 (67.5)	10.0 ±0.7 (91.8)
Safety/Truth	1.3 ±0.5 (67.2)	5.4 ±2.0 (59.4)	4.2 ±0.3 (65.3)
Science/Tech	1.5 ±0.1 (85.0)	6.2 ±0.2 (66.2)	7.9 ±0.4 (84.0)
Total	1.4 ±0.1 (79.2)	6.1 ±0.4 (55.0)	5.4 ±0.1 (77.4)

Table 22: Trained from Instruct - High Data Scenario: Backward Transfer (Part 2 of 3)

Category	OpenCodeReasoning-Nemotron-1.1-7B	OpenThinker3-7B	OpenThinker-7B
Common Sense	5.2 ±0.5 (54.3)	6.7 ±0.0 (73.2)	5.2 ±1.7 (53.5)
Culture	4.7 ±2.3 (19.2)	2.4 ±1.9 (17.8)	6.5 ±1.4 (33.6)
Logic	14.0 ±1.1 (58.2)	14.7 ±2.7 (71.9)	13.5 ±0.3 (67.4)
Knowledge/QA	5.3 ±0.2 (42.6)	14.2 ±8.6 (61.0)	7.7 ±0.9 (43.0)
Language	4.8 ±0.9 (23.3)	5.6 ±1.8 (32.8)	8.2 ±1.8 (37.6)
Liberal Arts	4.3 ±0.3 (52.8)	5.6 ±0.4 (67.6)	5.6 ±0.7 (59.2)
Math	10.1 ±0.7 (56.4)	13.9 ±1.6 (81.5)	11.0 ±2.2 (71.3)
Safety/Truth	8.7 ±1.1 (42.5)	6.8 ±0.8 (48.4)	7.5 ±0.7 (50.7)
Science/Tech	5.4 ±0.4 (49.3)	9.6 ±0.7 (69.4)	7.4 ±0.1 (58.7)
Total	6.7 ±0.4 (46.8)	8.5 ±1.5 (59.0)	7.7 ±0.4 (54.4)

Table 23: Trained from Instruct - High Data Scenario: Backward Transfer (Part 3 of 3)

Category	OpenMath2-Llama3.1-8B
Common Sense	9.0 ±2.8 (34.1)
Culture	1.7 ±1.2 (6.5)
Logic	14.0 ±0.4 (18.3)
Knowledge/QA	6.4 ±0.7 (25.3)
Language	12.4 ±2.3 (19.1)
Liberal Arts	5.2 ±1.3 (25.5)
Math	4.7 ±0.8 (15.0)
Safety/Truth	5.2 ±0.4 (15.9)
Science/Tech	3.4 ±0.7 (17.6)
Total	7.9 ±0.4 (21.2)

E.2 INSTRUCTION TUNING (TASK)

E.3 DOMAIN-CONTINUAL PRETRAINING

E.4 REASONING MODELS TRAINED FROM BASE

F REASONING MODELS TRAINED FROM INSTRUCT - HIGH DATA SCENARIO

G REASONING MODELS TRAINED FROM INSTRUCT - LOW DATA SCENARIO

Table 24: Trained from Instruct - Low Data Scenario: Forgetting (Part 1 of 2)

Category	s1.1-14B	LIMO	s1.1-32B
Common Sense	6.5 ±1.5 (76.9)	4.6 ±0.6 (80.1)	6.0 ±1.2 (80.1)
Culture	10.1 ±0.3 (65.0)	3.4 ±1.0 (60.9)	7.4 ±0.6 (60.9)
Logic	4.1 ±0.6 (75.2)	3.1 ±0.4 (80.2)	2.0 ±0.2 (83.3)
Knowledge/QA	2.4 ±0.3 (45.2)	2.1 ±1.1 (47.8)	2.3 ±1.5 (26.1)
Language	6.8 ±0.9 (51.6)	3.6 ±0.5 (55.2)	4.9 ±0.6 (60.1)
Liberal Arts	5.2 ±0.6 (79.1)	2.4 ±0.2 (81.4)	3.0 ±0.3 (87.4)
Math	5.2 ±0.5 (73.5)	1.1 ±0.1 (80.2)	3.7 ±0.4 (80.8)
Safety/Truth	5.6 ±0.6 (59.8)	3.0 ±0.5 (64.7)	3.6 ±0.3 (62.1)
Science/Tech	4.7 ±0.4 (74.8)	2.1 ±0.1 (75.9)	3.0 ±0.3 (79.3)
Total	5.3 ±0.3 (69.5)	2.7 ±0.1 (72.2)	3.8 ±0.2 (71.6)

Table 25: Trained from Instruct - Low Data Scenario: Forgetting (Part 2 of 2)

Category	s1.1-7B
Common Sense	9.3 ±0.1 (73.9)
Culture	11.3 ±5.4 (42.7)
Logic	8.8 ±0.1 (55.3)
Knowledge/QA	4.5 ±2.1 (57.8)
Language	7.9 ±0.8 (35.4)
Liberal Arts	7.5 ±0.4 (73.1)
Math	6.3 ±0.6 (66.8)
Safety/Truth	6.6 ±0.7 (53.5)
Science/Tech	7.4 ±0.1 (67.4)
Total	7.7 ±0.8 (61.6)

Table 26: Trained from Instruct - Low Data Scenario: Backward Transfer (Part 1 of 2)

Category	s1.1-14B	LIMO	s1.1-32B
Common Sense	5.1 ±0.1 (75.1)	4.4 ±0.5 (79.8)	4.5 ±0.7 (78.1)
Culture	3.0 ±0.9 (54.2)	7.8 ±0.5 (66.1)	5.1 ±0.7 (56.7)
Logic	9.3 ±0.5 (81.1)	6.0 ±0.0 (84.6)	5.5 ±0.1 (87.7)
Knowledge/QA	14.7 ±1.5 (59.4)	15.3 ±1.0 (62.4)	22.0 ±3.0 (48.3)
Language	7.5 ±1.3 (53.7)	5.5 ±0.5 (58.5)	6.6 ±0.2 (63.4)
Liberal Arts	5.1 ±0.3 (78.9)	4.5 ±0.6 (84.2)	3.8 ±0.2 (88.6)
Math	11.6 ±0.4 (81.9)	9.3 ±0.8 (90.7)	9.2 ±0.7 (88.2)
Safety/Truth	7.4 ±0.5 (62.2)	4.9 ±0.4 (67.1)	3.9 ±0.2 (62.5)
Science/Tech	7.7 ±0.2 (78.9)	7.4 ±0.2 (83.2)	7.5 ±0.4 (85.3)
Total	7.4 ±0.3 (71.9)	6.7 ±0.3 (77.2)	7.0 ±0.3 (75.4)

Table 27: Trained from Instruct - Low Data Scenario: Backward Transfer (Part 2 of 2)

Category	s1.1-7B
Common Sense	6.2 ±0.5 (69.8)
Culture	4.5 ±1.3 (29.1)
Logic	14.3 ±0.4 (62.4)
Knowledge/QA	5.4 ±1.3 (58.9)
Language	8.9 ±0.8 (37.2)
Liberal Arts	6.3 ±0.9 (71.5)
Math	11.9 ±0.6 (74.5)
Safety/Truth	8.8 ±0.2 (56.4)
Science/Tech	8.4 ±0.2 (68.8)
Total	8.0 ±0.2 (61.5)

H TRAINED FROM INSTRUCT - SFT

Table 28: Trained from Instruct - SFT: Forgetting (Part 1 of 2)

Category	s1.1-14B	OpenThinker2-32B	s1.1-32B
Common Sense	6.5 ±1.5 (76.9)	3.9 ±0.2 (84.6)	6.0 ±1.2 (80.1)
Culture	10.1 ±0.3 (65.0)	7.3 ±1.9 (61.4)	7.4 ±0.6 (60.9)
Logic	4.1 ±0.6 (75.2)	1.0 ±0.1 (81.2)	2.0 ±0.2 (83.3)
Knowledge/QA	2.4 ±0.3 (45.2)	2.9 ±0.5 (60.4)	2.3 ±1.5 (26.1)
Language	6.8 ±0.9 (51.6)	4.2 ±0.3 (57.1)	4.9 ±0.6 (60.1)
Liberal Arts	5.2 ±0.6 (79.1)	2.4 ±0.1 (80.8)	3.0 ±0.3 (87.4)
Math	5.2 ±0.5 (73.5)	0.9 ±0.3 (80.2)	3.7 ±0.4 (80.8)
Safety/Truth	5.6 ±0.6 (59.8)	3.4 ±0.1 (64.3)	3.6 ±0.3 (62.1)
Science/Tech	4.7 ±0.4 (74.8)	2.0 ±0.2 (76.0)	3.0 ±0.3 (79.3)
Total	5.3 ±0.3 (69.5)	2.9 ±0.3 (74.2)	3.8 ±0.2 (71.6)

Table 29: Trained from Instruct - SFT: Forgetting (Part 2 of 2)

Category	OpenThinker3-7B	OpenThinker-7B	s1.1-7B
Common Sense	9.4 ±0.5 (76.8)	21.2 ±4.0 (74.7)	9.3 ±0.1 (73.9)
Culture	15.8 ±3.9 (43.8)	14.9 ±4.4 (46.4)	11.3 ±5.4 (42.7)
Logic	5.6 ±0.8 (59.1)	7.5 ±0.3 (60.7)	8.8 ±0.1 (55.3)
Knowledge/QA	7.3 ±3.1 (54.8)	13.1 ±1.0 (51.2)	4.5 ±2.1 (57.8)
Language	10.7 ±4.3 (38.7)	10.1 ±1.2 (40.1)	7.9 ±0.8 (35.4)
Liberal Arts	9.8 ±0.4 (73.3)	16.1 ±1.1 (73.1)	7.5 ±0.4 (73.1)
Math	3.9 ±0.4 (68.4)	8.4 ±1.1 (67.6)	6.3 ±0.6 (66.8)
Safety/Truth	8.1 ±0.6 (50.1)	9.6 ±0.8 (53.6)	6.6 ±0.7 (53.5)
Science/Tech	6.5 ±0.3 (65.2)	12.2 ±0.4 (65.2)	7.4 ±0.1 (67.4)
Total	8.3 ±0.4 (59.8)	13.2 ±1.6 (62.1)	7.7 ±0.8 (61.6)

Table 30: Trained from Instruct - SFT: Backward Transfer (Part 1 of 2)

Category	s1.1-14B	OpenThinker2-32B	s1.1-32B
Common Sense	5.1 ±0.1 (75.1)	4.2 ±0.3 (85.0)	4.5 ±0.7 (78.1)
Culture	3.0 ±0.9 (54.2)	2.9 ±0.4 (54.0)	5.1 ±0.7 (56.7)
Logic	9.3 ±0.5 (81.1)	6.2 ±0.1 (88.4)	5.5 ±0.1 (87.7)
Knowledge/QA	14.7 ±1.5 (59.4)	4.6 ±0.7 (62.5)	22.0 ±3.0 (48.3)
Language	7.5 ±1.3 (53.7)	6.6 ±0.5 (61.3)	6.6 ±0.2 (63.4)
Liberal Arts	5.1 ±0.3 (78.9)	5.6 ±0.3 (85.2)	3.8 ±0.2 (88.6)
Math	11.6 ±0.4 (81.9)	10.0 ±0.7 (91.8)	9.2 ±0.7 (88.2)
Safety/Truth	7.4 ±0.5 (62.2)	4.2 ±0.3 (65.3)	3.9 ±0.2 (62.5)
Science/Tech	7.7 ±0.2 (78.9)	7.9 ±0.4 (84.0)	7.5 ±0.4 (85.3)
Total	7.4 ±0.3 (71.9)	5.4 ±0.1 (77.4)	7.0 ±0.3 (75.4)

Table 31: Trained from Instruct - SFT: Backward Transfer (Part 2 of 2)

Category	OpenThinker3-7B	OpenThinker-7B	s1.1-7B
Common Sense	6.7 ±0.0 (73.2)	5.2 ±1.7 (53.5)	6.2 ±0.5 (69.8)
Culture	2.4 ±1.9 (17.8)	6.5 ±1.4 (33.6)	4.5 ±1.3 (29.1)
Logic	14.7 ±2.7 (71.9)	13.5 ±0.3 (67.4)	14.3 ±0.4 (62.4)
Knowledge/QA	14.2 ±8.6 (61.0)	7.7 ±0.9 (43.0)	5.4 ±1.3 (58.9)
Language	5.6 ±1.8 (32.8)	8.2 ±1.8 (37.6)	8.9 ±0.8 (37.2)
Liberal Arts	5.6 ±0.4 (67.6)	5.6 ±0.7 (59.2)	6.3 ±0.9 (71.5)
Math	13.9 ±1.6 (81.5)	11.0 ±2.2 (71.3)	11.9 ±0.6 (74.5)
Safety/Truth	6.8 ±0.8 (48.4)	7.5 ±0.7 (50.7)	8.8 ±0.2 (56.4)
Science/Tech	9.6 ±0.7 (69.4)	7.4 ±0.1 (58.7)	8.4 ±0.2 (68.8)
Total	8.5 ±1.5 (59.0)	7.7 ±0.4 (54.4)	8.0 ±0.2 (61.5)

DISCLAIMER FOR USE OF LLMS

We primarily used LLMs in coding co-pilot applications to facilitate experimentation and help with plotting code for result presentation. LLMs were also used as writing tools to assist in refining the paper. However, the final version was carefully reviewed and finalized by the authors. No LLMs were used in ideation and experimental design.