CAPABILITY BUDGETING: ACQUIRING MATHEMATICAL REASONING WHILE RETAINING CAPABILITIES IN MULTILINGUAL CONTINUAL PRETRAINING

Anonymous authors
Paper under double-blind review

ABSTRACT

Multilingual continual pretraining must balance acquiring target skills with retaining useful capabilities under a limited training budget. We introduce Capability Budgeting, a framework that connects role-aware data allocation to mathematical acquisition under explicit per-capability retention requirements. To acquire Thai, Bengali, and Swahili mathematics, training tokens are allocated across language-serving, general-English, and mathematical-English pools; trained candidates are assessed by their target gains and allowed losses on protected tasks. Across Qwen3-4B-Base and Qwen3-8B-Base, three allocations and three training seeds at 16.384M tokens per run yield 18 models that all improve target-language average mathematics by 2.27–4.27 percentage points, full-test English GSM8K, and Swahili Global-MMLU. Every matched-seed comparison contains a candidate combining mathematical gains with all nine reference retention requirements; full-test HumanEval also improves in all six code-evaluated 4B models. A controlled exchange of mathematical for general English data at a fixed combined share produces distinct capability profiles, with repeated changes in target-language mathematics at 4B and GPQA at 8B. Retention requirements change the highest-mathematics choice in every 4B seed comparison. These results establish fixed-budget mathematical acquisition with budgeted capability retention and show why the role of training data and the requirements on individual capabilities jointly matter for choosing an adaptation recipe.

1 INTRODUCTION

Continual pretraining (CPT) adapts language models to new domains and languages without repeating pretraining from scratch (Gururangan et al., 2020; Ibrahim et al., 2024). Under a limited training budget, useful multilingual adaptation must acquire the intended skills while retaining capabilities that its users already rely on. This makes data allocation a joint acquisition and retention problem: the training budget must support target gains while keeping losses on other relevant tasks within acceptable limits.

We introduce *Capability Budgeting* to address this joint objective (Figure 1). Data roles organize the allocation of training tokens, and per-capability loss budgets express the retention requirements. Together, they connect the construction of adaptation recipes to the choice among their trained outcomes. In our study, language-serving, general-English, and mathematical-English data support the acquisition of Thai, Bengali, and Swahili mathematical reasoning. English mathematics, multilingual knowledge, and broader reasoning supply individual retention requirements. The framework identifies candidates that acquire the target skill while satisfying these requirements, then compares their acquisition.

Our controlled study compares three allocations on Qwen3-4B-Base and Qwen3-8B-Base, with 16.384M training tokens and three optimization seeds per allocation. Uniform provides an equal-share reference across six fixed pools. Replay changes the overall allocation, and English-role exchanges mathematical for general English text while holding the four language-serving shares

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

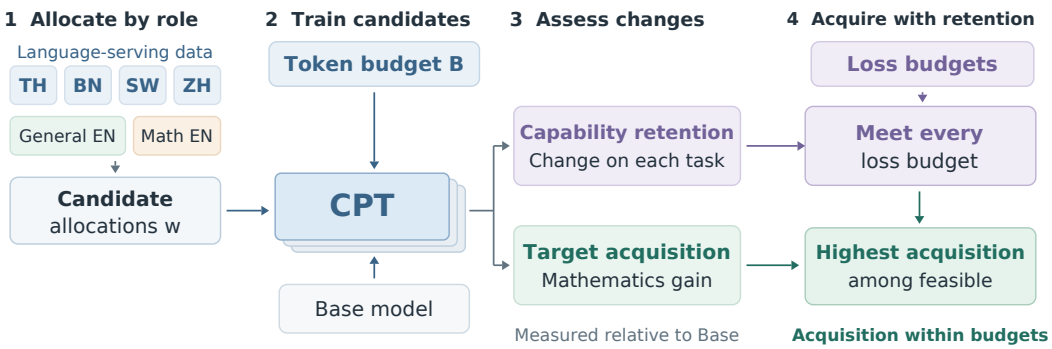


Figure 1: **Capability Budgeting workflow.** Language-serving, general-English, and mathematical-English pools receive shares of a fixed token budget. CPT candidates from a shared Base are assessed for target acquisition and changes in protected capabilities. Per-capability loss budgets define the feasible set among evaluated candidates and Base; selection maximizes acquisition within this set, retaining ties (Equation 2).

and combined dedicated English-pool share fixed. These comparisons establish the acquisition and retention outcomes of CPT and test how a change in data roles alters the capability profile.

The results demonstrate useful acquisition and retention combinations at both model sizes. All 18 trained models improve target-language average mathematics, full-test English GSM8K, and Swahili Global-MMLU. Uniform improves every target language and GSM8K in all six runs. Every matched-seed comparison contains a CPT candidate that improves target mathematics while meeting all nine reference retention budgets. The six code-evaluated 4B models also improve full-test HumanEval by 6.10–7.93 percentage points, extending the benefits to programming.

Data roles and capability requirements explain which outcomes to choose. The English-role exchange raises Thai mathematics and lowers Swahili mathematics in every 4B seed, while raising GPQA scores in every 8B seed. Thus, equal combined English shares can yield different capability profiles. Retention requirements change the highest-mathematics choice in every 4B seed comparison; at 8B, the highest-mathematics candidates already meet them. These results give data allocation and capability requirements complementary roles in obtaining a useful adaptation.

Our contributions are threefold:

- **A role-aware acquisition and retention framework.** Capability Budgeting connects fixed-budget data allocation to explicit per-capability loss requirements, making target acquisition and capability retention a joint criterion for comparing trained adaptation recipes.
- **A controlled finding on data roles.** Exchanging general English and mathematics at a fixed combined share produces repeated, task-specific changes. The capability profile depends on what the English data support, beyond their total share.
- **Cross-size evidence and consequential capability budgets.** The matched 4B/8B study establishes mathematical gains with qualifying candidates in every seed comparison. We show when retention requirements change the preferred candidate and locate each candidate's budget boundary.

2 Capability Budgeting for Acquisition and Retention

The objective is to acquire target-language mathematics within a fixed training budget while satisfying individual capability requirements. Role-aware allocation assigns training tokens to pools serving different purposes; retention budgets express the allowed task losses. We assess the resulting trained candidates jointly on acquisition and retention. Let $w \in \Delta^5$ allocate tokens over six source pools, with $\Delta^5 = \{w \in \mathbb{R}^6_{\geq 0} : \sum_i w_i = 1\}$. For a fixed Base model θ_0, token budget B, and training recipe $\mathcal{T}$, a run $p = (w, \xi)$ with optimization seed ξ produces $\theta_B(p) = \mathrm{CPT}(\theta_0, w, B; \mathcal{T}, \xi)$.

For target languages $\mathcal{L} = \{\mathrm{TH}, \mathrm{BN}, \mathrm{SW}\}$, define

$$A(p) = \frac{1}{|\mathcal{L}|} \sum_{\ell \in \mathcal{L}} [s_\ell(\theta_B(p)) - s_\ell(\theta_0)], \qquad W(p) = \min_{\ell \in \mathcal{L}} s_\ell(\theta_B(p)). \tag{1}$$

A measures acquisition relative to Base; W retains the weakest target. For protected capability j, let $R_j(p) = r_j(\theta_B(p)) - r_j(\theta_0)$ and let $\epsilon_j \geq 0$ be the allowed drop. Scores are measured on the fixed evaluation sets. Given an evaluated candidate set $\mathcal{C}$, including the Base anchor, the budget-feasible set and acquisition-based selection are

$$\mathcal{F}(\epsilon) = \{p \in \mathcal{C} : R_j(p) \geq -\epsilon_j \ \forall j\}, \qquad \mathcal{P}^*(\epsilon) = \arg\max_{p \in \mathcal{F}(\epsilon)} A(p). \tag{2}$$

Ties are retained. Base has $A = R_j = 0$ and meets every nonnegative budget. Throughout, meeting a budget means satisfying these measured-score constraints. Paired-item intervals describe fixed-checkpoint uncertainty, and training-seed variation describes repeated training; neither is treated as a guarantee for a new run. Both are reported separately in Appendix B. The experiment compares candidates within each model size and matched training seed, then summarizes acquisition and budget satisfaction across the three seeds of each allocation.

Here the nine protected outcomes are English MGSM, GSM8K, four Global-MMLU languages, C-Eval, GPQA, and CMMLU mathematics. Code supplies additional capability outcomes. For each evaluated model, these budgets identify whether its mathematical gains meet the retention requirements. Equation 2 then compares acquisition among the qualifying models. Individual task gains, losses, and weakest-target scores remain available alongside this decision. Section 4.5 varies the requirements continuously; Appendix D gives auxiliary Pareto comparisons.

3 Experimental Setup

Models and allocations. We study Qwen3-4B-Base and Qwen3-8B-Base (Yang et al., 2025; Qwen Team, 2025). Each CPT run consumes 16,384,000 tokens over 250 steps. Six pools—TH, BN, SW, ZH, EN-general, and EN-math—supply the data. The first four serve their named languages through synthetic mathematical reasoning, monolingual text, parallel translation, and bilingual Wikipedia contexts. Their English and bilingual material supports the named language. Wikipedia contexts adapt the related-bilingual-text construction of CrossIC-PT (Wu et al., 2025). Pool weights specify training-token allocations.

Uniform assigns equal shares. Replay increases the combined dedicated English-pool share from one third to 37%, split equally between general text and mathematics, while retaining an 18% Swahili allocation. English-role holds these four language-serving shares fixed and transfers 6.5 points from EN-math to EN-general. These comparisons test allocation balance and English data roles (Table 1).

Table 1: Pool allocations (%) shared by the 4B and 8B experiments. Uniform assigns exactly one sixth to each pool. English-role keeps the two dedicated English pools at a combined 37%. Each listed seed is run at both model sizes.

Allocation	TH	BN	SW	ZH	EN-general	EN-math	Seeds
Uniform	16.67	16.67	16.67	16.67	16.67	16.67	42, 43, 44
Replay	16	15	18	14	18.5	18.5	42, 43, 44
English-role	16	15	18	14	25	12	42, 43, 44

Controlled comparison. Each size uses its own fixed Base; all primary runs share the processed pools, tokenizer, token budget, optimizer, and schedule. Corresponding allocations reuse the same packed tokens across sizes. Repeated optimization seeds reuse the same pack and order. Every allocation is evaluated with three training seeds at both sizes, yielding 18 CPT checkpoints with complete mathematical and broad-task scores. Code covers Base and two seeds per allocation at each size. Cross-size comparisons use the same token budget; training FLOPs depend on model size. Data construction and implementation appear in Appendix A.

Outcomes. MGSM (Shi et al., 2023) uses 250 items per TH/BN/SW/EN language. Target-language average accuracy is the unweighted TH/BN/SW mean; worst-target accuracy is their minimum. We calculate these for each checkpoint before summarizing seeds. English GSM8K (Cobbe et al., 2021) uses the full 1,319-question test set and provides the principal English mathematical outcome; MGSM-English retains the English member of the multilingual comparison. Global-MMLU EN/BN/SW/ZH (Singh et al., 2025) and C-Eval (Huang et al., 2023) use 250 items each; GPQA Diamond (Rein et al., 2024) uses 198 and CMMLU mathematics (Li et al., 2024) 15. HumanEval (Chen et al., 2021) covers all 164 problems, while MBPP (Austin et al., 2021) uses a fixed 100-item subset. Both report deterministic pass@1. Seed means and sample SDs summarize repeated training; paired-item bootstrap intervals summarize fixed-checkpoint score differences (Appendix B).

Reference capability budgets. We use allowed absolute drops of 4 percentage points for MGSM-English, GSM8K, and each Global-MMLU task, and 5 for C-Eval, GPQA, and CMMLU. These researcher-set reference budgets were recorded in the early study protocol and retained across models. They express common absolute-score requirements, while sensitivity analysis shows how choices change as the requirements tighten. Raw task scores are reported throughout.

4 Acquiring Mathematics while Retaining Capabilities

We establish mathematical acquisition and joint retention, examine how data roles change the capability profile, and then consider broader benefits and the effect of tightening the retention budgets.

4.1 Mathematical gains extend across languages and English GSM8K

All 18 CPT checkpoints improve target-language average accuracy over their own Base, with gains of 2.27–4.27 percentage points (Table 2). At 4B, Uniform, Replay, and English-role raise the 53.87% Base to three-seed means of 57.29%, 57.51%, and 57.87%. At 8B, the corresponding means are 67.29%, 66.71%, and 66.67%, compared with 63.60% for Base. Uniform improves TH, BN, and SW in every seed at both sizes, with target-average sample SDs of 0.15 and 0.54 percentage points.

Table 2: MGSM accuracy (%). Target avg. averages Thai, Bengali, and Swahili; Δ is its gain over the same-size Base in percentage points. Allocation rows report seed means, with sample SD for target average. Every language uses the same 250 items. n is the number of training seeds.

Size	Allocation	n	TH	BN	SW	Target avg.	Δ	EN
4B	Base	–	70.80	65.60	25.20	53.87	–	82.00
4B	Uniform	3	77.47	67.20	27.20	57.29 $\pm$ 0.15	+3.42	78.93
4B	Replay	3	77.20	67.73	27.60	57.51 $\pm$ 0.67	+3.64	79.47
4B	English-role	3	79.33	68.13	26.13	57.87 $\pm$ 0.35	+4.00	78.00
8B	Base	–	77.20	71.20	42.40	63.60	–	88.00
8B	Uniform	3	81.20	73.47	47.20	67.29 $\pm$ 0.54	+3.69	84.67
8B	Replay	3	81.60	72.67	45.87	66.71 $\pm$ 0.80	+3.11	85.47
8B	English-role	3	81.73	73.07	45.20	66.67 $\pm$ 0.35	+3.07	85.33

English mathematical gains accompany target-language acquisition. All 18 checkpoints also improve full-test GSM8K (Table 3). At 4B, Uniform, Replay, and English-role reach 90.73%, 89.84%, and 89.51%, gaining 3.92, 3.03, and 2.70 points over the 86.81% Base. At 8B, their means are 91.53%, 91.46%, and 91.15%, gaining 1.84, 1.77, and 1.47 points over 89.69%. The Base-relative paired 95% intervals have positive lower bounds in all nine 4B runs and six of nine 8B runs. Uniform therefore improves all three target-language MGSM scores and English GSM8K in all six runs.

The allocations achieve similar target averages through different language profiles. Replay-minus-Uniform average differences are +0.93, +0.13, and -0.40 points at 4B, versus +0.80, -0.80, and -1.73 at 8B. The two allocations therefore have no consistent acquisition ordering across seeds and sizes. Their individual capability profiles determine which is preferable under a given set of retention requirements (Section 4.2); pairwise intervals appear in Appendix B.

4.2 MATHEMATICAL ACQUISITION WITHIN CAPABILITY BUDGETS

Mathematical acquisition and capability retention are jointly achievable at both sizes: every matched-seed group contains a trained model with positive target acquisition that meets all nine reference budgets. Table 3 reports the protected knowledge and reasoning scores alongside English GSM8K.

Table 3: Knowledge and reasoning accuracy (%), averaged over three training seeds for every allocation. GSM8K uses all 1,319 test items. GMMLU denotes Global-MMLU; its subsets and C-Eval each use 250 items, GPQA uses 198, and CMMLU math uses 15. Complete seed scores and paired intervals are in Appendix B.

Size	Allocation	n	GSM8K	GMMLU-EN	GMMLU-BN	GMMLU-SW
4B	Base	–	86.81	68.40	49.60	32.00
4B	Uniform	3	90.73	68.80	46.53	34.93
4B	Replay	3	89.84	69.20	45.33	34.53
4B	English-role	3	89.51	69.20	46.67	35.07
8B	Base	–	89.69	74.40	53.20	40.00
8B	Uniform	3	91.53	74.67	51.73	40.93
8B	Replay	3	91.46	75.60	50.67	41.60
8B	English-role	3	91.15	74.67	50.80	42.00

Size	Allocation	n	GMMLU-ZH	C-Eval	GPQA	CMMLU math
4B	Base	–	70.00	76.80	34.34	40.00
4B	Uniform	3	69.73	77.73	33.16	46.67
4B	Replay	3	69.20	76.67	32.83	46.67
4B	English-role	3	70.13	77.07	34.18	53.33
8B	Base	–	70.80	79.60	37.37	66.67
8B	Uniform	3	68.67	79.73	37.71	71.11
8B	Replay	3	68.53	80.13	37.71	68.89
8B	English-role	3	68.93	80.00	38.89	71.11

These requirements also distinguish the preferred candidates. Table 4 compares selection by mathematics alone with selection under the reference budgets; tighter budgets are examined in Section 4.5. At 4B, the target-only winner is excluded in every seed: Replay and English-role tie at seed 42, Replay wins at seed 43, and English-role wins at seed 44. The budget-qualified choices are Uniform, Uniform/English-role, and Replay, respectively. At 8B, the corresponding target-only winners already meet the reference budgets.

Table 4: Highest target-average candidate before and after imposing retention budgets, within model size and matched seed. Entries give the selected allocation(s) and target accuracy (%). U: Uniform; R: Replay; E: English-role. All exact ties are retained. Base is an eligible no-training option.

Size	Seed	Target only	$\lambda = 1$	$\lambda = .9$	$\lambda = .75$
4B	42	R/E (58.13)	U (57.20)	U (57.20)	Base (53.87)
4B	43	R (57.60)	U/E (57.47)	U/E (57.47)	U (57.47)
4B	44	E (58.00)	R (56.80)	R (56.80)	Base (53.87)
8B	42	R (67.47)	R (67.47)	R (67.47)	R (67.47)
8B	43	U (67.60)	U (67.60)	U (67.60)	Base (63.60)
8B	44	U (67.60)	U (67.60)	U (67.60)	U (67.60)

A concrete case shows why several protected tasks matter. At 4B seed 42, Replay and English-role both reach 58.13% target accuracy. Replay meets the English-MGSM budget but loses 4.4 points on Bengali Global-MMLU, exceeding its 4-point budget. English-role meets that Bengali requirement but loses 4.8 points on English MGSM. Uniform, at 57.20%, satisfies both and the remaining constraints. A single English-retention score would miss the constraint that excludes Replay.

4.3 English data roles redistribute gains and costs

The Replay–English-role comparison isolates a joint change of data roles: the four language-serving pool shares remain fixed, and 6.5 percentage points move from EN-math to EN-general within a 37% combined dedicated English-pool share. This holds total dedicated English exposure constant while changing its role. Relative to matched Replay runs, all three 4B contrasts improve Thai mathematics (+3.2/+2.4/+0.8 points) and decrease Swahili mathematics ($-2.0/-2.0/-0.4$). At 8B, GPQA improves in all three contrasts (+2.53/+0.51/+0.51 points), averaging +1.18 points (Table 22). The fixed-share exchange produces repeated, task-specific changes; the combined dedicated English-pool share alone does not describe the resulting capability profile.

The exchange also changes the retention profile. English MGSM falls in all three 4B contrasts by 1.2, 2.4, and 0.8 points; at 8B it changes by -2.0, -0.4, and +2.0, averaging -0.13 points. Bengali Global-MMLU gains average +1.33 points at 4B and Swahili Global-MMLU gains +0.40 at 8B, with positive differences in two of three seeds each. These profiles show why the English data role belongs in the description of an adaptation recipe alongside its total dedicated English share.

4.4 Knowledge and programming benefits extend the capability profile

Swahili knowledge gains recur across the entire matrix: every CPT checkpoint improves Global-MMLU SW by 0.4–3.6 percentage points over its own Base. Together with the mathematics results, all 18 models improve target-language average accuracy, English GSM8K, and Swahili knowledge. Bengali presents a different profile: Uniform improves its mathematics in all six runs, while Bengali Global-MMLU decreases throughout the matrix by 0.8–4.8 points. Mathematical improvement thus cannot serve as a proxy for knowledge retention in the same language. The full task scores make this distinction visible alongside the concurrent Swahili gains.

Full-test HumanEval gains repeat at 4B. The six code-evaluated 4B checkpoints improve HumanEval by 6.10–7.93 points; every paired 95% interval has a positive lower bound. Uniform, Replay, and English-role reach two-seed means of 61.59%, 62.50%, and 61.89%, respectively, compared with 54.88% for Base (Table 5). At 8B, checkpoint changes range from 0 to +2.44 points. MBPP has a different response: its 4B means are 62%, 63%, and 61%, respectively, versus a 63% Base. The full HumanEval evaluation thus extends the repeated 4B benefits to programming.

Table 5: Code pass@1 (%) on all 164 HumanEval problems and a fixed 100-item MBPP subset. Allocation rows show mean ± sample SD over seeds 42/43; Δ is the change from the same-size Base in percentage points. Paired intervals appear in Table 18.

Size	Allocation	HumanEval	MBPP	Δ HumanEval	Δ MBPP
4B	Base	54.88	63.00	–	–
4B	Uniform	61.59 ± 0.86	62.00 ± 1.41	+6.71	-1.00
4B	Replay	62.50 ± 0.43	63.00 ± 1.41	+7.62	+0.00
4B	English-role	61.89 ± 0.43	61.00 ± 0.00	+7.01	-2.00
8B	Base	61.59	65.00	–	–
8B	Uniform	63.11 ± 1.29	65.50 ± 2.12	+1.52	+0.50
8B	Replay	62.80 ± 0.86	66.50 ± 0.71	+1.22	+1.50
8B	English-role	61.89 ± 0.43	66.50 ± 2.12	+0.30	+1.50

Chinese mathematics provides an additional diagnostic. On the 15-item CMMLU mathematics subset, all three 4B English-role seeds improve from 40.0% to 53.3%, adding two correct answers. At 8B, two English-role seeds improve from 66.7% to 73.3% and one matches Base.

4.5 Budget boundaries characterize the available choices

For the positive reference margins used here, scale the allowed losses to $\lambda\epsilon_j$ and define the minimum required scale

$$\rho(p) = \max_j \frac{\max\{0, -R_j(p)\}}{\epsilon_j}. \tag{3}$$

A checkpoint meets the scaled budgets exactly when $\rho(p) \leq \lambda$. The reference is $\lambda = 1$; $\lambda = .90$ and .75 illustrate 10% and 25% tighter allowed drops. Each ρ identifies the complete eligibility boundary for that candidate (Figure 2).

Table 6: Retention-budget sensitivity over all nine protected tasks. ρ is the minimum common margin scale required by a checkpoint; smaller is better. Counts report seeds satisfying every retention constraint at each scale λ. The original margins correspond to $\lambda = 1$. Base has $\rho = 0$.

		ρ by seed			Feasible seeds at λ		
Size	Allocation	42	43	44	1.00	.90	.75
4B	Uniform	0.80	0.60	1.20	2/3	2/3	1/3
4B	Replay	1.10	1.20	0.90	1/3	1/3	0/3
4B	English-role	1.20	0.90	1.20	1/3	1/3	0/3
8B	Uniform	1.00	0.90	0.60	3/3	2/3	1/3
8B	Replay	0.60	0.90	0.80	3/3	3/3	1/3
8B	English-role	0.70	1.00	0.60	3/3	2/3	2/3

At the reference budgets, 4B admits 2/3 Uniform, 1/3 Replay, and 1/3 English-role runs; all nine 8B runs qualify. At $\lambda = .90$, Replay qualifies in all three 8B seeds, and Uniform and English-role in two each. At $\lambda = .75$, English-role qualifies in two 8B seeds, and the other allocations in one each. The boundaries expose how each allocation responds to stronger retention requirements: comparable acquisition means can correspond to different sets of acceptable models.

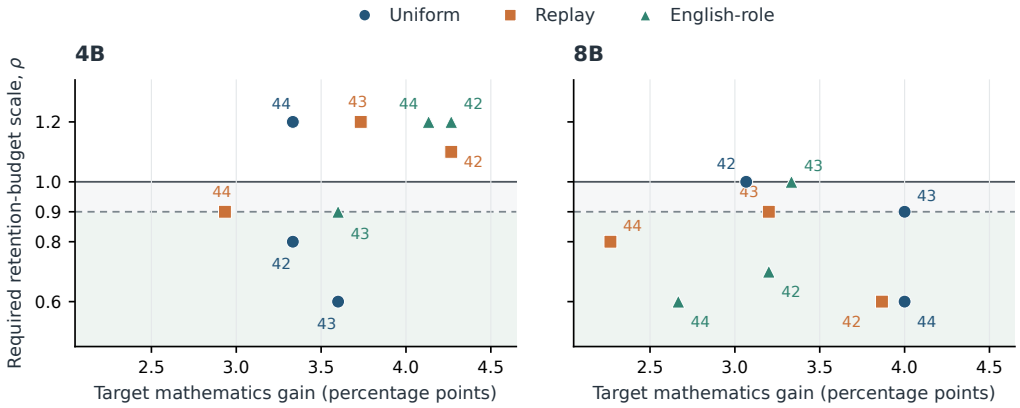


Figure 2: Mathematical acquisition and the minimum retention-budget scale required by each checkpoint, relative to its same-size Base. Points on or below the solid line meet the reference budgets; points on or below the dashed line meet budgets tightened by 10%. Symbols identify allocations and numbers identify training seeds.

5 Related Work

Domain and lifelong pretraining address corpus adaptation (Gururangan et al., 2020; Jin et al., 2022); replay and scheduling address retention (Ibrahim et al., 2024; Abbes et al., 2026; De Lange et al., 2022; Scialom et al., 2022). BLOOM+1 and Glot500 expand multilingual coverage (Yong et al., 2023; Imani et al., 2023), while ATLAS studies transfer across model sizes and languages (Longpre et al., 2026). We study mathematical acquisition together with explicit capability retention requirements at two model sizes under matched token budgets.

DoReMi learns domain weights from proxy-model excess loss (Xie et al., 2023); RegMix uses regression (Liu et al., 2025); Data Mixing Laws predicts mixture performance and studies CPT forgetting (Ye et al., 2025). IDEAL adjusts multi-domain SFT quantities for multiple capabilities (Ming et al., 2026). These methods estimate or optimize data mixtures. Capability Budgeting connects

role-aware allocations over fixed pools to mathematical acquisition under explicit retention requirements. Per-capability requirements connect the controlled comparisons to the choice among trained candidates.

CrossIC-PT constructs semantically related bilingual contexts and evaluates multilingual reasoning, knowledge, and English retention (Wu et al., 2025). We adapt its Wikipedia-context construction within our language-serving pools. MixCPT compares monolingual, bilingual, and code-augmented CPT, identifying task-dependent effects on classification, translation, and language modeling (Li et al., 2025). Our study jointly adapts several target languages under a fixed 16.384M-token budget, establishing repeated mathematical gains and identifying candidates that satisfy individual retention requirements. A matched-seed intervention exchanges general-text and mathematical data while holding the combined dedicated English-pool share fixed, connecting data roles to the resulting capability profile.

6 Discussion and Conclusion

Implications for data allocation. Fixed-budget multilingual adaptation can acquire mathematical skills while meeting explicit requirements on existing capabilities. Uniform provides a strong equal-share starting point, with gains in every target language and seed. The controlled English-role exchange shows that the content role, alongside the total share, changes which capabilities benefit. Retention requirements then determine which of these outcomes are useful. The Bengali results illustrate why these requirements should be task-specific: mathematical improvement and knowledge loss coexist within the same language. Role-aware allocation and individual capability budgets therefore address complementary parts of the adaptation problem.

Scope and interpretation. The study covers two sizes of one model family, three allocations, and one token budget. The controlled role exchange varies pool shares within a common recipe; the components inside each composite language pool are not separately ablated. Budget decisions compare evaluated candidates under stated score requirements, with reference margins treated as research choices and their sensitivity reported in Section 4.5. Statistical coverage and the multilingual resampling limitation are detailed in Appendix B; data construction and availability appear in Appendix A. Testing other model families, training budgets, and data-role exchanges would establish how broadly these acquisition and retention patterns hold.

Conclusion. Capability Budgeting makes mathematical acquisition and capability retention a joint objective for fixed-budget multilingual adaptation. Across 18 CPT models, the allocations improve target-language average mathematics, full-test English GSM8K, and Swahili knowledge; all six code-evaluated 4B models also improve full-test HumanEval. At both sizes, every seed comparison contains a trained candidate combining mathematical gains with all nine reference retention budgets. Holding the combined English share fixed reveals that changing data roles changes the capability profile, while the retention requirements determine which candidate to choose. The resulting framework connects what the training data are intended to support with what an adapted model must acquire and retain.

AI use statement

Generative AI was used to produce synthetic mathematical reasoning traces included in the training data. AI assistance also contributed to experimental orchestration, literature discovery, refinement of the conceptual framing and interpretation of results, manuscript drafting and translation, figure and validation-code preparation, and reference formatting. Automated manuscript checks verify tables and derived quantities against the recorded experiment reports. The authors are responsible for the final claims, citations, and artifacts.

Reproducibility statement

Section 3 and Appendix A describe the training, evaluation, and data sources. Individual seeds and available paired-item intervals appear in Appendix B; study coverage and budget comparisons appear in Appendices C–D. Data availability is stated in Appendix A.

REFERENCES

Istabrak Abbes, Gopeshh Subbaraj, Matthew Riemer, Nizar Islah, Tsuguchika Tabaru, Hiroaki Kingetsu, Sarath Chandar, and Irina Rish. Revisiting replay and gradient alignment for continual pre-training of large language models. In *Proceedings of the 4th Conference on Lifelong Learning Agents*, volume 330 of *Proceedings of Machine Learning Research*, pp. 465–486. PMLR, 2026. URL https://proceedings.mlr.press/v330/abbes26a.html.

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

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

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

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. URL https://doi.org/10.1109/TPAMI.2021.3057446.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 8342–8360, 2020. doi: 10.18653/v1/2020.acl-main.740. URL https://aclanthology.org/2020.acl-main.740/.

Yuzhen Huang, Yuzhuo Bai, Zhihao Zhu, Junlei Zhang, Jinghan Zhang, Tangjun Su, Junteng Liu, Chuancheng Lv, Yikai Zhang, Jiayi Lei, Yao Fu, Maosong Sun, and Junxian He. C-Eval: A multi-level multi-discipline chinese evaluation suite for foundation models. In *Advances in Neural Information Processing Systems*, volume 36, pp. 62991–63010, 2023. doi: 10.52202/075280-2749. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/c6ec1844bec96d6d32ae95ae694e23d8-Abstract-Datasets_and_Benchmarks.html.

Adam Ibrahim, Benjamin Thérien, Kshitij Gupta, Mats L. Richter, Quentin Anthony, Timothée Lesort, Eugene Belilovsky, and Irina Rish. Simple and scalable strategies to continually pre-train large language models. *Transactions on Machine Learning Research*, 2024. URL https://arxiv.org/abs/2403.08763.

Ayyoob Imani, Peiqin Lin, Amir Hossein Kargaran, Silvia Severini, Masoud Jalili Sabet, Nora Kassner, Chunlan Ma, Helmut Schmid, André Martins, François Yvon, and Hinrich Schütze. Glot500: Scaling multilingual corpora and language models to 500 languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1082–1117, 2023. doi: 10.18653/v1/2023.acl-long.61. URL https://aclanthology.org/2023.acl-long.61/.

Xisen Jin, Dejiao Zhang, Henghui Zhu, Wei Xiao, Shang-Wen Li, Xiaokai Wei, Andrew Arnold, and Xiang Ren. Lifelong pretraining: Continually adapting language models to emerging corpora. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 4764–4780, 2022. doi: 10.18653/v1/2022.naacl-main.351. URL https://aclanthology.org/2022.naacl-main.351/.

Haonan Li, Yixuan Zhang, Fajri Koto, Yifei Yang, Hai Zhao, Yeyun Gong, Nan Duan, and Timothy Baldwin. CMMLU: Measuring massive multitask language understanding in chinese. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 11260–11285, 2024. doi: 10.18653/v1/2024.findings-acl.671. URL https://aclanthology.org/2024.findings-acl.671/.

Zihao Li, Shaoxiong Ji, Hengyu Luo, and Jörg Tiedemann. Rethinking multilingual continual pretraining: Data mixing for adapting LLMs across languages and resources. In *Conference on Language Modeling*, 2025. URL https://openreview.net/forum?id=mpTIzK4Zca.

Qian Liu, Xiaosen Zheng, Niklas Muennighoff, Guangtao Zeng, Longxu Dou, Tianyu Pang, Jing Jiang, and Min Lin. RegMix: Data mixture as regression for language model pre-training. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=5BjQOUXq7i.

Shayne Longpre, Sneha Kudugunta, Niklas Muennighoff, I-Hung Hsu, Isaac Caswell, Alex Pentland, Sercan Arik, Chen-Yu Lee, and Sayna Ebrahimi. Atlas: Adaptive transfer scaling laws for multilingual pretraining, finetuning, and decoding the curse of multilinguality. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/35c26a810471039b3427e524565cbef9-Abstract-Conference.html.

Chenlin Ming, Chendi Qu, Qizhi Pei, Zhuoshi Pan, Yu Li, Xiaoming Duan, Lijun Wu, and Conghui He. Ideal: Data equilibrium adaptation for multi-capability language model alignment. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/0ced4d30a9e2d50312991c6cd131a9f5-Abstract-Conference.html.

Qwen Team. Qwen3-8B-Base model card. Hugging Face model repository, 2025. URL https://huggingface.co/Qwen/Qwen3-8B-Base.

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.

Thomas Scialom, Tuhin Chakrabarty, and Smaranda Muresan. Fine-tuned language models are continual learners. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 6107–6122, 2022. doi: 10.18653/v1/2022.emnlp-main.410. URL https://aclanthology.org/2022.emnlp-main.410/.

Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, Dipanjan Das, and Jason Wei. Language models are multilingual chain-of-thought reasoners. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=fR3wGCk-IXp.

Shivalika Singh, Angelika Romanou, Clémentine Fourrier, David Ifeoluwa Adelani, Jian Gang Ngui, Daniel Vila-Suero, Peerat Limkonchotiwat, Kelly Marchisio, Wei Qi Leong, Yosephine Susanto, Raymond Ng, Shayne Longpre, Sebastian Ruder, Wei-Yin Ko, Antoine Bosselut, Alice Oh, Andre Martins, Leshem Choshen, Daphne Ippolito, Enzo Ferrante, Marzieh Fadaee, Beyza Ermis, and Sara Hooker. Global MMLU: Understanding and addressing cultural and linguistic biases in multilingual evaluation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 18761–18799, 2025. doi: 10.18653/v1/2025.acl-long.919. URL https://aclanthology.org/2025.acl-long.919/.

Linjuan Wu, Hao-Ran Wei, Huan Lin, Tianhao Li, Baosong Yang, Fei Huang, and Weiming Lu. Enhancing LLM language adaption through cross-lingual in-context pre-training. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 27152–27166. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.emnlp-main.1380. URL https://aclanthology.org/2025.emnlp-main.1380/.

Sang Michael Xie, Hieu Pham, Xuanyi Dong, Nan Du, Hanxiao Liu, Yifeng Lu, Percy Liang, Quoc V. Le, Tengyu Ma, and Adams Wei Yu. DoReMi: Optimizing data mixtures speeds up language model pretraining. In *Advances in Neural Information Processing Systems*, volume 36, pp. 69798–69818, 2023. doi: 10.52202/075280-3059. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/dcba6be91359358c2355cd920da3fcbd-Abstract-Conference.html.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. URL https://arxiv.org/abs/2505.09388.

Jiasheng Ye, Peiju Liu, Tianxiang Sun, Jun Zhan, Yunhua Zhou, and Xipeng Qiu. Data mixing laws: Optimizing data mixtures by predicting language modeling performance. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=jjCB27TMK3.

Zheng Xin Yong, Hailey Schoelkopf, Niklas Muennighoff, Alham Fikri Aji, David Ifeoluwa Adelani, Khalid Almubarak, M Saiful Bari, Lintang Sutawika, Jungo Kasai, Ahmed Baruwa, Genta Winata, Stella Biderman, Edward Raff, Dragomir Radev, and Vassilina Nikoulina. BLOOM+1: Adding language support to BLOOM for zero-shot prompting. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 11682–11703, 2023. doi: 10.18653/v1/2023.acl-long.653. URL https://aclanthology.org/2023.acl-long.653/.

A Training Data and Implementation

Data construction. Table 7 summarizes pool construction. TH/BN/SW/ZH name the languages served by the pools; each also contains English or bilingual material. All primary comparisons use the same processed pools.

Table 7: Sources and construction of the six allocation pools. Dataset names identify upstream sources in the construction records; they do not imply that all upstream rows enter training. The language pools also contain English material supporting their target language.

Pool	Source components
TH	Synthetic Thai mathematical reasoning and English–Thai math pairs; English–Thai Wikipedia contexts; FineWeb2 Thai and FineWeb-Edu English; SCB-MT parallel text.
BN	Synthetic Bengali mathematical reasoning and English–Bengali math pairs; English–Bengali Wikipedia contexts; Bengali monolingual text (recorded as CulturaX), FineWeb-Edu English, and NLLB parallel text.
SW	Synthetic Swahili mathematical reasoning and English–Swahili math pairs; English–Swahili Wikipedia contexts; CulturaX Swahili, FineWeb-Edu English, and NLLB parallel text.
ZH	Synthetic Chinese mathematical reasoning and English–Chinese math pairs; English–Chinese Wikipedia contexts; CCI3-HQ Chinese, FineWeb-Edu English, and OPUS-100 parallel text.
EN-general	FineWeb-Edu English, subject to the shared source-risk and quality screening.
EN-math	English mathematical reasoning, combining retained historical traces with validated generated alternatives for source questions.

Wikipedia article pairs are matched through language links, interleaved by paragraph, and split into length-bounded samples, adapting the bilingual context principle of CrossIC-PT (Wu et al., 2025). Mathematical bilingual samples pair English and target-language reasoning by question. Parallel data are formatted in both translation directions. The recorded EN/BN/SW/ZH Wikipedia inputs are labeled 2023-11-01.

The English-mathematics pool has 9,826 rows after byte-level deduplication: 5,680 earlier rows plus 4,189 validated DeepSeek-generated candidates, minus 43 duplicates. Most additions are alternative reasoning traces for existing questions. English mathematics is capped at three deterministic

exposure passes during candidate construction. Provenance and quality screening precede packing; examples with unresolved source risk or confirmed normalized benchmark overlap are excluded. Table 8 reports the actual candidate text instances admitted before token packing. The four language-serving pools contribute the same instance counts to Replay and English-role; the English instance counts reflect the role exchange and repeated exposure.

Table 8: Text instances in the final candidate files before token packing, shared by the corresponding 4B/8B runs. Counts include repeated exposure during candidate construction; they are not unique questions or source-pool inventories. Training shares are defined in tokens (Table 1), and every run processes 16,384,000 tokens.

Allocation	TH	BN	SW	ZH	EN-general	EN-math	Total
Uniform	6509	5506	9485	9429	3256	25889	60074
Replay	6243	4949	10254	7956	3609	28727	61738
English-role	6243	4949	10254	7956	4772	18639	52813

A normalized lexical-overlap audit for HumanEval compared the actual three admitted candidate texts with all 164 problems' prompts, canonical solutions, and tests. NFKC normalization, case folding, and whitespace folding precede exact and literal-substring checks on fields of at least 32 characters; these checks found no matches.

Training details. Training uses four H800 80GB GPUs, full-shard FSDP, bf16, sequence length 2,048, gradient checkpointing, per-device microbatch one, and gradient accumulation eight. Thus $2048 \times 1 \times 8 \times 4 = 65,536$ tokens per step and 250 steps yield 16,384,000 tokens. Adafactor disables `scale_parameter`, `relative_step`, and `warmup_init`. Peak learning rate is 10^{-5} with 50 linear warmup steps followed by cosine decay; weight decay is 0.1, maximum gradient norm 1.0, and data-loader workers zero. Repeated seeds at the same allocation reuse the same tokenized pack and order. Checkpoints were independently reloaded before evaluation. The environment uses Python 3.12.3, PyTorch 2.10.0, CUDA 12.8, and Transformers 4.53.2. The tokenizer has 151,669 effective entries and EOS/PAD ID 151643. Each size uses a fixed Base snapshot, shared by all allocations and optimization seeds at that size.

Evaluation details. MGSM uses eight localized exemplars; GSM8K uses four fixed English chain-of-thought exemplars. Generation is greedy, with a 512-new-token cap and batch size four. GSM8K evaluates every model on all 1,319 test questions. Global-MMLU and CMMLU use zero-shot A–D conditional log-probability scoring; C-Eval uses its formatted A–D choices. Global-MMLU selects the first 250 rows after a seed-42 shuffle of each language test file; C-Eval selects the first 250 rows of a seed-42 shuffle of its source records. The selected items are fixed across all model comparisons. GPQA choice order is fixed by seed 42. HumanEval covers all 164 tasks with the same prompt, generation, extraction, and execution settings for Base and candidates. The earlier 100 predictions and remaining 64 are combined in a verified seed-42 shuffle order with unique task IDs. Generated programs run in fresh temporary directories with a five-second timeout. MBPP retains the first 100 items of its seed-42 shuffle.

Documentation and material availability. Tables 7 and 8, together with the procedures above, document data construction, candidate composition, training, and evaluation. The authors retain construction and screening scripts, source specifications, configurations, and records identifying candidate texts and token packs. This submission provides methodological descriptions and aggregate statistics; processed corpora and complete versioned reconstruction manifests are not included.

B Complete Scores and Statistical Results

Tables 9–13 report every primary checkpoint score and the mathematical seed summaries. Accuracies are percentages and differences are percentage points unless a caption states accuracy fractions. GSM8K and HumanEval use their full test sets throughout these primary tables.

Table 9: Complete MGSM accuracy (%) for both model sizes. Worst target is the minimum of TH/BN/SW. Each trained checkpoint uses 16.384M tokens.

Size	Checkpoint	TH	BN	SW	Target avg.	Worst	EN
4B	Base	70.80	65.60	25.20	53.87	25.20	82.00
4B	Uniform 42	78.00	66.40	27.20	57.20	27.20	80.00
4B	Uniform 43	77.20	68.40	26.80	57.47	26.80	79.60
4B	Uniform 44	77.20	66.80	27.60	57.20	27.60	77.20
4B	Replay 42	76.40	69.60	28.40	58.13	28.40	78.40
4B	Replay 43	77.20	68.40	27.20	57.60	27.20	80.80
4B	Replay 44	78.00	65.20	27.20	56.80	27.20	79.20
4B	English-role 42	79.60	68.40	26.40	58.13	26.40	77.20
4B	English-role 43	79.60	67.60	25.20	57.47	25.20	78.40
4B	English-role 44	78.80	68.40	26.80	58.00	26.80	78.40
8B	Base	77.20	71.20	42.40	63.60	42.40	88.00
8B	Uniform 42	80.80	74.00	45.20	66.67	45.20	84.00
8B	Uniform 43	81.20	73.60	48.00	67.60	48.00	84.40
8B	Uniform 44	81.60	72.80	48.40	67.60	48.40	85.60
8B	Replay 42	82.40	73.20	46.80	67.47	46.80	87.20
8B	Replay 43	81.20	73.60	45.60	66.80	45.60	84.40
8B	Replay 44	81.20	71.20	45.20	65.87	45.20	84.80
8B	English-role 42	81.20	73.20	46.00	66.80	46.00	85.20
8B	English-role 43	82.40	72.80	45.60	66.93	45.60	84.00
8B	English-role 44	81.60	73.20	44.00	66.27	44.00	86.80

Table 10: MGSM mean $\pm$ sample SD (percentage points) across optimization seeds: three for every allocation. The target-average SD is computed after averaging languages within each checkpoint.

Size	Allocation	TH	BN	SW	Target avg.	EN
4B	Uniform	77.47 ± 0.46	67.20 ± 1.06	27.20 ± 0.40	57.29 ± 0.15	78.93 ± 1.51
4B	Replay	77.20 ± 0.80	67.73 ± 2.27	27.60 ± 0.69	57.51 ± 0.67	79.47 ± 1.22
4B	English-role	79.33 ± 0.46	68.13 ± 0.46	26.13 ± 0.83	57.87 ± 0.35	78.00 ± 0.69
8B	Uniform	81.20 ± 0.40	73.47 ± 0.61	47.20 ± 1.74	67.29 ± 0.54	84.67 ± 0.83
8B	Replay	81.60 ± 0.69	72.67 ± 1.29	45.87 ± 0.83	66.71 ± 0.80	85.47 ± 1.51
8B	English-role	81.73 ± 0.61	73.07 ± 0.23	45.20 ± 1.06	66.67 ± 0.35	85.33 ± 1.40

Table 11: Complete 4B broad-capability accuracy (%). All allocations cover seeds 42/43/44. GSM8K uses 1,319 items; the remaining tasks retain their task-specific evaluation sets.

Checkpoint	GSM8K	GMMLU-EN	GMMLU-BN	GMMLU-SW
Base	86.81	68.40	49.60	32.00
Uniform 42	90.98	68.80	46.40	34.80
Uniform 43	90.60	68.80	47.60	35.60
Uniform 44	90.60	68.80	45.60	34.40
Replay 42	89.31	68.80	45.20	35.20
Replay 43	90.45	69.20	44.80	33.60
Replay 44	89.76	69.60	46.00	34.80
English-role 42	89.54	69.60	48.00	35.20
English-role 43	89.69	69.20	47.20	35.60
English-role 44	89.31	68.80	44.80	34.40

Checkpoint	GMMLU-ZH	C-Eval	GPQA Diamond	CMMLU math
Base	70.00	76.80	34.34	40.00
Uniform 42	70.00	77.60	33.33	46.67
Uniform 43	69.20	78.40	32.83	46.67
Uniform 44	70.00	77.20	33.33	46.67
Replay 42	68.40	76.80	32.83	46.67
Replay 43	69.20	77.20	33.33	46.67
Replay 44	70.00	76.00	32.32	46.67
English-role 42	70.40	77.60	34.85	53.33
English-role 43	70.40	76.40	32.83	53.33
English-role 44	69.60	77.20	34.85	53.33

Table 12: Complete 8B broad-capability accuracy (%). All allocations cover seeds 42/43/44. GSM8K uses 1,319 items; the remaining tasks retain their task-specific evaluation sets.

Checkpoint	GSM8K	GMMLU-EN	GMMLU-BN	GMMLU-SW
Base	89.69	74.40	53.20	40.00
Uniform 42	91.43	75.20	51.20	41.60
Uniform 43	91.58	74.40	51.60	40.40
Uniform 44	91.58	74.40	52.40	40.80
Replay 42	91.81	76.00	50.80	41.20
Replay 43	91.66	75.60	50.80	41.20
Replay 44	90.90	75.20	50.40	42.40
English-role 42	91.05	75.60	51.60	42.40
English-role 43	91.51	74.00	50.00	42.00
English-role 44	90.90	74.40	50.80	41.60

Checkpoint	GMMLU-ZH	C-Eval	GPQA Diamond	CMMLU math
Base	70.80	79.60	37.37	66.67
Uniform 42	68.80	80.00	38.89	73.33
Uniform 43	68.40	79.60	35.86	73.33
Uniform 44	68.80	79.60	38.38	66.67
Replay 42	68.40	80.40	36.87	66.67
Replay 43	68.80	80.00	37.88	66.67
Replay 44	68.40	80.00	38.38	73.33
English-role 42	68.00	80.40	39.39	73.33
English-role 43	70.00	79.20	38.38	73.33
English-role 44	68.80	80.40	38.89	66.67

Table 13: Complete code-task scores: pass@1 (%) on all 164 HumanEval problems and the fixed 100-item MBPP subset.

Size	Checkpoint	HumanEval	MBPP
4B	Base	54.88	63.00
4B	Uniform 42	60.98	61.00
4B	Uniform 43	62.20	63.00
4B	Replay 42	62.20	62.00
4B	Replay 43	62.80	64.00
4B	English-role 42	62.20	61.00
4B	English-role 43	61.59	61.00
8B	Base	61.59	65.00
8B	Uniform 42	62.20	64.00
8B	Uniform 43	64.02	67.00
8B	Replay 42	62.20	66.00
8B	Replay 43	63.41	67.00
8B	English-role 42	61.59	68.00
8B	English-role 43	62.20	65.00

Paired-item intervals. Base-relative broad-task and code intervals, and the primary MGSM contrasts, use 10,000 paired resamples with seed 20260901. The earlier 8B English-role seed-43 target and interpolation contrasts use 20,000 with seed 20260903. Pairing preserves aligned question or task identities. Reported two-sided intervals use the 2.5% and 97.5% bootstrap percentiles. For target-language averages, languages are resampled independently, which omits dependence between translations of the same question. The multilingual aggregate intervals retain that reporting convention and are not used to establish joint multilingual significance. Intervals quantify uncertainty for fixed checkpoints; sample SDs across three training seeds describe optimization variation. The intervals are marginal and do not adjust for candidate selection or constitute simultaneous coverage claims.

Full-test GSM8K and HumanEval Base-relative two-sided intervals are available for all evaluated CPT checkpoints. Other broad-task intervals are unavailable for English-role44 at both sizes and English-role43 at 8B; 8B Replay43 and English-role43 also lack MBPP intervals. Corresponding scores remain in the complete tables. The expanded GSM8K report provides no one-sided bound, so Table 20 summarizes only available bounds and marks incomplete checks. These bounds are not substituted for the measured-score budget decisions in the main text.

Table 14: 4B broad-retention differences from its Base, in percentage points: point difference [paired-item 95% CI]. GSM8K: 1,319 items; Global-MMLU: 250 items per language. Dashes indicate unavailable paired intervals.

Checkpoint	GSM8K	GMMLU-EN	GMMLU-BN	GMMLU-SW
Uniform 42	+4.17 [+2.50, +5.91]	+0.40 [-2.80, +3.60]	-3.20 [-6.40, +0.00]	+2.80 [-0.80, +6.40]
Uniform 43	+3.79 [+2.12, +5.53]	+0.40 [-2.40, +3.20]	-2.00 [-5.60, +1.60]	+3.60 [+0.00, +7.60]
Uniform 44	+3.79 [+2.12, +5.46]	+0.40 [-2.41, +3.60]	-4.00 [-7.20, -0.80]	+2.40 [-1.60, +6.40]
Replay 42	+2.50 [+0.83, +4.17]	+0.40 [-2.40, +3.20]	-4.40 [-7.61, -1.20]	+3.20 [-0.40, +6.80]
Replay 43	+3.64 [+1.97, +5.31]	+0.80 [-2.00, +3.60]	-4.80 [-8.40, -1.20]	+1.60 [-2.40, +5.60]
Replay 44	+2.96 [+1.29, +4.62]	+1.20 [-1.20, +4.00]	-3.60 [-7.20, +0.00]	+2.80 [-1.20, +6.80]
English-role 42	+2.73 [+1.06, +4.40]	+1.20 [-1.20, +4.00]	-1.60 [-5.60, +2.40]	+3.20 [-0.80, +7.20]
English-role 43	+2.88 [+1.21, +4.55]	+0.80 [-2.00, +3.60]	-2.40 [-6.40, +1.20]	+3.60 [-0.40, +7.60]
English-role 44	+2.50 [+0.83, +4.17]	–	–	–

Table 15: 4B broad-retention differences from its Base, in percentage points: point difference [paired-item 95% CI]. Global-MMLU and C-Eval: 250 items each; GPQA: 198; CMMLU math: 15. Dashes indicate unavailable paired intervals.

Checkpoint	GMMLU-ZH	C-Eval	GPQA Diamond	CMMLU math
Uniform 42	+0.00 [-2.80, +2.80]	+0.80 [-1.20, +2.80]	-1.01 [-5.56, +3.54]	+6.67 [+0.00, +20.00]
Uniform 43	-0.80 [-3.60, +2.00]	+1.60 [-0.40, +4.00]	-1.52 [-5.56, +2.53]	+6.67 [+0.00, +20.00]
Uniform 44	+0.00 [-2.80, +2.80]	+0.40 [-1.60, +2.40]	-1.01 [-5.05, +3.03]	+6.67 [+0.00, +20.00]
Replay 42	-1.60 [-4.40, +1.20]	+0.00 [-2.40, +2.00]	-1.52 [-6.06, +3.03]	+6.67 [+0.00, +20.00]
Replay 43	-0.80 [-3.60, +2.00]	+0.40 [-1.20, +2.00]	-1.01 [-5.56, +3.54]	+6.67 [+0.00, +20.00]
Replay 44	+0.00 [-2.40, +2.40]	-0.80 [-3.20, +1.20]	-2.02 [-6.57, +2.53]	+6.67 [+0.00, +20.00]
English-role 42	+0.40 [-2.00, +3.20]	+0.80 [-1.20, +2.80]	+0.51 [-4.04, +5.05]	+13.33 [+0.00, +33.33]
English-role 43	+0.40 [-2.00, +2.80]	-0.40 [-2.00, +1.20]	-1.52 [-6.06, +2.53]	+13.33 [+0.00, +33.33]
English-role 44	–	–	–	–

Table 16: 8B broad-retention differences from its Base, in percentage points: point difference [paired-item 95% CI]. GSM8K: 1,319 items; Global-MMLU: 250 items per language. Dashes indicate unavailable paired intervals.

Checkpoint	GSM8K	GMMLU-EN	GMMLU-BN	GMMLU-SW
Uniform 42	+1.74 [+0.23, +3.26]	+0.80 [-1.20, +2.80]	-2.00 [-5.20, +1.20]	+1.60 [-2.80, +6.00]
Uniform 43	+1.90 [+0.38, +3.41]	+0.00 [-2.00, +2.00]	-1.60 [-5.20, +2.00]	+0.40 [-3.60, +4.40]
Uniform 44	+1.90 [+0.38, +3.41]	+0.00 [-2.00, +2.00]	-0.80 [-4.00, +2.40]	+0.80 [-3.60, +5.20]
Replay 42	+2.12 [+0.61, +3.64]	+1.60 [-0.40, +4.00]	-2.40 [-5.60, +0.80]	+1.20 [-2.80, +5.20]
Replay 43	+1.97 [+0.38, +3.49]	+1.20 [-0.80, +3.20]	-2.40 [-5.60, +0.80]	+1.20 [-3.20, +5.60]
Replay 44	+1.21 [-0.30, +2.73]	+0.80 [-1.20, +2.80]	-2.80 [-6.00, +0.40]	+2.40 [-2.00, +6.80]
English-role 42	+1.36 [-0.15, +2.88]	+1.20 [-0.80, +3.20]	-1.60 [-4.80, +1.60]	+2.40 [-2.00, +6.80]
English-role 43	+1.82 [+0.23, +3.41]	–	–	–
English-role 44	+1.21 [-0.30, +2.73]	–	–	–

Table 17: 8B broad-retention differences from its Base, in percentage points: point difference [paired-item 95% CI]. Global-MMLU and C-Eval: 250 items each; GPQA: 198; CMMLU math: 15. Dashes indicate unavailable paired intervals.

Checkpoint	GMMLU-ZH	C-Eval	GPQA Diamond	CMMLU math
Uniform 42	-2.00 [-4.00, +0.00]	+0.40 [-0.80, +2.00]	+1.52 [-2.02, +5.05]	+6.67 [+0.00, +20.00]
Uniform 43	-2.40 [-4.80, -0.40]	+0.00 [-1.60, +1.60]	-1.52 [-4.55, +2.02]	+6.67 [+0.00, +20.00]
Uniform 44	-2.00 [-4.00, -0.40]	+0.00 [-1.60, +1.60]	+1.01 [-2.53, +4.55]	+0.00 [+0.00, +0.00]
Replay 42	-2.40 [-4.80, -0.40]	+0.80 [-0.80, +2.40]	-0.51 [-4.55, +3.03]	+0.00 [+0.00, +0.00]
Replay 43	-2.00 [-4.40, +0.40]	+0.40 [-1.20, +2.00]	+0.51 [-3.54, +4.55]	+0.00 [+0.00, +0.00]
Replay 44	-2.40 [-4.80, +0.00]	+0.40 [-0.80, +1.60]	+1.01 [-3.03, +5.05]	+6.67 [+0.00, +20.00]
English-role 42	-2.80 [-5.20, -0.80]	+0.80 [-0.80, +2.40]	+2.02 [-1.01, +5.05]	+6.67 [+0.00, +20.00]
English-role 43	–	–	–	–
English-role 44	–	–	–	–

Table 18: Code differences from the same-size Base in percentage points: point difference [paired-item 95% CI]. HumanEval uses 164 problems; MBPP uses 100. Dashes mark unavailable MBPP intervals.

Size	Checkpoint	HumanEval	MBPP
4B	Uniform 42	+6.10 [+0.61, +11.59]	-2.00 [-6.00, +2.00]
4B	Uniform 43	+7.32 [+2.44, +12.80]	+0.00 [-3.00, +3.00]
4B	Replay 42	+7.32 [+1.83, +12.80]	-1.00 [-5.00, +2.00]
4B	Replay 43	+7.93 [+2.44, +13.41]	+1.00 [-2.00, +4.00]
4B	English-role 42	+7.32 [+2.44, +12.80]	-2.00 [-6.00, +2.00]
4B	English-role 43	+6.71 [+2.44, +11.59]	-2.00 [-6.00, +2.00]
8B	Uniform 42	+0.61 [-4.27, +5.49]	-1.00 [-5.00, +3.00]
8B	Uniform 43	+2.44 [-1.83, +7.32]	+2.00 [-2.00, +6.00]
8B	Replay 42	+0.61 [-4.27, +5.49]	+1.00 [-2.00, +5.00]
8B	Replay 43	+1.83 [-2.44, +6.71]	–
8B	English-role 42	+0.00 [-4.27, +4.27]	+3.00 [-1.00, +8.00]
8B	English-role 43	+0.61 [-3.05, +4.88]	–

Table 19: Additional 8B target-language-average paired-item 95% intervals (accuracy fractions). English-role seed-43 and alpha contrasts use 20,000 resamples; other rows use 10,000. Resampling details are given at the start of this appendix.

Contrast	Target avg. difference	95% CI
Earlier pool − Base	+.0293	[.0000,+.0587]
Uniform42 − Base	+.0307	[.0000,+.0613]
Replay42 − Base	+.0387	[+.0093,+.0680]
Replay43 − Base	+.0320	[+.0027,+.0613]
English-role 43 − Uniform 43	−.0067	[−.0253, +.0120]
English-role 43 − Replay 43	+.0013	[−.0160, +.0187]
Alpha .75 − Uniform43	−.0147	[−.0400, +.0093]
Alpha .75 − English-role 43	−.0080	[−.0333, +.0173]

Table 20: Available one-sided retention checks over the eight broad tasks. Slack is the minimum paired 95% lower bound plus its margin among tasks with a reported one-sided bound, in percentage points. A negative available slack rules out passing all eight checks; otherwise missing bounds leave the result unknown. Expanded GSM8K supplies two-sided intervals only. MGSM-EN is separate.

Size	Checkpoint	Bounds	Avail. min. slack	Limiting avail. task	All-broad pass
4B	Uniform 42	7/8	-2.00	GMMLU-BN	no
4B	Uniform 43	7/8	-1.20	GMMLU-BN	no
4B	Uniform 44	7/8	-2.80	GMMLU-BN	no
4B	Replay 42	7/8	-3.20	GMMLU-BN	no
4B	Replay 43	7/8	-4.00	GMMLU-BN	no
4B	Replay 44	7/8	-2.80	GMMLU-BN	no
4B	English-role 42	7/8	-0.80	GMMLU-BN	no
4B	English-role 43	7/8	-1.60	GMMLU-BN	no
4B	English-role 44	0/8	–	–	unknown
8B	Uniform 42	7/8	-0.40	GMMLU-BN	no
8B	Uniform 43	7/8	-0.40	GMMLU-BN/GMMLU-ZH	no
8B	Uniform 44	7/8	+0.40	GMMLU-BN/GMMLU-ZH	unknown
8B	Replay 42	7/8	-1.20	GMMLU-BN	no
8B	Replay 43	7/8	-1.20	GMMLU-BN	no
8B	Replay 44	7/8	-1.60	GMMLU-BN	no
8B	English-role 42	7/8	-0.80	GMMLU-ZH	no
8B	English-role 43	0/8	–	–	unknown
8B	English-role 44	0/8	–	–	unknown

Table 21: Replay minus matched-seed Uniform, in percentage points. Entries are point difference [paired-item 95% CI]. Intervals are conditional on fixed checkpoints and do not represent training-seed uncertainty.

Size	Seed	Target avg.	Worst target	MGSM-EN
4B	42	+0.93 [-0.93, +2.80]	+1.20 [-2.80, +5.20]	-1.60 [-4.40, +1.20]
4B	43	+0.13 [-1.87, +2.00]	+0.40 [-2.80, +4.00]	+1.20 [-2.00, +4.40]
4B	44	-0.40 [-2.27, +1.47]	-0.40 [-3.60, +2.80]	+2.00 [-1.20, +5.20]
8B	42	+0.80 [-0.93, +2.53]	+1.60 [-2.00, +5.20]	+3.20 [+0.80, +6.00]
8B	43	-0.80 [-2.67, +1.07]	-2.40 [-6.00, +1.20]	+0.00 [-2.80, +2.80]
8B	44	-1.73 [-3.60, +0.13]	-3.20 [-6.80, +0.40]	-0.80 [-3.20, +1.60]

Table 22: English-role minus matched-seed Replay, in percentage points. The allocation moves 6.5 points from EN-math to EN-general while their combined share remains 37%. GSM8K and HumanEval use the complete test sets; code covers seeds 42/43.

	4B			8B		
Capability	42	43	44	42	43	44
TH	+3.20	+2.40	+0.80	-1.20	+1.20	+0.40
BN	-1.20	-0.80	+3.20	+0.00	-0.80	+2.00
SW	-2.00	-2.00	-0.40	-0.80	+0.00	-1.20
Target avg.	+0.00	-0.13	+1.20	-0.67	+0.13	+0.40
MGSM-EN	-1.20	-2.40	-0.80	-2.00	-0.40	+2.00
C-Eval	+0.80	-0.80	+1.20	+0.00	-0.80	+0.40
CMMLU math	+6.67	+6.67	+6.67	+6.67	+6.67	-6.67
Global-MMLU BN	+2.80	+2.40	-1.20	+0.80	-0.80	+0.40
Global-MMLU EN	+0.80	+0.00	-0.80	-0.40	-1.60	-0.80
Global-MMLU SW	+0.00	+2.00	-0.40	+1.20	+0.80	-0.80
Global-MMLU ZH	+2.00	+1.20	-0.40	-0.40	+1.20	+0.40
GPQA Diamond	+2.02	-0.51	+2.53	+2.53	+0.51	+0.51
GSM8K	+0.23	-0.76	-0.45	-0.76	-0.15	+0.00
HumanEval	+0.00	-1.22	–	-0.61	-1.22	–
MBPP	-1.00	-3.00	–	+2.00	-2.00	–

C Study Coverage and Local Diagnostics

The study began with exploratory 8B comparisons whose outcomes informed continuation. Subsequent 4B comparisons and evaluation extensions fixed their allocations, added seeds, and evaluation plans before their new results were observed. The final matrix reports all three seeds for every primary allocation at both sizes; code covers two seeds each.

Swahili-heavy neighbor. The 8B neighbor moves two percentage points from TH to SW relative to Replay, giving weights 14/15/20/14/18.5/18.5. At seed 42 it reaches TH/BN/SW .812/.720/.464, target average .6653, and MGSM-EN .832. Against Replay42, target average changes by $-.0093$ $[-.0253, +.0067]$, worst target by $-.004$ $[-.036, +.028]$, and MGSM-EN by $-.040$ $[-.068, -.012]$.

Interpolation toward Base. The 8B diagnostic uses $\theta_{.75} = \theta_0 + .75(\theta_{\text{Uniform43}} - \theta_0)$. It yields TH/BN/SW .808/.748/.428, target average .6613, and MGSM-EN .840. The target average is below the study's .6660 continuation threshold. Both local diagnostics have MGSM results only; the single interpolation point does not isolate an update-magnitude mechanism.

D Budget and Candidate Comparisons

Reference budgets. The 4/5-point budgets were recorded in the initial protocol after inspection of historical experiments and before the subsequent prospective training. They are researcher-set reference requirements. Their common absolute units hold the allowed score losses constant across model sizes. Scaling them by λ describes stricter requirements with these relative task weights held fixed; ρ permits any nonnegative scale to be examined. CMMLU mathematics has a coarse 15-item resolution, with one answer worth 6.67 points. Every primary checkpoint matches or improves its Base score, so omitting this diagnostic leaves all reported ρ values, eligibility sets, and selected candidates unchanged.

Limiting capabilities. At 4B, Bengali Global-MMLU sets ρ for Uniform42, all Replay seeds, and English-role44. MGSM-English sets it for Uniform43/44 and English-role42/43. At 8B, Bengali and Chinese Global-MMLU jointly limit Replay42; English MGSM and Chinese Global-MMLU jointly limit English-role42; Bengali Global-MMLU limits English-role44. English MGSM limits the remaining primary checkpoints. At $\lambda = .75$, only Uniform43 qualifies at 4B; at 8B, Uniform44, Replay42, and English-role42/44 qualify. Base meets all nonnegative budget scales.

Auxiliary Pareto comparison. For completeness, Table 23 reports nondominance within each model size and matched seed in (A, W, S), where $S(p) = \min_j\{R_j(p) + \epsilon_j\}$ is the smallest remaining absolute-score allowance. This auxiliary view retains weakest-target performance alongside acquisition and slack. The main selection uses Equation 2; ρ expresses the required proportional budget scale. A full task vector preserves benefits that a minimum can hide (Figure 3); code is omitted symmetrically from seed-44 comparisons because those models have no code evaluation.

Table 23: Acquisition A, worst-target accuracy W, and minimum retention slack S, in percentage points (accuracy percent for W). Nondominance is computed in (A, W, S) within each model size and matched seed, including Base. All groups contain Base and the three allocations.

Size	Checkpoint	$A\uparrow$	$W\uparrow$	$S\uparrow$	Nondominated
4B	Base	0.00	25.20	4.00	yes
4B	Uniform 42	+3.33	27.20	+0.80	yes
4B	Uniform 43	+3.60	26.80	+1.60	yes
4B	Uniform 44	+3.33	27.60	-0.80	yes
4B	Replay 42	+4.27	28.40	-0.40	yes
4B	Replay 43	+3.73	27.20	-0.80	yes
4B	Replay 44	+2.93	27.20	+0.40	yes
4B	English-role 42	+4.27	26.40	-0.80	no
4B	English-role 43	+3.60	25.20	+0.40	no
4B	English-role 44	+4.13	26.80	-0.80	yes
8B	Base	0.00	42.40	4.00	yes
8B	Uniform 42	+3.07	45.20	+0.00	no
8B	Uniform 43	+4.00	48.00	+0.40	yes
8B	Uniform 44	+4.00	48.40	+1.60	yes
8B	Replay 42	+3.87	46.80	+1.60	yes
8B	Replay 43	+3.20	45.60	+0.40	no
8B	Replay 44	+2.27	45.20	+0.80	no
8B	English-role 42	+3.20	46.00	+1.20	no
8B	English-role 43	+3.33	45.60	+0.00	no
8B	English-role 44	+2.67	44.00	+1.60	no

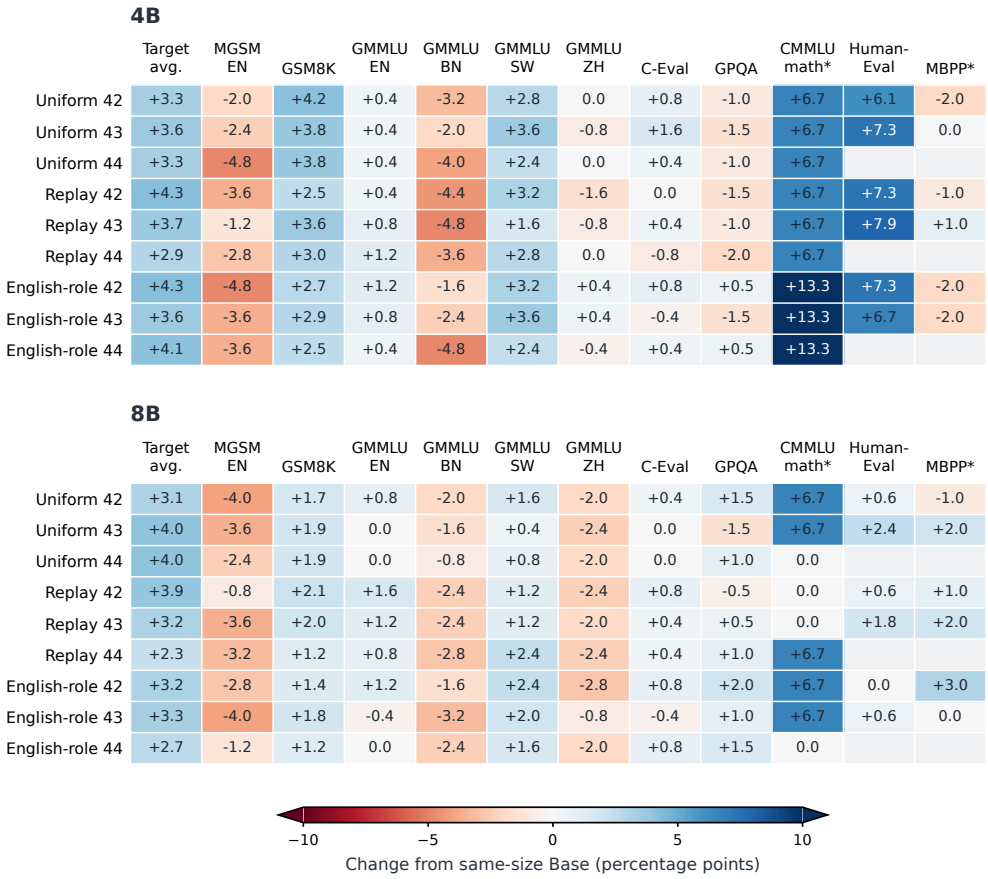


Figure 3: Per-checkpoint capability changes from the same-size Base, in percentage points. GSM8K and HumanEval use all 1,319 and 164 test items, respectively. Asterisks mark the 15-item CMMLU mathematics and 100-item MBPP diagnostics. Code is available for seeds 42/43; seed-44 code cells are blank. Colors show score differences.