

Which Loss Prices the English Share? What Target-Language Validation Loss Can and Cannot See in Continued Pretraining for Language Adaptation, Why Two Training Regimes Disagree About What English Buys, and the Instruments a Mixture Choice Needs

Pranay Mahendrakar

SONYTECH · ORCID 0009-0003-7224-029X

pranaymahendrakar@sonytech.in · mahendrakarpranay@gmail.com · +91 6361723454

Abstract. When an English-centric language model is adapted to a new language by continued pretraining, practitioners usually keep a share of English in the training mixture, and a growing family of mixture laws chooses that share by predicting validation loss. A 2025 study found that adding English left target-language validation perplexity unchanged while deciding whether few-shot ability in the target language emerged at all, and the result invites the reading that perplexity cannot predict emergent abilities. In this analysis the author reviews and synthesises published results to show that the evidence supports a narrower and more useful conclusion. A loss measured on held-out text from the target corpus registers only what is expressed in that text. It cannot rank mixtures, because what the English share changes is expressed elsewhere, and even within one run it can miss an early collapse of an ability it does not measure. The same study's loss on few-shot answer labels moved within the first few steps of training, thousands of steps before accuracy did; and an independent 2026 study of bilingual pretraining from scratch, over roughly a thousand runs, found that target loss chose good checkpoints within a run but could not rank bilingual against monolingual training. The two regimes then disagree about what English supplies. In continued pretraining it acts mainly as a damper on early parameter drift that protects in-context learning, and a curriculum that drops it after the first tenth of training, or weight averaging without it, recovers most of the downstream benefit. Trained from scratch on scarce target data, English supplies knowledge and diversity the target corpus lacks; measured as equivalent unique target data, mixing was worth two to three times on validation loss and two to thirteen times on benchmark accuracy. The paper consolidates the measured values in two tables and a figure, states the mixture choice as a decision procedure keyed to the instrument each comparison needs, and names the experiments that would separate the accounts.

Keywords: continued pretraining; language adaptation; data mixtures; validation loss; in-context learning; catastrophic forgetting; multilingual language models; emergent abilities

SONYTECH ORCID: 0009-0003-7224-029X

1. Introduction

Most large language models are trained on corpora dominated by English, and their ability in other languages falls off with the share of those languages in pretraining. The cheapest way to repair that for a particular language is to keep training the existing model on text in that language, a procedure called continued pretraining (CPT). It has produced adapted models for Basque (Etxaniz et al., 2024), Japanese (Fujii et al., 2024), Bulgarian and German (Alexandrov et al., 2024), Dutch (Meeus et al., 2024), Urdu (Hassan, Ahmed and Awais, 2026), several South-East Asian languages (Dou et al., 2024) and hundreds of languages at once (Lin et al., 2024; Ji et al., 2024). Many of these recipes share one design decision: a portion of English, or of an approximation to the original pretraining distribution, is kept in the mixture alongside the target-language text. Latxa used 500,000 documents from the Pile; Qalb added 140 million tokens of English Wikipedia to 1.84 billion tokens of Urdu; Sailor replayed SlimPajama and SkyPile.

Where a reason is given, it usually comes from the continual learning literature: replaying old data prevents catastrophic forgetting (Kirkpatrick et al., 2017; Ibrahim et al., 2024).

How large the English share should be is now often decided by a fitted law rather than by habit. D-CPT Law (Que et al., 2024) fits the validation loss on a general corpus and on a domain corpus as functions of the mixture ratio, model size and data size, and reads the optimal ratio off the fit. The CMR scaling law (Gu et al., 2024) fits a power law between loss, mixture ratio and training tokens to locate a critical mixture ratio. Data mixing laws (Ye et al., 2024), RegMix (Liu et al., 2024), DoReMi (Xie et al., 2023) and the mixture scaling laws of Shukor et al. (2025) all predict loss from domain weights. Sailor, an adapted model for South-East Asian languages, chose its language weights by fitting a linear regression from 64 proxy runs to a joint loss across languages (Dou et al., 2024). The multilingual scaling laws of He et al. (2024) and the cross-lingual transfer matrix of ATLAS (Longpre et al., 2025) are likewise measured in test loss. Behind them stands an

assumption, well supported in monolingual pretraining on a fixed corpus, that loss tracks capability (Xia et al., 2023; Du et al., 2024; Gadre et al., 2024).

Elhady, Agirre and Artetxe (2025) report a result that does not fit this picture. They continued pretraining Llama 2 7B on Basque with and without a 20 percent share of English. The two runs reached almost the same Basque validation perplexity, 3.35 with English and 3.58 without, and followed nearly identical perplexity curves throughout. Their downstream behaviour did not match. With English, accuracy on Basque multiple-choice benchmarks rose by 8 points between steps 2,000 and 4,000; without it, accuracy never took off. On a language-agnostic in-context learning probe the authors built for the purpose, the run without English lost most of the base model's ability within the first few steps, falling from 44.67 to 20.12 by the end of training, while the run with English held 43.43. The authors connect the loss to a large early shift in parameters and show that two alternatives, including English only during the first 10 percent of steps or averaging weights over time without English, recover most of the downstream benefit.

The finding invites a short summary: "validation perplexity does not predict emergent abilities". In this analysis the author reviews and synthesises published results to show that the summary is both too strong and too weak. It is too strong because the same paper contains a perplexity that did see the damage, and early: the perplexity of answer labels on few-shot prompts in the target language spiked in the first steps of the run without English, at the same time as the in-context probe collapsed. What failed was not perplexity as a kind of measurement. It was perplexity on one particular distribution, held-out text drawn from the same corpus as the training data, which is the distribution on which a mixture law's target term is fitted. It is too weak because the problem does not stop at emergent abilities. A loss on held-out target text cannot, in principle, price anything a mixture changes that is not expressed in target text, and an English share changes several such things.

The argument has five parts. The first concerns scope. The claim that models of equal loss perform equally downstream was established for models trained on a fixed corpus (Du et al., 2024; Xia et al., 2023). Two arms of a mixture experiment do not share a corpus. The Basque result is therefore outside the stated scope of the claim it is said to contradict, and it is better read as a demonstration of why that scope condition matters for mixture choice.

The second part shows that two independent regimes converge on the same failure: target validation loss cannot rank mixtures that differ in English share. Jeha et al. (2026) pretrained Arabic models from scratch, from 150 million to 1.43 billion parameters, with and without

English, over approximately a thousand runs. They found that Arabic validation loss "systematically underestimates mixing's value": on loss, more unique Arabic data always beat mixing in English, while on benchmarks the ranking reversed. They also found that within a single run, Arabic validation loss selected checkpoints whose accuracy was close to the run's best, with a median gap of 0.21 points (Arabic) and 0.13 (English) across about 700 bilingual runs. Their summary is that validation loss "is a reasonable proxy within a paradigm but cannot rank monolingual and bilingual models against each other". The across-mixture half of that summary is what the Basque result implies, reached in a regime where nothing is being forgotten because nothing was learned first. The within-run half holds less well in the Basque data, where target perplexity tracked closely in a run whose in-context ability collapsed, and Section 5 takes up the difference.

The third part examines what English buys, and here the regimes disagree. In CPT from an English-centric base, the evidence points to English as a damper on the early transition: the critical damage happens in the first thousand steps, parameter drift without English is 7 times larger by step 100 and 15 times larger by step 1,000, and on downstream accuracy either a short English curriculum or weight averaging substitutes for English throughout (Elhady et al., 2025). Alexandrov et al. (2024) report, for Llama 3 8B adapted to Bulgarian, that both learning and forgetting "strongly correlated with weight change magnitude". In from-scratch training on scarce target data, the evidence points instead to English as a continuing source of knowledge and diversity whose value grows with model size (Jeha et al., 2026; Seto et al., 2024). These two accounts make opposite predictions about curricula and about whether a regulariser can stand in for data. A third account, that English keeps in use an English-routed internal representation on which target-language processing depends (Wendler et al., 2024; Wu et al., 2024), makes predictions of its own.

The fourth part reads the Basque study's own tables more closely than its summary does. English was not sufficient to protect in-context learning in Arabic or Indonesian, where the probe fell from 44.67 to 32.60 and 30.79 even with English in the mixture. Weight averaging without English preserved more of it in Arabic, 42.76, than English did. For Llama 3.1 8B, whose base model was already far better at Basque, the downstream difference between the arms was 0.91 points. Lowering the learning rate without English did not repair the probe and slowed learning of the language. These rows do not undo the headline. They locate it: the English effect was largest where the base model was weakest in the target language, which is where the switch of distribution is most abrupt, and was small for base models already competent in it.

The fifth part turns the analysis into instruments. If the English share is to be chosen by a fitted law, the law needs a term that can see what the share changes. The paper sets out which loss can see which effect, gives the evidence for each, and proposes a decision procedure (Algorithm 1) and a set of experiments that would separate the accounts.

What the paper contributes is the reframing of the perplexity-emergence result as a statement about the support of the measuring distribution, the evidence that its across-mixture consequence holds in two regimes that otherwise disagree, and the separation of three accounts of what English buys by the predictions they make. Section 2 fixes terms. Section 3 states the Basque finding and the reading the author contests. Section 4 sets out the practice of pricing mixtures by loss. Section 5 presents the from-scratch regime. Section 6 sets out the three accounts of what English supplies. Section 7 reads the complicating rows. Section 8 asks why a loss on prompts saw what a loss on corpus text missed, and Section 9 asks whether a loss on English text would have seen it. Section 10 gathers the evidence in tables and a figure, Section 11 states the decision structure, Section 12 proposes the experiments, and the remaining sections answer objections, list what is not known, state limitations and conclude.

2. Scope and terms

Continued pretraining (CPT) for language adaptation. Training an existing pretrained language model further, with the next-token objective, on a corpus dominated by a target language the base model handles poorly. The paper's evidence concerns decoder-only models from 7 to 13 billion parameters in this setting, and from-scratch bilingual pretraining from 150 million to 1.43 billion parameters in the comparison regime. Instruction tuning and preference tuning appear only where their forgetting results bear on the mechanism.

Source share. The fraction of training tokens drawn from the base model's dominant language, usually English, or from an approximation of its original pretraining distribution. The continual learning literature calls this replay; the language adaptation literature often calls it mixing. The paper uses "source share" when the distinction does not matter and "English share" when the source is specifically English.

Arm. One training run, or one family of runs, defined by a fixed mixture and schedule. A within-arm comparison compares checkpoints, seeds or hyperparameters under one mixture. An across-arm comparison compares mixtures.

Four instruments. The paper distinguishes four measurements that are often collapsed into "perplexity" and "accuracy". Target validation loss (TVL) is next-token loss on held-out text drawn from the same corpus as the target-language training data; in Elhady et al. (2025) the

validation set is a random split of the training corpus, so it shares the training distribution. Source validation loss (SVL) is next-token loss on held-out text from the source distribution, typically English; D-CPT Law and the CMR law fit this term alongside the domain term, and Ibrahim et al. (2024) use it as their measure of forgetting. Prompt-distribution loss (PDL) is loss measured on the distribution of the use the model will be put to, for example the perplexity of answer labels given few-shot demonstrations and a question, or exact-match on a pattern-inference probe whose format is a few-shot prompt; Elhady et al. measure both kinds. Downstream accuracy (DA) is the thresholded outcome, usually the fraction of multiple-choice items on which the correct option scores highest.

Two regimes. Regime A is CPT from a base model trained mostly on English, with a target corpus that is seen a small number of times. Regime B is pretraining from random initialisation on a target corpus too small for the compute budget, so that the target text is repeated many times, with or without an auxiliary high-resource language. The regimes matter because forgetting is possible only in A, while heavy repetition is present mainly in B.

What the paper does not cover. Vocabulary extension and tokenizer replacement are a separate design axis (Fujii et al., 2024; Csaki et al., 2023; Yamaguchi et al., 2024) and are held fixed in the anchor study. Massive multi-language mixtures (Ji et al., 2024; Lin et al., 2024) raise a curse-of-multilinguality question (Chang et al., 2023) that the paper touches only in passing. The paper does not evaluate whether any particular deployed adapted model chose its share well. Broad surveys of continual learning in language models (Shi et al., 2024) and of multilingual language models (Zhu et al., 2024) cover replay and cross-lingual transfer in general terms and predate both the anchor study and the from-scratch comparison; neither addresses which loss should set the source share.

3. The finding and the reading this paper contests

Elhady et al. (2025) continued pretraining four base models on Basque and Llama 2 7B on Arabic and Indonesian, each for 10,000 steps with a peak learning rate of $1e-4$, cosine schedule, sequence length 4,096 and effective batch size 256. The target corpora were matched at 4.5 to 4.7 billion tokens. In the arm with English, English accounted for 20 percent of training tokens, drawn from 500,000 Pile documents. They report three quantities at the end of training: target validation perplexity, average accuracy on native target-language multiple-choice benchmarks, and accuracy on Copain, a seven-task benchmark of 1,050 items in which the model must infer from few-shot demonstrations which element of a short list of numbers or letters to output, with no instruction in natural language.

Their Table 2 contains the paper's central comparison. Across the six base-model and language pairs, the arm with English had lower validation perplexity in three cases and higher in three, with small differences in each. It had higher downstream accuracy in all six, by between 0.91 points (Llama 3.1 8B, Basque) and 7.32 points (Llama 2 13B, Basque). It had higher Copain accuracy in all six, by between 0.85 and 23.31 points. Over the course of training the two perplexity curves for Llama 2 7B on Basque are close to indistinguishable; the downstream curve for the English arm rises by 8 points between steps 2,000 and 4,000, and the curve for the other arm does not. The authors describe this as a challenge to "prior observations in monolingual settings that models with a similar perplexity tend to obtain similar performance in downstream tasks", citing Du et al. (2024) and Xia et al. (2023).

Two further measurements in the same study carry the argument of this paper. First, the authors recast each multiple-choice item as conditional generation and computed the perplexity of the correct answer label and of the incorrect labels, given the few-shot demonstrations and the question. With English, the perplexity of incorrect labels rose early in training while that of correct labels held steady or fell slightly. Without English, both spiked in the first few steps, stayed high relative to the other arm, and the gap between correct and incorrect remained much smaller. The damage, in the authors' words, "happens much earlier" than it shows in accuracy, around step 3,000. Second, they measured parameter drift from the initial model. Without English, the cumulative L2 distance was 7 times larger at step 100 and 15 times larger by step 1,000, and the run moved further in its first 10 steps than the English run did in all 10,000.

The reading this paper contests is that the result shows perplexity, as such, to be blind to emergence. Three features of the evidence tell against it.

The first is the scope of the claim being contradicted. Du et al. (2024) state their equal-loss, equal-performance result for models trained "with a fixed data corpus, tokenization, and model architecture". Xia et al. (2023) analysed checkpoints of one model family on one corpus. The two arms in Elhady et al. are trained on different corpora by construction: that is the experimental variable. A demonstration that two models trained on different data can share a loss on one held-out set while differing downstream is not a counterexample to a claim made for models that share their training data. Liu, Xie, Li and Ma (2022) had already shown that equal pretraining loss does not fix downstream performance even within one corpus, through implicit bias towards flatter minima; across corpora there is still less reason to expect it.

The second is that a perplexity did see the damage. The label perplexity measured in Section 4.3 of the Basque

study is a loss, computed by the same next-token machinery, on a different distribution: target-language few-shot prompts. It moved within the first few steps, at the same time as the Copain collapse and the parameter drift, and roughly 3,000 steps before downstream accuracy separated. On that distribution, loss was an early and sensitive instrument.

The third is the authors' own account of what the validation set measures. Their footnote 3 states that training and validation sets are random splits of the same corpus and "thus come from the same distribution". Section 4.3 of the paper describes the two arms as performing alike "when evaluated in the training distribution" while generalising differently to few-shot tasks "out of this distribution". The authors therefore frame the result in exactly the terms this paper uses. What is contested here is not their analysis but the shorthand the result invites, and its use as evidence about loss-based mixture selection in general.

Read with its scope restored, the finding is this. On held-out target corpus text, the effect of a 20 percent English share was small and of inconsistent sign across the six pairs. On target-language few-shot prompts, the effect was large and immediate. On downstream accuracy, the effect was large and delayed. The question the result poses for practice is which of these instruments a mixture law is fitted on.

4. How mixtures are priced, and what each price can see

The literature that chooses mixture weights by prediction is large and has converged on loss as its currency. This section sets out which losses the main methods use, so that the gap identified in Section 3 can be located precisely.

Laws fitted on target or domain loss alone. Data mixing laws (Ye et al., 2024) predict validation loss on each domain from the mixture proportions and choose proportions that minimise a weighted loss. RegMix (Liu et al., 2024) trains 512 models of 1 million parameters on 1 billion tokens each and regresses loss on mixture weights. Shukor et al. (2025) fit scaling laws that predict loss as a function of model size, data size and domain weights and extrapolate to unseen weights. Sailor (Dou et al., 2024) fitted its regression to a joint loss defined as the product of per-language losses across all its languages, which by the paper's description include English and Chinese as well as the South-East Asian targets. In each case the quantity optimised is a loss on held-out text from the training domains.

Laws fitted on a target term and a source term. D-CPT Law (Que et al., 2024) defines two validation losses, one on the general corpus and one on the domain corpus, and fits a law for each as a function of the mixture ratio; the user trades them off. The CMR law (Gu et al., 2024) defines a feasible mixture ratio as one under

which general loss stays within a stated tolerance, and takes the largest feasible domain share as its critical ratio; its authors list downstream evaluation among the limitations of the study. Wang et al. (2025) model the CPT loss curve as a transition between two hidden curves and use it to tune replay ratio and learning-rate schedule against general and domain validation loss. These methods include a source term, which is the term that could in principle register forgetting. Section 9 asks whether it would have registered the Basque damage.

Studies of replay that measure forgetting by loss.

Ibrahim et al. (2024) continued pretraining a 405 million parameter model, first trained on 300 billion Pile tokens, on 200 billion tokens of German Common Crawl, with replay shares of 1, 5, 10, 25 and 50 percent. They chose 25 percent for the German shift because several shares gave similar final validation loss and that share left more room for new tokens. They report English downstream results and a German HellaSwag score, but decline to report an average over their German evaluations because those sat near chance at that scale, so that loss was the main instrument with resolution on the target side. Gupta et al. (2023) studied learning-rate re-warming in the same framework, again with loss as the primary measure.

Studies that price cross-lingual transfer by loss. He et al. (2024) fit multilingual scaling laws at the level of language families using test cross-entropy. ATLAS (Longpre et al., 2025) measures how training on one language changes test loss on another, using a vocabulary-insensitive loss, and derives a 38 by 38 transfer matrix from it. Chang et al. (2023) trained more than 10,000 small models across 250 languages and report that a moderate amount of added multilingual data helps language modelling in a low-resource language by an amount comparable to increasing the low-resource dataset by up to 33 percent; their outcome measure is language-modelling performance.

Studies of language shift that report transfer in both directions. Gogoulou et al. (2024) continued training a monolingual English model on Danish, Icelandic and Norwegian in sequence and found forward transfer to the new languages largely positive and backward transfer to earlier ones positive or negative depending on language order and characteristics, with language contamination and syntactic similarity the best-fitting explanatory factors. Zhao et al. (2024), evaluating with C-Eval, MMLU, AGI-Eval and GAOKAO-Bench and reporting similar trends across thirteen low-resource languages, report that transferring LLaMA's generation and instruction-following abilities to a non-English language reached performance comparable to state-of-the-art transfer models with less than 1 percent of the pretraining data. Both measure what CPT changes in the languages involved; neither isolates the role of the source share.

Practice that uses both loss and benchmarks. Xi et al. (2024) chose the additional-language mixture ratio for continued pretraining of Llama 3 8B on Chinese by studying its interaction with learning rate under both validation loss and benchmark performance, and argue that a CMR-style choice of the largest feasible ratio may be suboptimal downstream.

Set beside Section 3, the picture is plain. Every fitted law listed prices the mixture by loss on held-out text from the training domains. Xi et al. (2024) are the one listed practice that already brings in a downstream instrument, though they do not fit it as a term of a law. Those that include a source term can see what the mixture does to the source distribution. None includes a term on the distribution of use, such as target-language few-shot prompts, which is where the Basque damage was visible first. The point is not that these methods are wrong. Within their own validation, several were checked against downstream outcomes: DoReMi reports a 6.5 point average gain in few-shot accuracy from loss-derived weights (Xie et al., 2023). The point is that their target term cannot see what an English share changes off the target distribution. The source term may or may not see it. That depends on whether the damage is expressed in source text, which is an empirical question.

One more practice deserves mention because it is the default. Many adaptation recipes fix the source share by convention and explain it only as a guard against catastrophic forgetting (Hassan et al., 2026). The Latxa paper is informative here because it records why. Its authors write that their "preliminary experiments showed that adding English data was critical for strong few-shot performance", and report that an early 7B version trained without English scored 23.67 on Belebele (Bandarkar et al., 2024), 13 points below the corresponding version with English. The convention was set by a downstream instrument and justified in the vocabulary of forgetting. Elhady et al. (2025) is the systematic study of what that preliminary observation had found.

5. The second regime: target-language loss underprices English when nothing is forgotten

If target validation loss failed to see the English effect only in CPT, the explanation could be specific to forgetting. Jeha et al. (2026) remove that possibility by studying a regime in which there is nothing to forget.

They pretrained LLaMA-style models of 150 million, 380 million, 600 million and 1.43 billion parameters from random initialisation, with Arabic as a data-constrained target and English as an auxiliary language, over approximately 1,000 runs. The unique Arabic budget in the main comparison was 200 million tokens, repeated as often as the compute budget required. Two levers were compared: tuning hyperparameters, chiefly weight decay and learning rate, against overfitting to repeated

Arabic, and mixing in English at a tuned ratio. The first lever follows Kim et al. (2025), who found that in data-constrained pretraining the optimal weight decay is about 30 times larger than standard practice. Downstream evaluation used ARC Easy at 5 shots in English and in an Arabic translation produced with NLLB-200.

On Arabic validation loss, mixing helped and its advantage grew with scale. Mixing improved loss over the monolingual baseline by 5.7, 8.7 and 12.9 percent at 380 million, 600 million and 1.43 billion parameters, while hyperparameter tuning alone improved it by 1.2, 0.9 and 3.1 percent. The optimal Arabic fraction fell from 0.33 at 380 million parameters to 0.10 at 1.43 billion. But on validation loss, more unique Arabic data always beat mixing: 2.5 times more Arabic data alone matched the best bilingual loss at 600 million parameters.

On benchmarks the ranking flipped. Monolingual Arabic accuracy barely moved with scale, from 29.7 percent at 150 million parameters to 31.1 or 32.3 percent at 1.43 billion depending on tuning, against a 25 percent chance level. Bilingual training improved Arabic accuracy over monolingual training at the same hyperparameters by 0.4 points at 380 million rising to 6.3 points at 1.43 billion. At 1.43 billion, a monolingual run that had seen 18.4 billion Arabic tokens reached 33.6 percent, while a bilingual run that had seen 9.2 billion reached 41.4 percent; on English ARC Easy the corresponding figures were 29.6 and 57.7 percent. The authors express the gap as a data multiplier, the amount of unique Arabic data a monolingual model would need to match the bilingual one. On validation loss that multiplier was 2 to 3 times; on benchmark accuracy it was 2 to 13 times, and at 1.43 billion parameters the bilingual model trained on 200 million Arabic tokens matched the benchmark accuracy of a monolingual model trained on 1.6 billion unique tokens, or 2.7 billion with hyperparameter tuning on top.

The authors' interpretation is that English plays two roles: it regularises by diversifying the training signal, and it supplies knowledge that the repeated target corpus does not contain. Validation loss on held-out target text registers the first and is blind to the second. They draw the practical conclusion directly: do not use validation loss alone to compare monolingual and bilingual strategies.

They also report the converse, which matters as much. Within a single bilingual run, the checkpoint with the best Arabic validation loss had accuracy close to that run's peak, with median absolute gaps of 0.21 and 0.13 points for Arabic and English. Validation loss is, in their words, "a reasonable proxy within a paradigm but cannot rank monolingual and bilingual models against each other."

Seto et al. (2024), from overlapping authors, studied the same question with German and other targets and motivate their study with the same asymmetry: a modest amount of target text, they suppose, suffices to reach a reasonable perplexity but not to perform well on knowledge-based downstream tasks. They report that gains from better English pretraining data carried over to target-language settings with limited data, particularly for languages close to English.

Two cautions belong with this evidence. The Arabic benchmark is a machine translation of an English benchmark, and Jeha et al. list "translated evaluation set" and "single-seed runs" among their own limitations. A translated benchmark may reward knowledge of English-language facts and English-shaped question structure more than a native one would, and translation artefacts are a known confound in cross-lingual evaluation; Elhady et al. chose native benchmarks to avoid them. The size of the benchmark multiplier may therefore be inflated in a way that the existence of the gap is not. Second, at these scales the absolute accuracies are low, and the monolingual baselines sit within a few points of chance, so the multipliers rest on small differences near the floor.

Even with these cautions, the convergence is the important fact. In regime A, where the base model knows English and can forget, target validation loss failed to rank arms that differed in English share. In regime B, where the model starts from nothing and cannot forget, target validation loss failed to rank arms that differed in English share. Within a run the picture is less symmetric. In regime B, checkpoint selection by target loss came close to each run's best benchmark accuracy, with a median gap of 0.21 points (Arabic) and 0.13 (English) across about 700 bilingual runs. In regime A, target perplexity followed nearly the same curve in both runs while in-context ability collapsed early in one of them, so even within that run the loss did not register the damage; the anchor study does not show whether choosing that run's checkpoint by target loss would have cost downstream accuracy. Within-run validity is therefore contingent: it holds where the abilities the loss does not measure happen to move with the ones it does, and an abrupt early shift is exactly where they need not. The common structure is not about forgetting and not about emergence. It is about what a loss on one distribution can register about changes that live on another.

6. What English buys: three accounts and what separates them

That target validation loss cannot price the English share in either regime does not mean English buys the same thing in both. Three accounts are on offer in the literature, and they are not exclusive. They do make different predictions, which is what allows the evidence to weigh them.

6.1 The damper account

On this account, the source share matters mainly at the moment of transition. Switching abruptly from a mostly English distribution to a target-language distribution produces large early gradient updates. These move the parameters far from the base model and disrupt capabilities that do not depend on the target language, in-context learning above all. A source share softens the switch, keeps the early updates small, and lets the model learn the target language without first breaking what it already had.

Elhady et al. (2025) state this account and supply most of its support. Their Copain trajectory for the run without English falls almost to zero within the first few steps and recovers only partially over 10,000 steps; their parameter-drift measurements put the difference between arms in the first 10 to 1,000 steps; their curriculum, English only for the first 1,000 steps, matches the full-English arm, with each winning half the comparisons on downstream accuracy and on Copain; and exponential moving averaging of weights with decay 0.92, without any English, is competitive downstream. Their summary of the curriculum result is that "the role of English is to provide a smoother transition to the target language distribution".

Independent work supports each link. The stability gap, a temporary loss of performance on earlier tasks at the start of training on a new one, followed by partial or full recovery, was identified in continual learning by De Lange, van de Ven and Tuytelaars (2023), who note that it had gone unnoticed because evaluations were run only at task boundaries. Caccia et al. (2022) traced a related effect to abrupt representation change when new classes first appear. Guo et al. (2024) observed the same pattern in continued pretraining of language models on medical text, and among three strategies for shortening it proposed training on a data mixture closer to the original pretraining distribution. In instruction tuning, Wang et al. (2024) found that fine-tuning on a single task reduces general in-context learning through what they call format specialization, and that it happens at the very beginning of fine-tuning. Zheng et al. (2025) argue that many apparent performance drops in continual learning reflect lost task alignment rather than lost knowledge, caused by early optimisation steps, and report that freezing the bottom layers helps. Kotha, Springer and Raghunathan (2024) show that capabilities apparently lost to fine-tuning, including in-context learning lost to instruction tuning, can be partly recovered by translating the prompt into another language, which suggests that what was lost was the model's inference about which task it was being asked to do, not the capability itself.

On the weight-space side, Alexandrov et al. (2024) report for Llama 3 8B adapted to Bulgarian that both learning, measured by Bulgarian benchmarks, and

forgetting, measured by English negative log-likelihood, "strongly correlated with weight change magnitude". Their branch-and-merge procedure, which trains on data slices in parallel and averages the results, achieved more learning per unit of weight change. Halving the learning rate gave a similar weight change to their method and reduced forgetting slightly less, at the cost of "severely reduced learning". Biderman et al. (2024) found that LoRA learns less and forgets less than full fine-tuning in continued pretraining on code and mathematics, and that full fine-tuning learns perturbations of 10 to 100 times higher rank than typical LoRA configurations. Elhady et al. found the same trade-off for Basque: LoRA without English preserved Copain at 39.61 but left downstream accuracy at 28.03, barely above the base model's 27.43. They suggest this also explains why Yong et al. (2023), adding languages to BLOOM without an English share, found adapter-based fine-tuning more effective than continued pretraining for the larger models. Mukhoti et al. (2024) report a related effect for pretrained foundation models, which they call concept forgetting: end-to-end fine-tuning reduces the model's ability to recognise concepts on tasks other than the fine-tuning one, and a method that penalises distance from the pretrained features reduces the loss. Weight averaging as a guard against forgetting has a longer history: WiSE-FT interpolates fine-tuned and zero-shot weights to retain robustness (Wortsman et al., 2022), and WARP uses an exponential moving average as a moving anchor during reinforcement learning from human feedback (Rame et al., 2024). Morales-Brotons, Vogels and Hendrikx (2024) document the good early performance of EMA, which is the property the damper account relies on.

The damper account predicts that (a) the English share matters most early and can be withdrawn after a short window; (b) any intervention that limits early drift, whether weight averaging, a smaller update rank or a gentler schedule, can substitute for English if it does not also block learning; (c) the benefit should shrink when the base model is already competent in the target language, because the switch is then less abrupt; and (d) the composition of the source share matters to the extent that it resembles the distribution whose capabilities are being protected.

6.2 The knowledge account

On this account, the source share matters throughout training because it supplies content. A target corpus of a few billion tokens, or a few hundred million, contains far less factual and procedural knowledge than the English web. Mixing English in lets the model learn things it could not learn from the target text, and cross-lingual transfer lets it use them in the target language. Repetition of a small target corpus also overfits, and a second language diversifies the signal.

Jeha et al. (2026) state this account for regime B and support it with the scale dependence of the benefit: the benchmark multiplier grows steeply with model size, as capacity becomes available to absorb the extra knowledge, and monolingual accuracy stays near chance however the hyperparameters are tuned. The large English-benchmark gains of their bilingual models, 47 percent on English ARC Easy at 1.43 billion parameters when checkpoints are selected by Arabic loss, against 28 percent for monolingual training, show that the English content is being learned, not merely used as a regulariser. Seto et al. (2024) add that better English data, in the sense of more information-rich filtering, transfers to the target language, which is what the knowledge account predicts and the damper account does not. Etxaniz et al. (2024) make a related observation for regime A when they note that the capabilities an adapted model shows in a language appear not to be determined by its linguistic competence in that language, so that stronger English models may yield stronger low-resource models.

The knowledge account predicts that (a) withdrawing English after a short window should cost something, increasingly so as training continues; (b) a regulariser that does not add content, such as weight decay or weight averaging, cannot substitute for English, a prediction Jeha et al. confirm for weight decay in regime B, where data mixing beat aggressive hyperparameter tuning at every scale from 380 million parameters up; (c) the gain should be larger on benchmarks that test knowledge, and larger on translated benchmarks whose content originates in English than on native ones; and (d) the source share's quality, in the sense of its information content, should matter more than its resemblance to the base model's original distribution.

6.3 The pivot account

On a third account, the source share matters because the model's internal processing of other languages is routed through a representation organised around its dominant language. Wendler et al. (2024) found that in the Llama 2 family, intermediate representations for non-English prompts pass through a region from which the English version of the answer is more decodable than the target-language version, before the final layers map it into the target language. Wu et al. (2024) report that representations of semantically equivalent inputs in different languages and modalities converge in intermediate layers and can be read through the dominant pretraining language. Tang et al. (2024) attribute a model's proficiency in a given language mainly to a small subset of language-specific neurons, situated chiefly in the top and bottom layers, and show that switching those neurons on or off can steer the output language. If adaptation to a new language disturbs the shared middle while training only the target language, the target-language interface to everything the model knows may be damaged even if the target

language itself is learned.

An older line of work makes the pivot account more plausible by showing how much of a model's machinery can be reused across languages unchanged. Artetxe, Ruder and Yogatama (2020) transferred a monolingual English masked language model to new languages by learning only new lexical embeddings, with the transformer body frozen, and the transferred model held its own against multilingual BERT on cross-lingual classification and on XQuAD, a question answering set they released. De Vries and Nissim (2021) did the same for GPT-2 in Italian and Dutch. Chen et al. (2023) found that periodically resetting the embedding layer during pretraining produces models that converge faster when adapted to a new language and do better in low-data adaptation, particularly for languages distant from English. If the body of a model can serve a new language through new embeddings alone, then a full-parameter update on target text alone is doing more than it needs to, and what it does beyond the interface is a candidate for the damage.

The pivot account predicts that (a) the English share protects capabilities that are language-independent, such as in-context pattern inference, which is what Copain measures; (b) the share could be replaced by another high-resource language only if that language shares the pivot, a test Elhady et al. list as future work; and (c) interventions that confine updates to the outer layers, where language-specific components concentrate, should preserve the middle. The last prediction has partial support from Kunz (2025), who reports that the trouble some adapted Icelandic models had with longer contexts "can be mitigated by adapting only the final layers", and from layer-swapping work in which Bandarkar et al. (2025) compose a language expert and a mathematics expert by exchanging the top and bottom transformer layers.

6.4 What the evidence currently favours

For regime A, the evidence favours the damper account as the main mechanism, with two qualifications. The curriculum and weight-averaging results are its strongest support: if English were mainly supplying content, removing it after the first 1,000 of 10,000 steps should have cost more than it did, and weight averaging, which adds no content, should not have matched it on downstream accuracy. In the Basque, Arabic and Indonesian results, the curriculum and weight averaging each fell short of full English on downstream accuracy by at most about 1.7 points, and in some cases exceeded it. The first qualification is that the downstream benchmarks in regime A are native target-language tests of proficiency, trivia, exams and reading, on which an English-centric base model already holds most of the relevant world knowledge; the knowledge account's distinctive gains would not be expected to show there. The second is that the pivot account is not separated

from the damper account by any experiment in the literature the author located. Both predict early, language-independent damage that a short English window prevents.

For regime B, the evidence favours the knowledge account, again with a qualification. A model trained from scratch has no prior capability to protect, so the damper account has nothing to explain there, and the benchmark gains on English show content being acquired. The qualification is the translated benchmark, which tilts the measurement towards knowledge originating in English.

The accounts therefore divide the territory more than they compete for it. The practical consequence is that the right English schedule is regime-dependent: a short early window may suffice when adapting a strong English-centric base to a language it barely knows, and a sustained share is warranted when the target corpus is too small to carry the knowledge the model needs. Either way, the instrument that chooses the schedule cannot be target validation loss alone.

7. What the anchor's own tables complicate

The headline of Elhady et al. (2025) is accurate, and the authors themselves state several of the qualifications below in their conclusion. They are collected here because the shorthand version of the result drops them, and each one narrows where the result applies.

English is not sufficient to protect in-context learning. In Basque with Llama 2 7B, the English arm kept Copain at 43.43 against a base of 44.67, 97 percent of the base model's score. In Arabic and Indonesian the base model scored 44.67 on Copain and the English arms fell to 32.60 and 30.79, 73 and 69 percent of base. The arms without English fell further, to 23.80 and 27.34. The authors note this directly: including English is "not a universal solution", as shown by the drop in Copain for Arabic and Indonesian. On the damper account this means a 20 percent share was not enough to damp the transition for those languages, or that something other than the transition was damaging in-context learning there.

A regulariser beat English on the probe English was supposed to protect. In Arabic, weight averaging without English kept Copain at 42.76, against 32.60 for the English arm, a difference of about 10 points. In Basque with Llama 2 13B the ordering reversed, 42.99 against 47.80. The authors report that the averaging interval had to be tuned per language. On the damper account the Arabic result is unsurprising: if what matters is limiting drift, a direct limit on drift can do better than an indirect one. On the knowledge account it is hard to explain.

The benefit shrinks as the base model's competence in the target language grows. For Llama 3.1 8B and

Gemma 2 9B, whose base validation perplexities on Basque were 2.18 and 2.28 against 23.64 for Llama 2 7B, the downstream differences between arms were 0.91 and 3.44 points and the Copain differences 0.85 and 6.64 points. The authors attribute this to a smaller distribution shift for base models already able to model the target language, which is the damper account's prediction (c) and is the opposite of what one would expect if English were supplying knowledge that a stronger English base already has.

Lowering the learning rate did not substitute for English. In preliminary runs without English, lowering the peak learning rate from $1e-4$ to $5e-5$ and $1e-5$ raised final Copain only from 20.12 to 26.43 and 29.43, while Basque perplexity worsened sharply, from 3.58 to 8.76 and 8.29. A uniform reduction of step size limited learning far more than it protected in-context learning. Alexandrov et al. (2024) found the same for Bulgarian. This complicates a naive reading of the damper account: it is not the overall size of updates that matters but where they go, which is what a curriculum, a moving average or a merged model changes and a smaller learning rate does not.

The generative evaluation sharpens the picture. On MGSM-eu, a human-translated mathematics benchmark used with chain-of-thought, the base model scored 2.40, the English arm 9.60, the arm without English 1.86 and weight averaging without English 5.60. Continued pretraining on Basque alone left the model slightly below its starting point on this task. These are small numbers on a hard task and the paper reports them in an appendix, but they are the only generative result in the study, and they go the same way as the multiple-choice results.

A reporting discrepancy, and its resolution. For Llama 2 13B on Basque with weight averaging, Table 4 of Elhady et al. gives an average downstream accuracy of 41.39, while the detailed Table 8 gives an average of 40.39 beside four component scores, 40.80, 40.30, 32.67 and 51.77. The four components average to 41.385, so Table 4's figure is correct and the Table 8 cell appears to be a typographical error. The two tables also print the same row for the weight-averaging variants with and without curriculum for this model. This paper uses the Table 4 value.

The tables report single values. No variance across seeds is reported for the main comparisons. Differences of a few points between arms on benchmarks ranging from 352 items (EusReading) to 16,774 (EusExams), and 1,050 for Copain, carry some sampling uncertainty, as does run-to-run training variation, and the smaller differences in the table, under one point for Llama 3.1, should not be read as effects without replication. The large differences, 18 to 23 points on Copain for Llama 2 in Basque, are well beyond what item sampling alone would produce.

Both arms repeat the target corpus. The arithmetic below is the author's, from the stated hyperparameters, and assumes every sequence is packed to its full 4,096 tokens. Ten thousand steps of 256 sequences of 4,096 tokens give about 10.5 billion token positions. Against a target corpus of 4.5 to 4.7 billion tokens, that is roughly 2.2 passes for the target-only arm and roughly 1.8 for the arm with a 20 percent English share. Muennighoff et al. (2023) found that up to about four passes over repeated data produced negligible change in loss relative to unique data. Regime A as studied is therefore not data-constrained in the sense of regime B, which is one reason the two regimes can disagree about what English buys without either being wrong.

8. Why a loss on prompts saw what a loss on corpus text missed

Section 3 showed that a loss on target-language few-shot prompts registered the damage early, while a loss on target corpus text did not register it at all. This section asks why, because the answer determines what kind of instrument a mixture law would need.

The general relation between loss and downstream accuracy has been studied at length, with results on both sides. Xia et al. (2023) found, across OPT checkpoints from 125 million to 175 billion parameters, that perplexity was a strong predictor of in-context performance on 74 BIG-Bench multiple-choice tasks regardless of model size. Gadre et al. (2024) fitted a power law from perplexity to average top-1 error across downstream tasks, over 104 models on three data distributions, and used it to predict the error of larger runs. Du et al. (2024) found thresholds in pretraining loss below which particular abilities appear. Against this, Lourie, Hu and Cho (2025) conducted a meta-analysis of published downstream scaling data and found predictable scaling in 39 percent of cases, with seemingly benign changes to the setting able to change the behaviour completely. Isik et al. (2025) found that when pretraining and downstream data are misaligned, translation quality can fluctuate or worsen with more pretraining while downstream cross-entropy keeps improving. Tay et al. (2022) showed that model shape, not only size, matters for downstream fine-tuning, so that upstream loss does not fix downstream quality. Liu et al. (2022) constructed models with identical pretraining loss and different downstream transfer. Wei et al. (2022) and Schaeffer, Miranda and Koyejo (2023) disagreed about whether emergent abilities are a property of models or of thresholded metrics; Hu et al. (2024) showed that sufficiently fine-grained evaluation reveals steady task improvements in small models that coarse evaluation misses, while also identifying a kind of accelerated emergence that standard scaling functions do not fit.

Two strands of this literature explain the Basque observation directly.

The first concerns what a multiple-choice score depends on. Schaeffer et al. (2024) trace the unpredictability of downstream accuracy to the sequence of transformations from per-token log-likelihood to accuracy. Their central finding is that accuracy depends on how the right option's probability compares with that of the few particular wrong options offered, so that predicting it requires knowing how probability mass moves on the incorrect options, not only on the correct one. That is precisely the quantity Elhady et al. plotted in their Figure 3: with English, the perplexity of incorrect labels rose while that of correct labels held, widening the margin; without English, both rose and the margin stayed narrow. A loss on corpus text averages over every token of running prose and contains no term for the margin between a correct and an incorrect label in a few-shot prompt. A loss on the labels does.

The second concerns in-context learning specifically. Olsson et al. (2022) found that the sudden appearance of in-context learning in transformer language models coincides with the formation of induction heads and is visible as a bump in the training loss. Chan et al. (2022) found that in-context learning emerges when training data has particular distributional properties, burstiness and many rare classes among them. Singh et al. (2023) found that in-context learning can emerge and then disappear in favour of in-weights learning while the training loss continues to decrease. Hernandez et al. (2022) found that repeating a small fraction of data damages induction heads and copying. Taken together, these results suggest that in-context ability is a property of a model's mechanisms that can be built or lost while aggregate next-token loss on the training distribution moves smoothly or not at all. A target-language corpus of running text contains few sequences in which a pattern must be inferred from demonstrations and applied to a new instance; its held-out loss has little weight on the behaviour that Copain probes.

Two recent lines of work suggest that the missing instrument need not be expensive. Brandfonbrener et al. (2024) show that a loss on one distribution can be predicted from a loss on another, including from pretraining data to downstream task data, with fits that extrapolate to 20 times the largest compute budget used to fit them. Snell et al. (2024) show that fine-tuning a model on a task shifts the point at which ability on that task emerges towards less capable models, so that emergence can be forecast from small models by fitting how the shift depends on fine-tuning data. Ruan, Maddison and Hashimoto (2024) build scaling laws from about 100 public models by mapping benchmark results into a low-dimensional capability space. None of these was applied to mixture selection for language adaptation, but each is a way of adding a distribution-of-use term to a law without training a full model per candidate mixture.

Two cautions follow. First, the Basque label perplexity was measured on the same benchmarks whose accuracy it was used to explain, so its early movement is evidence of sensitivity, not of independent predictive validity; a mixture law fitted on it would need a held-out set of prompts. Second, a loss on prompts is only as broad as the prompts. A probe built around few-shot pattern inference will see damage to in-context learning and may miss damage to, say, long-context use or instruction following. Kunz (2025), adapting a model to Icelandic with parameter-efficient methods, found that some adapted models struggled with longer contexts, a capability no short-prompt probe would see.

9. Would a loss on English text have seen it?

The mixture laws that include a source term would register the Basque damage if it showed up in English validation loss. Elhady et al. (2025) do not report English validation loss for either arm. This is the most important unmeasured quantity in the evidence base, and the literature points both ways.

For the view that it would have registered: Alexandrov et al. (2024), in their Bulgarian adaptation runs with a minimal replay of English web text, observed that target-language benchmark performance first rose and then fell as English negative log-likelihood rose, and that a curated replay mix held English likelihood down and target performance up. In that setting the source-language loss and the target-language downstream decline moved together. Ibrahim et al. (2024) found that even 1 percent replay substantially reduced the rise in Pile validation loss under an English to German shift, which shows that source loss is a sensitive detector of the general forgetting a source share prevents.

For the view that it would not: the Basque damage was concentrated in a format-dependent, language-independent ability, and several results suggest such abilities can be lost while aggregate likelihood on running text is little affected. Singh et al. (2023) observed in-context learning vanishing during a phase in which training loss fell. Zheng et al. (2025) and Kotha et al. (2024) argue that much apparent forgetting is a loss of task alignment or task inference, recoverable by the right prompt, rather than a loss of knowledge, which is what source-text loss primarily measures. Wang et al. (2024) attribute lost in-context performance to format specialisation, a change in how the model responds to a prompt's format, not in what it knows. If the run without English lost Copain mainly through such a change, its English text loss might have risen only slightly.

The composition of the replay mix bears on this. In Alexandrov et al. (2024), the curated approximate-replay mix contained instruction-tuning data, about 15 percent by sampling probability, alongside web text, Wikipedia and code. Moving Mistral 7B's Bulgarian adaptation from minimal to approximate replay raised average Bulgarian

benchmark scores from 43.71 to 52.16 and English scores from 51.44 to 62.03. The authors note that because their replay mix contained instruction examples, it could raise English benchmark scores above the base model's, and they therefore used held-out English document likelihood as their measure of forgetting. A replay mix rich in prompt-and-response structure protecting target-language performance is what the damper account, read as format protection, would predict, but the comparison changes the information content of the replay at the same time, so the knowledge account predicts it too.

The question has a direct empirical answer that nobody has reported: rerun the Basque arms and log English validation loss alongside Copain and label perplexity at the same early checkpoints. If English loss rises with the Copain collapse, the two-term laws already contain an instrument that can price the damage, and the gap identified in this paper concerns only the target-only laws. If it does not, every published mixture law for language adaptation is blind to the effect, and a third term is needed.

10. The evidence in two tables and a figure

The measured values the argument rests on are collected in Table 1, one source result per row. Table 2 sets the three accounts of Section 6 against the observations that separate them. Figure 1 re-plots the anchor study's own tables to show, in panel (a), that the arms with and without English differ little in validation perplexity and a great deal in downstream accuracy and in-context learning, and in panel (b), which interventions preserved in-context learning in which language.

Panel (a) of Figure 1 makes the support point of Section 3 visible. The perplexities printed under each pair differ by at most 0.23 and change sign across pairs, while every bar is positive. Panel (b) makes the complication of Section 7 visible. In Basque, English throughout and English early only keep the probe at 94 to 97 percent of the base model's score, and weight averaging at 87 to 96 percent. In Indonesian, none of the four arms rises above 75 percent. In Arabic, English throughout keeps only 73 percent, while the curriculum keeps 89 percent and weight averaging 96 percent.

11. The decision structure

The analysis reduces to a rule about instruments: use the loss whose distribution contains the effect being priced. Algorithm 1 states it as a procedure for choosing and validating a source share. Its first branch is the distinction of Sections 3 and 5, its middle branches are the three accounts of Section 6, and its last branch is the uncovered case, in which no published instrument has been shown to see the effect.

Step 2 is the part of the procedure the evidence settles: across-arm selection by target loss alone failed in both

Table 1. Measured values used in the argument, one source result per row. "+En" is the arm with an English share and "none" the arm without; for Elhady et al. (2025) PPL is target validation perplexity, "Dwn" is average native-benchmark accuracy and "Cop" is Copain accuracy, all at the end of training, with the base model's Copain score of 44.67 for Llama 2 7B. Values are as reported by each source; ratios in the Jeha et al. rows are the source's own.

Source	Setting	Arms compared	Loss instrument	Downstream or ICL
Elhady et al. 2025	Llama 2 7B, Basque	+ En vs none	PPL 3.35 vs 3.58	Dwn 34.14 vs 28.89; Cop 43.43 vs 20.12
Elhady et al. 2025	Llama 2 13B, Basque	+ En vs none	PPL 2.82 vs 2.79	Dwn 42.52 vs 35.20; Cop 47.80 vs 29.43
Elhady et al. 2025	Llama 3.1 8B, Basque	+ En vs none	PPL 1.73 vs 1.82	Dwn 55.75 vs 54.84; Cop 42.04 vs 41.19
Elhady et al. 2025	Gemma 2 9B, Basque	+ En vs none	PPL 1.52 vs 1.48	Dwn 49.39 vs 45.95; Cop 50.23 vs 43.59
Elhady et al. 2025	Llama 2 7B, Arabic	+ En vs none	PPL 2.09 vs 2.12	Dwn 34.34 vs 32.67; Cop 32.60 vs 23.80
Elhady et al. 2025	Llama 2 7B, Indonesian	+ En vs none	PPL 3.25 vs 3.05	Dwn 30.79 vs 26.92; Cop 30.79 vs 27.34
Elhady et al. 2025	Llama 2 7B, Basque	En in first 10% only	PPL 3.08	Dwn 35.12; Cop 42.94
Elhady et al. 2025	Llama 2 7B, Arabic	EMA, no English	PPL 2.03	Dwn 33.36; Cop 42.76
Elhady et al. 2025	Llama 2 7B, Basque	none, LR 5e-5 / 1e-5	PPL 8.76 / 8.29	Dwn 27.98 / 27.42; Cop 26.43 / 29.43
Elhady et al. 2025	Llama 2 7B, Basque	LoRA, no English	PPL 3.68	Dwn 28.03; Cop 39.61
Elhady et al. 2025	Llama 2 7B, Basque	Drift, none vs + En	L2 distance 7x at step 100, 15x at 1k	Label PPL spikes in first steps
Etzaniz et al. 2024	Latxa 7B, early version	none vs + En	not reported	Belebele 23.67, 13 points lower
Alexandrov et al. 2024	Mistral 7B, Bulgarian	minimal vs curated replay	English NLL used for forgetting	BG 43.71 vs 52.16; EN 51.44 vs 62.03
Ibrahim et al. 2024	405M, English to German	0 to 50% replay	1% replay cuts Pile loss rise	German evaluations near chance
Jeha et al. 2026	1.43B from scratch, Arabic	bilingual vs monolingual	Mixing 12.9% better, tuning 3.1%	ARC-E ar 41.4 vs 33.6; en 57.7 vs 29.6 (peak accuracy; mono saw 18.4B Arabic tokens, bilingual 9.2B)
Jeha et al. 2026	380M to 1.43B, Arabic	data multiplier of mixing	2 to 3x on loss	2 to 13x on benchmark
Jeha et al. 2026	Bilingual runs	best-loss vs peak checkpoint	Within-run selection by loss	Median gap 0.21 / 0.13 points

Table 2. Three accounts of what the English share supplies, with the observation each predicts. "Fits" and "strains" record whether the reported result matches the prediction; the sources are those in Table 1 and Section 6.

Account	English supplies	English only early	Regulariser, no English	Stronger base model	Best support
Damper	A gentle transition; limits early drift	Should suffice: fits (eu, ar, id)	Should substitute: fits downstream	Less benefit: fits (L3.1, G2)	Elhady 2025; Alexandrov 2024
Knowledge	Content and diversity absent from target text	Should cost: strains in regime A	Should not substitute: fits regime B	Same or more benefit: strains	Jeha 2026; Seto 2024
Pivot	Keeps English-routed middle layers in use	Should suffice if damage is early	Partial at most	Less benefit if target already routed	Wendler 2024; Wu 2024; Tang 2024

regimes. Step 1 is supported directly only in regime B, and is hedged for regime A by the requirement to log a prompt-distribution loss, because the anchor's run without English shows target loss missing an early collapse within a single run. Step 4 encodes the damper reading and the open question of Section 9 as a branch, so that whichever way the English-loss measurement comes out, the procedure says what follows. Step 5 encodes the knowledge reading with the translated-benchmark caution. Step 6 exists because the literature contains no demonstration that any loss prices every effect of a source share, and a procedure that pretended otherwise would hide the gap this paper is about.

12. Experiments that would settle the open parts

The open questions are specific enough to be closed by experiments of modest size. Each is stated with the outcome that would favour each reading.

E1. Log the source loss in the anchor setting. Repeat the Basque arms with Llama 2 7B and record English validation loss at the checkpoints where Copain and label perplexity were recorded. If English loss rises in step with the Copain collapse in the arm without English and not in the arm with it, two-term laws such as D-CPT and CMR already contain the needed instrument. If

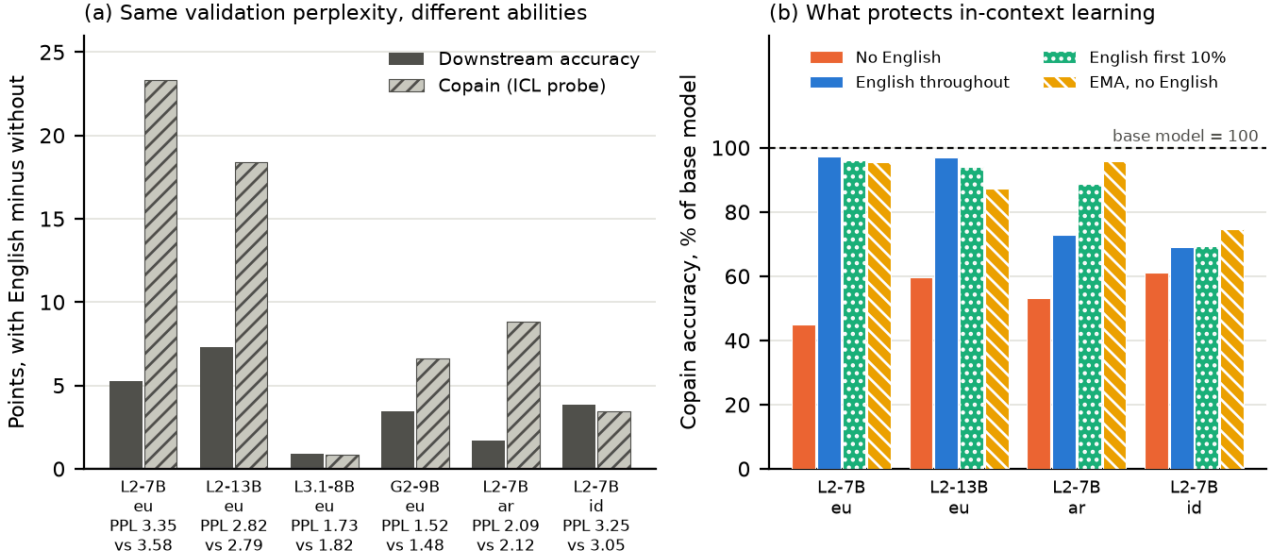


Figure 1. (a) Difference between the continued-pretraining arms with and without a 20 percent English share, in points, for downstream accuracy (dark) and the Copain in-context learning probe (hatched), for six base-model and language pairs; the target validation perplexity of each arm is printed under the pair (with English vs without). (b) Copain accuracy as a percentage of the base model's score for four Llama 2 runs under four arms: no English, English throughout, English in the first 10 percent of steps only, and weight averaging (EMA) without English. All values from Elhady, Agirre and Artetxe (2025), Tables 2, 3 and 4; differences and percentages computed by the author from those tables.

Algorithm 1. Choosing and validating a source-language share for language adaptation

```

Input: base model B, target corpus T, source corpus S, candidate shares R, schedules K
Instruments: TVL (held-out T text), SVL (held-out S text),
            PDL (target-language few-shot prompts: format probe + answer-label margin),
            DA (native target-language benchmarks; translated ones reported separately)

Step 1: If the comparison is within one arm (checkpoints, seeds, step size at a fixed share):
    TVL is admissible for checkpoint choice (Jeha 2026); in regime A also log PDL
Step 2: Otherwise the comparison is across arms (different shares or schedules):
    TVL alone is inadmissible. # Elhady 2025; Jeha 2026
Step 3: Regime check:
    if B is English-centric and T is seen a few times -> regime A (forgetting possible)
    if B is trained from scratch and T is repeated many times -> regime B
Step 4: Regime A: log PDL and SVL at early steps (e.g. 10, 100, 1000) for share 0 and r > 0.
    if PDL collapses early at share 0 and not at r:
        damper reading; test K = {r early only} and {EMA, share 0}
        if either matches r throughout on DA: prefer the cheaper schedule
    if SVL rises with the PDL collapse: a two-term law (target + source loss) can price it
    else: SVL is blind here; add a PDL term to any fitted law
Step 5: Regime B: compare arms on DA, never on TVL.
    if DA gain grows with model size and persists on native benchmarks:
        knowledge reading; keep the share throughout training
    if gain appears only on translated benchmarks: report it as unconfirmed
Step 6: If no instrument separates the arms but DA differs: the effect is unpriced;
    report the share as chosen by DA directly and say so.
Output: share, schedule, and the instrument that justified them

```

English loss moves little, the damage is invisible to every loss term currently fitted, and a prompt-distribution term is required. This is the cheapest experiment in the list and the one with the largest consequence for practice.

E2. Fit a mixture law on a held-out prompt-distribution loss. Build a held-out set of target-language few-shot prompts disjoint from the evaluation benchmarks, compute the margin between correct and incorrect answer labels, and fit a mixture law to it alongside target and source validation loss. Compare the share it selects with the share selected by target loss

alone and by downstream accuracy. If the prompt-loss law selects the downstream-optimal share and the target-loss law does not, the instrument gap is closed for this effect.

E3. Cross the curriculum with the regime. Run the English-first-10-percent curriculum in regime B, from scratch on a data-constrained target, at the scales Jeha et al. used. The damper account predicts that the curriculum should do little there, because nothing is being protected; the knowledge account predicts it should forfeit most of the mixing benefit. Conversely,

run a sustained English share in regime A for much longer than 10,000 steps on a small target corpus, so that the target text is repeated many times. If English's value then grows over training, the knowledge channel is present in regime A too, and the curriculum result reflects the short horizon of the anchor study.

E4. Substitute format for language. In regime A, replace the English share with target-language text arranged as few-shot demonstrations, for example synthetic pattern-inference sequences or target-language question-answer pairs, at the same token share. If this protects Copain and downstream accuracy as well as English does, the damper account in its format-protection form is supported and the language of the source share is incidental. If it does not, either the pivot account or the knowledge account carries more weight than the current evidence assigns them. The replay-composition result of Alexandrov et al. (2024) suggests the substitution may work, but confounds format with information content.

E5. Substitute language for English. Use a high-resource language other than English, close to the target or distant from it, as the source share. Elhady et al. propose this as future work. The pivot account predicts that a language the base model does not route through should help less than English; the damper account predicts that any distribution close to the base model's original one should help; the knowledge account predicts that information content should matter more than language.

E6. Native versus translated evaluation for the knowledge account. Repeat the regime B comparison of Jeha et al. with a natively authored target-language knowledge benchmark alongside the translated one. If the bilingual advantage on the native benchmark is a large fraction of that on the translated one, the knowledge account's support does not depend on translation artefacts. If it shrinks sharply, part of the benchmark multiplier was measuring English-shaped content.

E7. Seeds. Replicate the smaller differences in Table 1, the 0.91 point downstream difference for Llama 3.1 and the 1.67 point difference for Arabic, across several seeds. Without this, the claim that the English effect shrinks with base-model competence rests partly on single runs.

13. The strongest case against this paper

"This is a restatement of what Elhady et al. already said." The anchor's authors do frame their result in terms of the training distribution versus out-of-distribution prompts, and they report the label perplexity that this paper treats as the key instrument. The contribution here is not the observation but its consequences: that the result is outside the scope of the claim it is said to contradict; that the same failure to rank mixtures appears, independently, in a regime where

forgetting is impossible; that the mixture laws now used for language adaptation are fitted on exactly the instrument that missed the effect; and that the three accounts of what English buys are separable by experiments that have not been run. A reader who already held all four of those conclusions will find little new here. The author did not find them stated together in the literature reviewed.

"Loss-based mixture laws work in practice; DoReMi and RegMix improve downstream accuracy." They do, in the settings they were validated on, which are mixtures of English domains for pretraining from scratch. The argument here is not that loss-based mixture selection fails in general. It is that a target term fitted on held-out target text cannot see effects off that distribution, and that in language adaptation the main effect of the source share is off that distribution. Where a method includes a source term, the question becomes the empirical one of Section 9.

"Jeha et al. measured a different thing; their result concerns knowledge, not forgetting." That is this paper's point in Section 6. The two regimes disagree about what English buys and agree about what target loss can see. The convergence on the instrument does not depend on agreement about the mechanism.

"The translated benchmark undermines the regime B evidence." It weakens the size of the benchmark multiplier and leaves its direction. Monolingual Arabic training stayed within a few points of chance on that benchmark however it was tuned, while bilingual training rose steadily with scale; for that gap to be wholly an artefact, a monolingual Arabic model would have to have learned the relevant knowledge and be unable to show it on a translated test, which is possible but would itself be a finding. E6 tests it.

"The damper account is just catastrophic forgetting with a new name." In part it is, and the paper uses the continual learning literature as its support. The distinction that matters is between forgetting that a source-text loss registers, which is what replay is usually tuned against, and forgetting of a format-dependent capability that it may not register. If E1 finds that English loss tracks the Copain collapse, the distinction collapses and the account is ordinary forgetting. That is a live possibility, and the paper does not claim otherwise.

"Small differences and single runs." Conceded for the smaller rows, which is why E7 is listed. The large rows, around 20 points on Copain and 5 to 7 points on downstream accuracy for Llama 2 in Basque, and the 13 point Belebele gap reported for Latxa, are not plausibly sampling noise.

14. What is not known

Whether English validation loss rose when in-context learning collapsed in the Basque run without English is not known; it was not reported. Whether a mixture law fitted on a prompt-distribution loss would select the downstream-optimal English share is not known; no such law has been fitted. Whether the curriculum result survives longer training or heavier repetition of the target corpus is not known. Whether a source share in a language other than English, or target-language text in few-shot format, can replace English is not known. Why English failed to protect in-context learning in Arabic and Indonesian when it protected it in Basque is not known; differences in script, corpus source and base-model competence are all candidates and none has been isolated. How much of the regime B benchmark multiplier survives a native benchmark is not known. Whether the pivot account is distinguishable from the damper account by any experiment yet performed is not known; the author located none that separates them.

15. Limitations

The analysis rests heavily on one study for regime A and one for regime B. Elhady et al. (2025) cover three target languages, four base models for one of them, and 10,000 steps of training; Jeha et al. (2026) cover one target language, four model sizes up to 1.43 billion parameters, and a translated benchmark. Corroboration from Latxa, branch-and-merge, Seto et al. and the continual learning literature is consistent but less direct. A larger or more varied body of evidence could shift the balance among the accounts in Section 6.

The four-instrument taxonomy of Section 2 is the author's. Other divisions are possible, for example separating losses on instruction-formatted text from losses on few-shot prompts, and a finer taxonomy might show that some instrument now classed as a prompt-distribution loss behaves like corpus loss.

The comparison of mixture laws in Section 4 is by method description. The claim that a target-only law would choose differently from downstream accuracy follows from the reported perplexities and accuracies; whether a fitted law does so is open, and experiment E2 would settle it.

The token arithmetic in Section 7 assumes full packing of sequences, which the anchor study does not state. If sequences were shorter on average, the number of passes over the target corpus was lower, which strengthens rather than weakens the contrast with regime B.

The paper says nothing about vocabulary extension, which interacts with the source share in ways the anchor study held fixed by design (Fujii et al., 2024; Csaki et al., 2024), or about the curse of multilinguality in massively multilingual adaptation (Chang et al., 2023; Longpre et al., 2025).

16. Conclusion

Adding English to a continued-pretraining mixture for a new language changed what the adapted model could do and left its loss on held-out target text where it was. The result invites the reading that perplexity fails to predict emergent abilities. The evidence reviewed here supports a different one. A loss on held-out text from the target corpus registers what is expressed in that text and nothing else. It cannot rank mixtures, because what the English share changes is not expressed there, and within a run it is reliable only where the abilities it does not measure move with the ones it does. A loss on target-language few-shot prompts registered the change early, and an independent study of bilingual pretraining from scratch found the same failure to rank mixtures where forgetting was impossible. The two regimes differ about what English supplies: a damper on early drift that protects in-context learning when adapting an English-centric model, and knowledge and diversity when training from scratch on scarce target text. The mixture laws now used to set the English share are fitted on the instrument that missed the effect, or on a source-language loss whose sensitivity to it has never been measured. Until it is, the share should be chosen with an instrument whose distribution contains the effect being priced, and the paper names the experiment that would show whether any existing loss can do that job.

References

- Alexandrov, A., Raychev, V., Mueller, M. N., Zhang, C., Vechev, M., & Toutanova, K. (2024). Mitigating Catastrophic Forgetting in Language Transfer via Model Merging. arXiv:2407.08699. doi:10.48550/arXiv.2407.08699
- Artetxe, M., Ruder, S., & Yogatama, D. (2020). On the Cross-lingual Transferability of Monolingual Representations. ACL 2020. arXiv:1910.11856. doi:10.48550/arXiv.1910.11856
- Bandarkar, L., Liang, D., Muller, B., Artetxe, M., Shukla, S. N., Husa, D., Goyal, N., Krishnan, A., Zettlemoyer, L., & Khabsa, M. (2024). The Belebele Benchmark: a Parallel Reading Comprehension Dataset in 122 Language Variants. ACL 2024. arXiv:2308.16884. doi:10.48550/arXiv.2308.16884
- Bandarkar, L., Muller, B., Yuvraj, P., Hou, R., Singhal, N., Lv, H., & Liu, B. (2025). Layer Swapping for Zero-Shot Cross-Lingual Transfer in Large Language Models. ICLR 2025. arXiv:2410.01335. doi:10.48550/arXiv.2410.01335
- Biderman, D., Portes, J., Gonzalez Ortiz, J. J., Paul, M., Greengard, P., Jennings, C., King, D., Havens, S., Chiley, V., Frankle, J., Blakeney, C., & Cunningham, J. P. (2024). LoRA Learns Less and Forgets Less. TMLR. arXiv:2405.09673. doi:10.48550/arXiv.2405.09673
- Brandfonbrener, D., Anand, N., Vyas, N., Malach, E., & Kakade, S. (2024). Loss-to-Loss Prediction: Scaling Laws for All Datasets. arXiv:2411.12925. doi:10.48550/arXiv.2411.12925
- Caccia, L., Aljundi, R., Asadi, N., Tuytelaars, T., Pineau, J., & Belilovsky, E. (2022). New Insights on Reducing Abrupt Representation Change in Online Continual Learning. ICLR 2022. arXiv:2104.05025. doi:10.48550/arXiv.2104.05025
- Chan, S. C. Y., Santoro, A., Lampinen, A. K., Wang, J. X., Singh, A., Richemond, P. H., McClelland, J., & Hill, F. (2022). Data Distributional Properties Drive Emergent In-Context Learning in Transformers. NeurIPS 2022. arXiv:2205.05055. doi:10.48550/arXiv.2205.05055

- Chang, T. A., Arnett, C., Tu, Z., & Bergen, B. K. (2023). When Is Multilinguality a Curse? Language Modeling for 250 High- and Low-Resource Languages. *arXiv:2311.09205*. doi:10.48550/arXiv.2311.09205
- Chen, Y., Marchisio, K., Raileanu, R., Adelani, D. I., Stenetorp, P., Riedel, S., & Artetxe, M. (2023). Improving Language Plasticity via Pretraining with Active Forgetting. *NeurIPS 2023*. *arXiv:2307.01163*. doi:10.48550/arXiv.2307.01163
- Csaki, Z., Pawakapan, P., Thakker, U., & Xu, Q. (2023). Efficiently Adapting Pretrained Language Models To New Languages. *arXiv:2311.05741*. doi:10.48550/arXiv.2311.05741
- Csaki, Z., Li, B., Li, J., Xu, Q., Pawakapan, P., Zhang, L., Du, Y., Zhao, H., Hu, C., & Thakker, U. (2024). SambaLingo: Teaching Large Language Models New Languages. *arXiv:2404.05829*. doi:10.48550/arXiv.2404.05829
- De Lange, M., van de Ven, G., & Tuytelaars, T. (2023). Continual evaluation for lifelong learning: Identifying the stability gap. *ICLR 2023*. *arXiv:2205.13452*. doi:10.48550/arXiv.2205.13452
- de Vries, W., & Nissim, M. (2021). As Good as New: How to Successfully Recycle English GPT-2 to Make Models for Other Languages. *Findings of ACL-IJCNLP 2021*. *arXiv:2012.05628*. doi:10.48550/arXiv.2012.05628
- Dou, L., Liu, Q., Zeng, G., Guo, J., Zhou, J., Lu, W., & Lin, M. (2024). Sailor: Open Language Models for South-East Asia. *arXiv:2404.03608*. doi:10.48550/arXiv.2404.03608
- Du, Z., Zeng, A., Dong, Y., & Tang, J. (2024). Understanding Emergent Abilities of Language Models from the Loss Perspective. *NeurIPS 2024*. *arXiv:2403.15796*. doi:10.48550/arXiv.2403.15796
- Elhady, A., Agirre, E., & Artetxe, M. (2025). Emergent Abilities of Large Language Models under Continued Pretraining for Language Adaptation. *ACL 2025*. *arXiv:2506.00288*. doi:10.48550/arXiv.2506.00288
- Etzaniz, J., Sainz, O., Perez, N., Aldabe, I., Rigau, G., Agirre, E., Ormazabal, A., Artetxe, M., & Soroa, A. (2024). Latxa: An Open Language Model and Evaluation Suite for Basque. *ACL 2024*. *arXiv:2403.20266*. doi:10.48550/arXiv.2403.20266
- Fujii, K., Nakamura, T., Loem, M., Iida, H., Ohi, M., Hattori, K., Shota, H., Mizuki, S., Yokota, R., & Okazaki, N. (2024). Continual Pre-Training for Cross-Lingual LLM Adaptation: Enhancing Japanese Language Capabilities. *arXiv:2404.17790*. doi:10.48550/arXiv.2404.17790
- Gadre, S. Y., Smyrnis, G., Shankar, V., Gururangan, S., Wortsman, M., Shao, R., et al. (2024). Language models scale reliably with over-training and on downstream tasks. *arXiv:2403.08540*. doi:10.48550/arXiv.2403.08540
- Gogoulou, E., Lesort, T., Boman, M., & Nivre, J. (2024). Continual Learning Under Language Shift. *TSD 2024*. *arXiv:2311.01200*. doi:10.48550/arXiv.2311.01200
- Gu, J., Yang, Z., Ding, C., Zhao, R., & Tan, F. (2024). CMR Scaling Law: Predicting Critical Mixture Ratios for Continual Pre-training of Language Models. *EMNLP 2024*. *arXiv:2407.17467*. doi:10.48550/arXiv.2407.17467
- Guo, Y., Fu, J., Zhang, H., Zhao, D., & Shen, Y. (2024). Efficient Continual Pre-training by Mitigating the Stability Gap. *arXiv:2406.14833*. doi:10.48550/arXiv.2406.14833
- Gupta, K., Therien, B., Ibrahim, A., Richter, M. L., Anthony, Q., Belilovsky, E., Rish, I., & Lesort, T. (2023). Continual Pre-Training of Large Language Models: How to (re)warm your model? *arXiv:2308.04014*. doi:10.48550/arXiv.2308.04014
- Hassan, M. T., Ahmed, J., & Awais, M. (2026). Qalb: Largest State-of-the-Art Urdu Large Language Model for 230M Speakers with Systematic Continued Pre-training. *arXiv:2601.08141*. doi:10.48550/arXiv.2601.08141
- He, Y., Benhaim, A., Patra, B., Vaddamanu, P., Ahuja, S., Chopra, P., Chaudhary, V., Zhao, H., & Song, X. (2024). Scaling Laws for Multilingual Language Models. *arXiv:2410.12883*. doi:10.48550/arXiv.2410.12883
- Hernandez, D., Brown, T., Conerly, T., DasSarma, N., Drain, D., El-Showk, S., et al. (2022). Scaling Laws and Interpretability of Learning from Repeated Data. *arXiv:2205.10487*. doi:10.48550/arXiv.2205.10487
- Hu, S., Liu, X., Han, X., Zhang, X., He, C., Zhao, W., et al. (2024). Predicting Emergent Abilities with Infinite Resolution Evaluation. *arXiv:2310.03262*. doi:10.48550/arXiv.2310.03262
- Ibrahim, A., Therien, B., Gupta, K., Richter, M. L., Anthony, Q., Lesort, T., Belilovsky, E., & Rish, I. (2024). Simple and Scalable Strategies to Continually Pre-train Large Language Models. *arXiv:2403.08763*. doi:10.48550/arXiv.2403.08763
- Isik, B., Ponomareva, N., Hazimeh, H., Paparas, D., Vassilvitskii, S., & Koyejo, S. (2025). Scaling Laws for Downstream Task Performance of Large Language Models. *ICLR 2025*. *arXiv:2402.04177*. doi:10.48550/arXiv.2402.04177
- Jeha, P., Sedova, A., Bethune, L., Seto, S., Frellsen, J., Ablin, P., & Schluter, N. (2026). Mix, Don't Tune: Bilingual Pre-Training Outperforms Hyperparameter Search in Data-Constrained Settings. *arXiv:2605.13225*. doi:10.48550/arXiv.2605.13225
- Ji, S., Li, Z., Paavola, J., Lin, P., Chen, P., O'Brien, D., Luo, H., Schuetze, H., Tiedemann, J., & Haddow, B. (2024). EMMA-500: Enhancing Massively Multilingual Adaptation of Large Language Models. *arXiv:2409.17892*. doi:10.48550/arXiv.2409.17892
- Kim, K., Kotha, S., Liang, P., & Hashimoto, T. (2025). Pre-training under infinite compute. *arXiv:2509.14786*. doi:10.48550/arXiv.2509.14786
- Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A. A., et al. (2017). Overcoming catastrophic forgetting in neural networks. *arXiv:1612.00796*. doi:10.48550/arXiv.1612.00796
- Kotha, S., Springer, J. M., & Raghunathan, A. (2024). Understanding Catastrophic Forgetting in Language Models via Implicit Inference. *ICLR 2024*. *arXiv:2309.10105*. doi:10.48550/arXiv.2309.10105
- Kunz, J. (2025). Train More Parameters But Mind Their Placement: Insights into Language Adaptation with PEFT. *NoDaLiDa 2025*. *arXiv:2412.12674*. doi:10.48550/arXiv.2412.12674
- Lin, P., Ji, S., Tiedemann, J., Martins, A. F. T., & Schuetze, H. (2024). MaLA-500: Massive Language Adaptation of Large Language Models. *arXiv:2401.13303*. doi:10.48550/arXiv.2401.13303
- Liu, H., Xie, S. M., Li, Z., & Ma, T. (2022). Same Pre-training Loss, Better Downstream: Implicit Bias Matters for Language Models. *arXiv:2210.14199*. doi:10.48550/arXiv.2210.14199
- Liu, Q., Zheng, X., Muennighoff, N., Zeng, G., Dou, L., Pang, T., Jiang, J., & Lin, M. (2024). RegMix: Data Mixture as Regression for Language Model Pre-training. *ICLR 2025*. *arXiv:2407.01492*. doi:10.48550/arXiv.2407.01492
- Longpre, S., Kudugunta, S., Muennighoff, N., Hsu, I., Caswell, I., Pentland, A., Arik, S., Lee, C., & Ebrahimi, S. (2025). ATLAS: Adaptive Transfer Scaling Laws for Multilingual Pretraining, Finetuning, and Decoding the Curse of Multilinguality. *ICLR 2026*. *arXiv:2510.22037*. doi:10.48550/arXiv.2510.22037
- Lourie, N., Hu, M. Y., & Cho, K. (2025). Scaling Laws Are Unreliable for Downstream Tasks: A Reality Check. *Findings of EMNLP 2025*. *arXiv:2507.00885*. doi:10.48550/arXiv.2507.00885
- Meeus, M., Rathe, A., Remy, F., Delobelle, P., Decorte, J., & Demeester, T. (2024). ChocoLlama: Lessons Learned From Teaching Llamas Dutch. *arXiv:2412.07633*. doi:10.48550/arXiv.2412.07633
- Morales-Brotons, D., Vogels, T., & Hendriks, H. (2024). Exponential Moving Average of Weights in Deep Learning: Dynamics and Benefits. *TMLR*. *arXiv:2411.18704*. doi:10.48550/arXiv.2411.18704
- Muennighoff, N., Rush, A. M., Barak, B., Le Scao, T., Piktus, A., Tazi, N., Pyysalo, S., Wolf, T., & Raffel, C. (2023). Scaling Data-Constrained Language Models. *arXiv:2305.16264*. doi:10.48550/arXiv.2305.16264

- Mukhoti, J., Gal, Y., Torr, P. H. S., & Dokania, P. K. (2024). Fine-tuning can cripple your foundation model; preserving features may be the solution. TMLR. arXiv:2308.13320. doi:10.48550/arXiv.2308.13320
- Olsson, C., Elhage, N., Nanda, N., Joseph, N., DasSarma, N., Henighan, T., et al. (2022). In-context Learning and Induction Heads. arXiv:2209.11895. doi:10.48550/arXiv.2209.11895
- Que, H., Liu, J., Zhang, G., Zhang, C., Qu, X., Ma, Y., et al. (2024). D-CPT Law: Domain-specific Continual Pre-Training Scaling Law for Large Language Models. arXiv:2406.01375. doi:10.48550/arXiv.2406.01375
- Rame, A., Ferret, J., Vieillard, N., Dadashi, R., Hussenot, L., Cedoz, P.-L., et al. (2024). WARP: On the Benefits of Weight Averaged Rewarded Policies. arXiv:2406.16768. doi:10.48550/arXiv.2406.16768
- Ruan, Y., Maddison, C. J., & Hashimoto, T. (2024). Observational Scaling Laws and the Predictability of Language Model Performance. NeurIPS 2024. arXiv:2405.10938. doi:10.48550/arXiv.2405.10938
- Schaeffer, R., Miranda, B., & Koyejo, S. (2023). Are Emergent Abilities of Large Language Models a Mirage? arXiv:2304.15004. doi:10.48550/arXiv.2304.15004
- Schaeffer, R., Schoelkopf, H., Miranda, B., Mukobi, G., Madan, V., Ibrahim, A., Bradley, H., Biderman, S., & Koyejo, S. (2024). Why Has Predicting Downstream Capabilities of Frontier AI Models with Scale Remained Elusive? arXiv:2406.04391. doi:10.48550/arXiv.2406.04391
- Seto, S., ter Hoeve, M., Bai, R. H., Schluter, N., & Grangier, D. (2024). Training Bilingual LMs with Data Constraints in the Targeted Language. arXiv:2411.12986. doi:10.48550/arXiv.2411.12986
- Shi, H., Xu, Z., Wang, H., Qin, W., Wang, W., Wang, Y., Wang, Z., Ebrahimi, S., & Wang, H. (2024). Continual Learning of Large Language Models: A Comprehensive Survey. arXiv:2404.16789. doi:10.48550/arXiv.2404.16789
- Shukor, M., Bethune, L., Busbridge, D., Grangier, D., Fini, E., El-Nouby, A., & Ablin, P. (2025). Scaling Laws for Optimal Data Mixtures. arXiv:2507.09404. doi:10.48550/arXiv.2507.09404
- Singh, A. K., Chan, S. C. Y., Moskvitz, T., Grant, E., Saxe, A. M., & Hill, F. (2023). The Transient Nature of Emergent In-Context Learning in Transformers. arXiv:2311.08360. doi:10.48550/arXiv.2311.08360
- Snell, C., Wallace, E., Klein, D., & Levine, S. (2024). Predicting Emergent Capabilities by Finetuning. arXiv:2411.16035. doi:10.48550/arXiv.2411.16035
- Tang, T., Luo, W., Huang, H., Zhang, D., Wang, X., Zhao, X., Wei, F., & Wen, J.-R. (2024). Language-Specific Neurons: The Key to Multilingual Capabilities in Large Language Models. ACL 2024. arXiv:2402.16438. doi:10.48550/arXiv.2402.16438
- Tay, Y., Dehghani, M., Rao, J., Fedus, W., Abnar, S., Chung, H. W., et al. (2022). Scale Efficiently: Insights from Pre-training and Fine-tuning Transformers. ICLR 2022. arXiv:2109.10686. doi:10.48550/arXiv.2109.10686
- Wang, X., Tissue, H., Wang, L., Li, L., & Zeng, D. D. (2025). Learning Dynamics in Continual Pre-Training for Large Language Models. ICML 2025. arXiv:2505.07796. doi:10.48550/arXiv.2505.07796
- Wang, Y., Si, S., Li, D., Lukasik, M., Yu, F., Hsieh, C.-J., Dhillion, I. S., & Kumar, S. (2024). Two-stage LLM Fine-tuning with Less Specialization and More Generalization. ICLR 2024. arXiv:2211.00635. doi:10.48550/arXiv.2211.00635
- Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., et al. (2022). Emergent Abilities of Large Language Models. TMLR. arXiv:2206.07682. doi:10.48550/arXiv.2206.07682
- Wendler, C., Veselovsky, V., Monea, G., & West, R. (2024). Do Llamas Work in English? On the Latent Language of Multilingual Transformers. arXiv:2402.10588. doi:10.48550/arXiv.2402.10588
- Wortsman, M., Ilharco, G., Kim, J. W., Li, M., Kornblith, S., Roelofs, R., et al. (2022). Robust fine-tuning of zero-shot models. CVPR 2022. arXiv:2109.01903. doi:10.48550/arXiv.2109.01903
- Wu, Z., Yu, X. V., Yogatama, D., Lu, J., & Kim, Y. (2024). The Semantic Hub Hypothesis: Language Models Share Semantic Representations Across Languages and Modalities. ICLR 2025. arXiv:2411.04986. doi:10.48550/arXiv.2411.04986
- Xi, N., Wu, Y., Fan, K., Chen, T., Gu, Q., & Ji, L. (2024). A Practice of Post-Training on Llama-3 70B with Optimal Selection of Additional Language Mixture Ratio. PAKDD 2025. arXiv:2409.06624. doi:10.48550/arXiv.2409.06624
- Xia, M., Artetxe, M., Zhou, C., Lin, X. V., Pasunuru, R., Chen, D., Zettlemoyer, L., & Stoyanov, V. (2023). Training Trajectories of Language Models Across Scales. ACL 2023. arXiv:2212.09803. doi:10.48550/arXiv.2212.09803
- Xie, S. M., Pham, H., Dong, X., Du, N., Liu, H., Lu, Y., Liang, P., Le, Q. V., Ma, T., & Yu, A. W. (2023). DoReMi: Optimizing Data Mixtures Speeds Up Language Model Pretraining. NeurIPS 2023. arXiv:2305.10429. doi:10.48550/arXiv.2305.10429
- Yamaguchi, A., Morishita, T., Villavicencio, A., & Aletras, N. (2024). Adapting Chat Language Models Using Only Target Unlabeled Language Data. TMLR. arXiv:2412.11704. doi:10.48550/arXiv.2412.11704
- Ye, J., Liu, P., Sun, T., Zhan, J., Zhou, Y., & Qiu, X. (2024). Data Mixing Laws: Optimizing Data Mixtures by Predicting Language Modeling Performance. ICLR 2025. arXiv:2403.16952. doi:10.48550/arXiv.2403.16952
- Yong, Z.-X., Schoelkopf, H., Muennighoff, N., Aji, A. F., Adelani, D. I., Almubarak, K., et al. (2023). BLOOM + 1: Adding Language Support to BLOOM for Zero-Shot Prompting. ACL 2023. arXiv:2212.09535. doi:10.48550/arXiv.2212.09535
- Zhao, J., Zhang, Z., Gao, L., Zhang, Q., Gui, T., & Huang, X. (2024). LLaMA Beyond English: An Empirical Study on Language Capability Transfer. arXiv:2401.01055. doi:10.48550/arXiv.2401.01055
- Zheng, J., Cai, X., Qiu, S., & Ma, Q. (2025). Spurious Forgetting in Continual Learning of Language Models. ICLR 2025. arXiv:2501.13453. doi:10.48550/arXiv.2501.13453
- Zhu, S., Supryadi, Xu, S., Sun, H., Pan, L., Cui, M., Du, J., Jin, R., Branco, A., & Xiong, D. (2024). Multilingual Large Language Models: A Systematic Survey. arXiv:2411.11072. doi:10.48550/arXiv.2411.11072