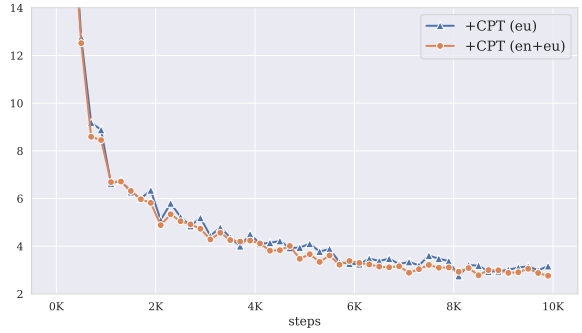
Emergent Abilities of Large Language Models under Continued Pretraining for Language Adaptation

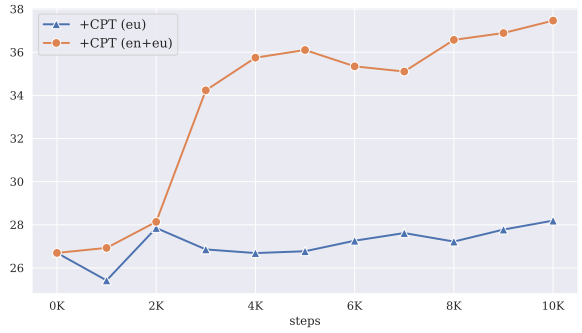
Anonymous ACL submission

Abstract

Continued pretraining (CPT) is a popular approach to adapt existing large language models (LLMs) to new languages. When doing so, it is common practice to include a portion of English data in the mixture, but its role has not been carefully studied to date. In this work, we show that including English does not impact validation perplexity, yet it is critical for the emergence of downstream capabilities in the target language. We introduce a language-agnostic benchmark for in-context learning (ICL), which reveals catastrophic forgetting early on CPT when English is not included. This in turn damages the ability of the model to generalize to downstream prompts in the target language as measured by perplexity, even if it does not manifest in terms of accuracy until later in training, and can be tied to a big shift in the model parameters. Based on these insights, we introduce curriculum learning and exponential moving average (EMA) of weights as effective alternatives to mitigate the need for English. All in all, our work sheds light into the dynamics by which emergent abilities arise when doing CPT for language adaptation, and can serve as a foundation to design more effective methods in the future.



(a) Validation perplexity (eu)



(b) Downstream accuracy (eu)

Figure 1: **Continued pretraining of Llama 2 7B on Basque data with and without including English data.** Both models exhibit similar validation perplexity on Basque (top), yet the variant including English significantly outperforms on downstream tasks (bottom).

1 Introduction

Despite achieving remarkable results in multilingual tasks like machine translation (Zhu et al., 2024), existing large language models (LLMs) are notoriously English-centric, and their performance has been reported to drop significantly in less-resourced languages (Shliazhko et al., 2024; Yong et al., 2023; Scao et al., 2023; Talat et al., 2022). This has motivated a large body of work to extend existing LLMs to new languages through continued pretraining (CPT) (Gogoulou et al., 2024; Etxaniz et al., 2024; Luukkonen et al., 2023; Yong et al., 2023). In its most basic form, CPT uses an existing LLM as initialization and fine-tunes all parameters on next-token prediction over a monolingual corpus in the target language.

Nevertheless, vanilla CPT is rarely used in practice. Instead, there are two techniques that are broadly used in the literature: (i) mixing target language data with English or other languages in the original mixture (Etxaniz et al., 2024; Gogoulou et al., 2024), and (ii) using LORA (Hu et al., 2021) or other parameter-efficient fine-tuning methods (Cui et al., 2024; Yong et al., 2023). There are inherent advantages to these techniques that have typically been used to justify their adoption, such

as preserving English performance when including data in this language (Fujii et al., 2024; Cui et al., 2024), or reducing memory requirements when performing parameter-efficient fine-tuning (Hu et al., 2021). Perhaps more intriguingly, there have also been isolated reports of these two techniques improving performance in the target language (Ji et al., 2024; Etxaniz et al., 2024).

Through systematic experiments, we corroborate that including English data when doing CPT is critical to obtain strong in-context learning (ICL) performance in the target language (§4.1). For instance, we obtain considerably better results on Basque downstream tasks when performing CPT of Llama 2 on a mixture of Basque and English data, as opposed to Basque alone (Figure 1b). But, to our surprise, we find that both mixtures perform at par in terms of Basque perplexity (Figure 1a). We find this to be counterintuitive: both models do equally well in terms of the pretraining objective in the target language,[1] yet downstream capabilities only emerge in one of the variants, challenging prior observations in monolingual settings that models with a similar perplexity tend to obtain similar performance in downstream tasks (Du et al., 2024; Xia et al., 2023).

We present an empirical study of the training dynamics that lead to this behavior. We introduce Copain, a new benchmark to evaluate ICL in a language-agnostic manner (§3), which reveals that CPT without English suffers from a catastrophic forgetting of its ICL capabilities in the first few steps of training (§4.2). We further show that the ability of this variant to generalize to downstream prompts gets severely damaged at this exact same period as measured by perplexity, even if it does not manifest in terms of accuracy until much later in training (§4.3). Finally, we show that this behavior can be tied to a strong shift in the model parameters when English is not included in the CPT mixture (§4.4).

Based on these insights, we explore two alternative approaches that mitigate the need for English. First, we show that including English in the first 10% training steps in a curriculum learning fashion is sufficient, confirming that the critical period is concentrated in the first stage of CPT (§5.1). Second, we eliminate the need for English by applying the exponential moving average (EMA) of weights, which acts as a regularizer to limit the parameter shift (§5.2).

All in all, our work sheds light into the dynamics that condition the emergence of downstream abilities when doing CPT for language adaptation, which transcend what is directly observable by inspecting the training loss. We validate our findings by designing two CPT variants that mitigate the need for English, and we hope that our analysis can serve as a foundation to design more effective language adaptation methods in the future.

2 Experimental Setup

Our focus is to analyze the CPT of English-centric LLMs for language adaptation. To that end, we conduct separate experiments on 3 diverse target languages: Basque, Arabic and Indonesian. We next detail our experimental settings.

Base models. We use Llama 2 7B (Touvron et al., 2023) as the base model for most of our experiments. We base this choice on the original pretraining of Llama 2 being notoriously English-centric,[2] making it a good fit to study CPT under language shift. To understand the impact of scale and different base models, we run additional experiments on Basque using Llama 2 13B, Llama 3.1 8B (Dubey et al., 2024) and Gemma 2 9B (Riviere et al., 2024).

Training data. For Basque, we use the Latxa corpus (Etxaniz et al., 2024), which consists of 4.7B tokens of high-quality Basque text. For Arabic and Indonesian, we randomly sample documents from their respective CulturaX corpus (Nguyen et al., 2023). We ensure all languages have token count parity (4.5~4.7B tokens per language). When including English in the CPT mixture, we use a random sample of 500k English documents from the Pile (Gao et al., 2020). For all languages, the English data accounts for 20% of the total CPT tokens.

Hyperparameters. All models are continued pretrained for 10k steps on 4×8 A100 GPUs. The learning rate is set to 1E-04 with cosine scheduling and a 10% warm-up ratio. The maximum sequence length is set to 4096 and the effective batch size to 256. These hyperparameters were chosen in accordance with Etxaniz et al. (2024); we did not

[1] Note that perplexity is the exponential of cross-entropy, which is used as the loss.

[2] According to the authors, their pretraining dataset included 0.03% of Indonesian and less than 0.005% of Basque and Arabic.

observe any significant impact when varying them in our early experiments (see Appendix A).

Evaluation. For all models, we report the perplexity on the validation split of their respective data. We assess the performance on downstream tasks using multiple-choice benchmarks. For Arabic and Indonesian, we report accuracy on ArabicMMLU (Koto et al., 2024) and IndoMMLU (Koto et al., 2023), respectively. Both benchmarks consist of 5 sub-tasks measuring language proficiency, reasoning ability, and cultural knowledge of their respective language. For Basque, we report average accuracy across EusTrivia, EusProficiency, EusExams and EusReading (Etxaniz et al., 2024). All benchmarks use multiple choice prompting (Robinson et al., 2023) with 5-shot examples, except for EusReading which uses 1-shot (see Appendix C for details). In addition, we report accuracy in Copain—a new language-agnostic ICL benchmark we introduce in §3—for all languages.

3 Copain: A Language-Agnostic ICL Benchmark

Multiple-choice benchmarks require both (i) a good level of ICL (so the LLM generates an answer in the expected format based on the few-shot demonstrations), and (ii) knowledge of the relevant task in the target language. Intuitively, (i) is not tied to any specific language, so the initial models should already be capable on it, while (ii) should improve as we perform CPT in the target language. However, the fact that these two aspects are conflated in downstream metrics makes it hard to understand why certain variants underperform others. Are they less effective at learning the target language? Or do they become weaker at ICL?

So as to evaluate ICL in a language-agnostic manner, we introduce the **Co**ntextual **pa**ttern **in**ference (Copain) benchmark. As shown in Table 1, the input of the task is a list of either numbers or characters, and the model needs to output the element in the list that meets certain criterion. However, there is no natural language instruction in the prompt, so the model needs to infer the task from the few-shot demonstrations.

The benchmark comprises 7 tasks. Each task consists of 150 examples, totaling 1050. The tasks are to identify: (i) the minimum/maximum/median number in a list of 3 integers, (ii) the even/odd number in a list of 4 integers, and (iii) the alphabetically first/last character in a list of 3 Latin characters. We use exact-match accuracy as the evaluation metric.

Task	Example Prompt
Max. integer in list	85, 24, 63: 85 29, 47, 79: 79 59, 77, 41: 77 19, 81, 88:
Min. integer in list	85, 24, 63: 24 29, 47, 79: 29 59, 77, 41: 41 19, 81, 88:
Median integer in list	85, 24, 63: 63 29, 47, 79: 47 59, 77, 41: 59 19, 81, 88:
Even integer among odd list	21, 71, 68, 95: 68 25, 35, 58, 83: 58 92, 71, 61, 29: 92 97, 66, 1, 3:
Odd integer among even list	24, 76, 60, 51: 51 83, 52, 22, 52: 83 32, 68, 10, 79: 79 64, 87, 0, 28:
First character alphabetically	w, y, a: a b, m, k: b v, h, p: h y, e, p:
Last character alphabetically	w, y, a: y b, m, k: m v, h, p: v y, e, p:

Table 1: **Example Copain tasks using 3-shot demonstrations.** The model's predictions are evaluated using exact match accuracy.

4 The Impact of English in CPT

In this section, we study the impact of English when doing CPT for language adaptation, covering the final performance at the end of CPT (§4.1), the learning trajectory (§4.2), the generalization behavior as measured by perplexity (§4.3), and the parameter shift (§4.4).

4.1 Final performance

We start by analyzing the final performance of the models at the end of CPT. As shown in Table 2, CPT brings big gains over all base models in terms of target language perplexity. Both CPT variants perform at par: the one with English wins in 3 instances and the one without wins in the remaining 3, but the differences are small in all cases. This suggests that including English does not directly help language modeling in the target language, at least in terms of the pretraining objective itself.

However, we do observe notable differences in downstream performance. Just as with perplexity, all CPT models outperform their corresponding base model. But, in this case, the variant including

	PPL	Dwn.	Cop.
Basque (eu)			
Llama 2 (7B)	23.64	27.43	**44.67**
+ CPT (eu+en)	**3.35**	**34.14**	43.43
+ CPT (eu)	3.58	28.89	20.12
Llama 2 (13B)	13.66	29.52	**49.23**
+ CPT (eu+en)	2.82	**42.52**	47.80
+ CPT (eu)	**2.79**	35.20	29.43
Llama 3.1 (8B)	2.18	42.31	41.32
+ CPT (eu+en)	**1.73**	**55.75**	**42.04**
+ CPT (eu)	1.82	54.84	41.19
Gemma 2 (9B)	2.28	42.22	**51.90**
+ CPT (eu+en)	1.52	**49.39**	50.23
+ CPT (eu)	**1.48**	45.95	43.59
Arabic (ar)			
Llama 2 (7B)	4.36	32.45	**44.67**
+ CPT (ar+en)	**2.09**	**34.34**	32.60
+ CPT (ar)	2.12	32.67	23.80
Indonesian (id)			
Llama 2 (7B)	6.27	26.65	**44.67**
+ CPT (id+en)	3.25	**30.79**	30.79
+ CPT (id)	**3.05**	26.92	27.34

Table 2: **Main results for each base model and its continued pretraining with and without English.** We report validation perplexity in the target language (*PPL*), average downstream accuracy in the target language (*Dwn*; see §2 for details), and Copain accuracy (*Cop*).

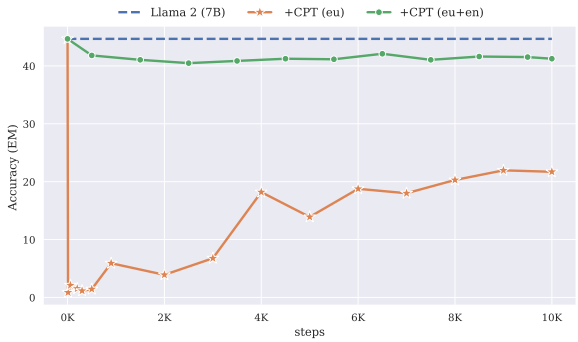


Figure 2: **Copain results for Llama 2 7B.** Including English during CPT retains over 94% of the original performance, while not including it results in catastrophic forgetting followed by a slow partial recovery.

English obtains considerably better results, beating the variant not including English in all cases. The weaker the base model is in a given language, the more helpful including English tends to be, with a difference over 7 points for Llama 2 13B in Basque.

Results in Copain show a similar trend: the CPT variant with English outperforms the one without in all cases. The differences tend to be large (e.g., around 20 points for Llama 2 in Basque), although they greatly vary across languages and base models. However, different from perplexity and downstream tasks, it is the initial model that obtains the best results in most cases (5 out of 6). This suggests that doing CPT for language adaptation tends to harm the ICL capabilities of LLMs, and including English helps mitigate this.

All in all, our results show that including English during CPT leads to considerably better downstream performance. However, this difference does not manifest in perplexity, and can instead be attributed to a better preservation of the ICL capabilities of the original model.

4.2 Learning Trajectory

Our results so far were limited to the final performance of the models. In this section, we will look at how their behavior evolves throughout CPT. To that end, we will focus on Llama 2 7B in Basque, for which we have previously observed one of the biggest differences in final performance.

As shown in Figure 1a, the learning curve for perplexity looks very similar regardless of whether English is included or not. In contrast, downstream performance shows an emergent behavior when including English, with a sudden improvement of 8 points between steps 2k and 4k, while it never takes off when English is not included (Figure 1b). This challenges prior findings that models with a similar perplexity obtain a similar downstream performance, with certain abilities emerging when perplexity falls below a certain threshold (Xia et al., 2023; Du et al., 2024). While those studies focused on monolingual pretraining, we show that this behavior does not hold more broadly when doing CPT for language adaptation.

Figure 2 shows that both CPT variants behave differently in Copain too. When English is not included, we observe catastrophic forgetting early on training, with performance plummeting to nearly zero in the first few steps. This is followed by a slow improvement throughout the rest of training, which is far from recovering the full performance of the original model. The CPT variant with English suffers a more progressive degradation in the first 2k steps, but it is very mild in comparison, and performance remains nearly constant after that.

Based on these results, we hypothesize that there is a critical period at the beginning of CPT, where the strong distribution shift from switching to a new language can result in a catastrophic forgetting of the ICL capabilities of the model. This would in turn prevent the emergence of downstream ca-

pabilities later on, even if not directly impacting the training objective as reflected by the validation perplexity. Including English data in the mixture would alleviate this distribution shift, mitigating the catastrophic forgetting.

4.3 Generalization Behavior

We have so far established that including English data in CPT is critical for good downstream and ICL performance, despite not having an impact on validation perplexity. In other words, the two CPT variants perform similarly when evaluated in the training distribution,[3] but generalize differently to few-shot tasks that are out of this distribution. However, the two aspects are evaluated using different metrics (perplexity vs. accuracy). The former is a function of the training loss, but the latter is not directly tied to it, making it difficult to understand the nature of this different generalization behavior. To overcome this, we next formulate downstream tasks as conditional text generation, and use perplexity to evaluate different models on it.

Given a set of few-shot demonstrations C and a question Q, the model predicts the probability of each answer label A conditioned on the prefix prompt:

$$\mathrm{P}(A|C,Q) = \prod_{t=1}^{T} p(A_t|C, Q < T) \qquad (1)$$

The perplexity of the answers aggregated over the entire test set can be computed as follows:

$$\mathrm{PPL}(A|Q,C) = 2^{-\frac{1}{N}\sum_{i=1}^{N}\log P(a_i|c_i,q_i)} \qquad (2)$$

We separately compute the perplexity of the correct and incorrect answers. Intuitively, we want the gap between the two to be as high as possible, as strong models should assign a higher probability to correct answers than to incorrect answers. As shown in Figure 3, CPT with English is effective at achieving this: the perplexity of incorrect answers increases early on training, while the perplexity of correct answers remains constant or even slightly goes down. When English data is not included, the perplexity of both answers spikes in the first few steps of CPT. Even if it tends to stabilize later on, it stays high compared to the variant with English. Moreover, the gap between the correct and the incorrect answers is much smaller.

[3] Our training and validation datasets are obtained by taking two random splits of the original corpora, and they thus come from the same distribution.

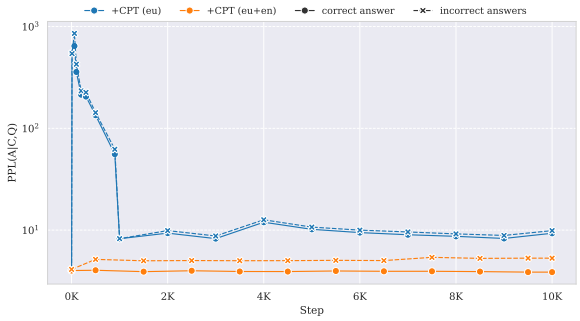


Figure 3: **Perplexity of choice labels on Basque downstream tasks for Llama 2 7B.** The variant without English experiences a spike in perplexity simultaneous with the drop in ICL (Figure 2). *PPL* of incorrect labels are averaged.

In conclusion, when it comes to their pure language modeling performance, both CPT variants behave similarly in the training distribution. But, when not including English, the ability to generalize to multiple choice prompts that are out of this distribution gets severely damaged after the first few steps. Even if the difference in downstream accuracy becomes prominent later on training (around step 3k in Figure 1b), this shows that the real damage happens much earlier, in line with the drop observed for Copain in §4.2.

4.4 Parameters Shift

Our experiments so far have shown that excluding English from the mixture causes catastrophic forgetting of some emergent abilities in the first few steps of CPT. In this section, we analyze the underlying training dynamics that cause this behavior. To that end, we measure how much the model parameters change with respect to their initial value as training progresses.

As shown in Figure 4, the variant without English experiences a stronger parameter shift. The shift rapidly builds up during the first few steps: at the 100th step, the cumulative L2 distance is 7x higher for the variant without English, reaching 15x by the 1000th step. In contrast, CPT with English shows a more steady and regularized parameter change. In fact, the variant without English undergoes a bigger change in the first 10 steps than the variant with English during the entire course of training.

We further run an experiment using LoRA with Basque data only.[4] We find that LoRA has a similar

[4] We set the rank to 512 following Talla et al. (2024), which corresponds to 20% additional trainable parameters.

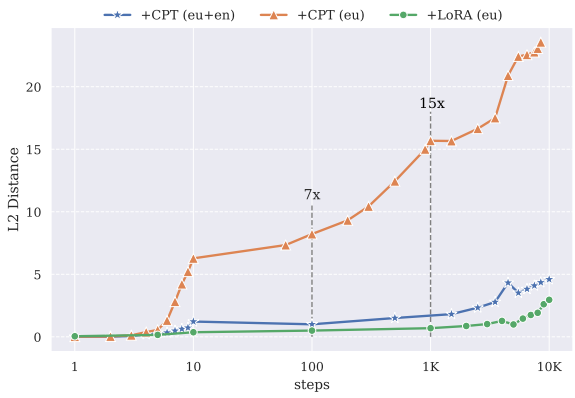


Figure 4: **L2 distance of model parameters** from the initial Llama 2 7B model throughout full-parameter CPT and using LoRA. The CPT variant without English data rapidly diverges from the initial weights during the first 1k steps. The divergence is slowed for the rest of the training steps.

	PPL	Dwn.	Cop.
Basque (eu)			
Llama 2 (7B)	23.64	27.43	**44.67**
+ CPT (full)	3.35	34.14	43.43
+ CPT (curr)	**3.08**	**35.12**	42.94
Llama 2 (13B)	13.66	29.52	**49.23**
+ CPT (full)	2.82	**42.52**	47.80
+ CPT (curr)	**2.65**	42.42	46.33
Arabic (ar)			
Llama 2 (7B)	4.36	32.45	**44.67**
+ CPT (full)	2.09	34.34	32.60
+ CPT (curr)	**2.00**	**34.53**	39.66
Indonesian (id)			
Llama 2 (7B)	6.27	26.65	**44.67**
+ CPT (full)	3.25	**30.79**	30.79
+ CPT (curr)	**3.14**	29.09	31.03

Table 3: **Results with English data added for all training steps (*full*) and for the first 10% steps (*curr*).** We report validation perplexity (*PPL*), average downstream accuracy (*Dwn*) and Copain accuracy (*Cop*).

effect to including English, with an even smaller shift in model parameters.[5] As reported in Appendix B, this approach is quite effective at preserving ICL performance, but barely improves over the initial model on downstream tasks. This suggests that overly constraining the parameter shift can hinder learning the target language, while giving too much flexibility can cause catastrophic forgetting of ICL capabilities. Given its effectiveness at reducing the parameter shift, this can explain why, with optimized hyperparameters, prior work has found LoRA to outperform full-parameter CPT in low-resource languages when English is not included (Ji et al., 2024; Yong et al., 2023).

5 Alternatives to Including English Data

Our analysis in §4 shows that (i) there is a critical period early on CPT where catastrophic forgetting occurs if not including English, and (ii) this phenomenon can be tied to a strong shift in the model's parameters. Based on these insights, we next explore two alternative CPT techniques that achieve the same effect.

5.1 Curriculum Learning

Even if the difference in downstream accuracy arises around step 3k (Figure 1b), our results on ICL capabilities (Figure 2), perplexity of downstream choice labels (Figure 3) and parameter shift (Figure 4) show that the divergence originates much earlier. This can presumably be attributed to the strong distribution shift when changing the training language in CPT, and including English would serve to mitigate this. But what if we make the shift more gradual in a curriculum learning fashion? Is English really needed after the initial critical period?

To answer this, we experiment with including English during the first 1k steps, and omitting it thereafter. As shown in Table 3, models trained with this approach (*curr*) exhibit similar performance compared to those where English is included throughout the entirety of CPT (*full*). More concretely, each variant wins in half of the cases for both downstream and Copain accuracy. Interestingly, the curriculum approach obtains the best perplexity results in all cases, which we speculate could be attributed to the additional training budget in the target language from omitting English. In any case, the differences are small in all cases.

All in all, these results corroborate that the role of English is to provide a smoother transition to the target language distribution. In line with our results in Table 2, this also explains why certain models like Llama 3.1 and Gemma 2 benefit less from including English: those initial models are already decent at modeling the target language distribution (as reflected by their lower validation perplexity), alleviating the distribution shift in CPT and making

[5] LoRA reparameterizes the model weights as $W = W_0 + AB$, where W_0 are the initial model weights (which are kept frozen), and A and B are learnable, low-rank matrices. This way, even if the number of learnable parameters is reduced, one can still measure the parameter shift from W_0 to W in a way that is comparable to full fine-tuning.

	PPL	Dwn.	Cop
Basque (eu)			
Llama 2 (7B)	23.64	27.43	**44.67**
+ CPT (eu+en)	3.35	34.14	43.43
+ CPT w/ EMA (eu)	**2.98**	**34.89**	42.66
Llama 2 (13B)	13.66	29.52	**49.23**
+ CPT (eu+en)	2.82	**42.52**	47.80
+ CPT w/ EMA (eu)	**2.71**	41.39	42.99
Arabic (ar)			
Llama 2 (7B)	4.36	32.45	**44.67**
+ CPT (ar+en)	2.09	**34.34**	32.60
+ CPT w/ EMA (ar)	**2.03**	33.36	42.76
Indonesian (id)			
Llama 2 (7B)	6.27	26.65	**44.67**
+ CPT (id+en)	3.25	**30.79**	30.79
+ CPT w/ EMA (id)	**2.97**	29.11	33.34

Table 4: **Results using EMA of model parameters without English data**. We report validation perplexity (*PPL*), average downstream accuracy (*Dwn*) and Copain accuracy (*Cop*).

the smoother transition from including English less necessary.

5.2 EMA of Model Parameters

As discussed in §4.4, including English significantly reduces the parameter shift during CPT, which can presumably explain why this variant is less prone to catastrophic forgetting. In this section, we explore taking the EMA of the parameters (Morales-Brotons et al., 2024; Cha et al., 2021) as an alternative approach to reduce the parameter shift without requiring any English data.

More concretely, every η steps EMA sets the model parameters to a weighted average between their current value and their value η steps ago:

$$\theta_t = \begin{cases} \theta'_t & \text{if } t \leq 0 \vee t \bmod \eta \neq 0 \\ \alpha\theta_{t-\eta} + (1-\alpha)\,\theta'_t & \text{otherwise} \end{cases}$$

where θ'_t and θ_t denote the model parameters at step t before and after EMA is applied, respectively, and α denotes the decay rate, which we set to 0.92 in all of our experiments. Unless otherwise indicated, we use $\eta = 1$ for Basque and Indonesian, and $\eta = 10$ for Arabic.

As shown in Table 4, EMA is competitive with conventional CPT without the need for any English data. More concretely, it obtains the best validation perplexity in all cases, and comparable results on downstream tasks. This corroborates that the benefit of including English can be tied to alleviating the parameter shift during CPT, and a similar effect can be obtained by using EMA as a regularizer.

However, differences in Copain are bigger and more inconsistent. For instance, EMA outperforms vanilla CPT by 10 points for Llama 2 7B in Arabic, but underperforms it by 5 points for Llama 2 13B in Basque. In relation to this, we found that the interval of applying EMA, η, had a big impact during our preliminary experiments: lower values result in more constrained parameter updates, which helps mitigate the catastrophic forgetting of ICL capabilities, but can potentially obstruct the learning of the target language. Intuitively, we want to set η so the shift in parameters is comparable to that of vanilla CPT with English. But, as shown in Figure 5, this requires different values of η depending on the language. While outside the scope of this work, this prompts for future work to develop more robust methods that can find a good trade-off without excessive hyperparameter tuning.

6 Related Work

CPT for Language Adaptation. The increased availability of LLMs lays a strong foundation for adapting models to new domains through CPT (Abnar et al., 2021). Lately, CPT has been employed to expand LLMs to new languages or boost their performance in languages where they previously struggled (Gogoulou et al., 2024). Compared to training from scratch, CPT achieves promising results by efficiently transferring the knowledge and abilities learned by English-centric LLMs to target languages (Fujii et al., 2024). Full-parameter CPT has been shown to be efficient, provided that sufficient data is available in the target language (Etxaniz et al., 2024; Luukkonen et al., 2023; Yong et al., 2023). For low-resource languages, LoRA is often leveraged in CPT. Separate LoRA weights can be trained for each target language (Fujii et al., 2024; Badola et al., 2023) or for all languages collectively (Ji et al., 2024) and then merged with the original weights. However, comprehensive investigations on the effectiveness of LoRA in CPT are limited.

Stability Gap in Continual Learning. Continual learning aims to accumulate knowledge in deep neural models (Parisi et al., 2019). It is often used to extend pretrained models to new domains, tasks, and languages. However, the ongoing distribution shift leads to catastrophic forgetting of previous capabilities (Ghunaim et al., 2023). A large body of literature focused on mitigating this forgetting, mentioning it often occurs as a transition phase to

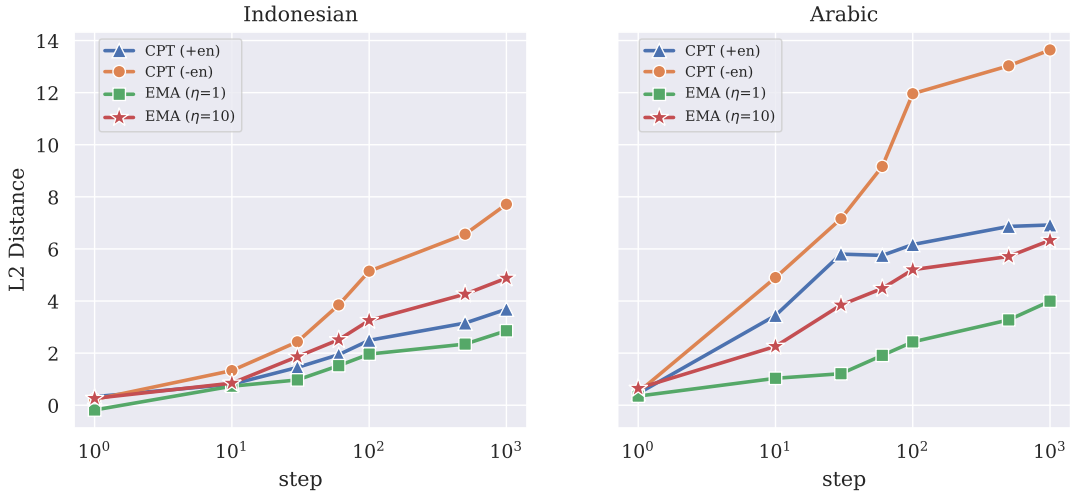


Figure 5: **L2 distance of model parameters from the initial Llama 2 (7B) during CPT** with and without English data, and using EMA with interval (η) of 1 and 10.

the new distribution (Lange et al., 2023; Caccia et al., 2022): the phase is referred to as the *Stability Gap*. During this gap, models lose performance on previously learned tasks before recovering during training, or sometimes not at all. Our analysis shows an analogous yet extreme case of the stability gap. Namely, we see a rapid loss in ICL on Copain (§3), from which the model struggles to recover.

EMA of Model Weights. EMA stabilizes the training of deep learning models. It is often employed in approaches that focus on improving the generalization of the final model or models close to convergence (Cha et al., 2021; Yang et al., 2019; Izmailov et al., 2018; Nikishin et al., 2018). Furthermore, EMA allows the use of higher learning rates, which is particularly beneficial for training LLMs with large effective batch sizes (Morales-Brotons et al., 2024). Lately, EMA gained wider use in alignment of LLMs (Ouyang et al., 2022). For example, it is used as a dynamic anchor in regularization to prevent forgetting of pretrained knowledge while optimizing for rewards during Reinforcement Learning from Human Feedback (RLHF) (Ramé et al., 2024) and Proximal Policy Optimization (PPO) (Schulman et al., 2017).

7 Conclusion

In this paper, we have shown that including English data in CPT can be critical for downstream capabilities to emerge in the new language, despite not having an impact in validation perplexity. This can be traced back to a critical period early on CPT, during which a drastic change in the training distribution when switching to a new language causes a big shift in the model parameters, which in turn results in the catastrophic forgetting of its ICL capabilities. Based on these insights, we have shown that curriculum learning and EMA can achieve the same effect while reducing—or fully eliminating—the need for English data, further validating our findings.

While the focus of our work was to analyze the dynamics by which emergent abilities arise during CPT, we believe that our insights can be helpful to develop better strategies for language adaptation in the future. In particular, one of our key findings is that controlling the degree of parameter shift is critical for good downstream performance: giving too much flexibility can result in the catastrophic forgetting of ICL, but overly constraining it can hinder the learning of the target language. Our results with both LoRA and EMA show that finding the right balance can be very sensitive to hyperparameters, and even including English is not a universal solution as reflected by the considerable drop in Copain performance in the case of Arabic and Indonesian (Table 2). In the future, we want to explore more robust CPT approaches that can find the optimal trade-off without the need for excessive hyperparameter tuning.

Limitations

Our analysis of emergent abilities was limited to multiple-choice downstream tasks and language-agnostic ICL. It would be interesting to extend

the study to other capabilities, both from the perspective of language-independent skills that might suffer from catastrophic forgetting, and target language skills that may or may not emerge depending on the training dynamics. However, the scarcity of relevant benchmarks, in particular in low-resource languages, hinders this study.

In addition, our experiments were limited to including English in combination with the target language. Experimenting with other high-resource languages could provide additional insights, in particular when closely related to the target language.

References

Samira Abnar, Mostafa Dehghani, Behnam Neyshabur, and Hanie Sedghi. 2021. Exploring the limits of large scale pre-training. *Preprint*, arXiv:2110.02095.

Kartikeya Badola, Shachi Dave, and Partha Talukdar. 2023. Parameter-efficient finetuning for robust continual multilingual learning. *Preprint*, arXiv:2209.06767.

Lucas Caccia, Rahaf Aljundi, Nader Asadi, Tinne Tuytelaars, Joelle Pineau, and Eugene Belilovsky. 2022. New insights on reducing abrupt representation change in online continual learning. *Preprint*, arXiv:2104.05025.

Junbum Cha, Sanghyuk Chun, Kyungjae Lee, Han-Cheol Cho, Seunghyun Park, Yunsung Lee, and Sungrae Park. 2021. Swad: Domain generalization by seeking flat minima. *Advances in Neural Information Processing Systems*, 34:22405–22418.

Yiming Cui, Ziqing Yang, and Xin Yao. 2024. Efficient and effective text encoding for chinese llama and alpaca. *Preprint*, arXiv:2304.08177.

Zhengxiao Du, Aohan Zeng, Yuxiao Dong, and Jie Tang. 2024. Understanding emergent abilities of language models from the loss perspective. *Preprint*, arXiv:2403.15796.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, and 516 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Julen Etxaniz, Oscar Sainz, Naiara Perez, Itziar Aldabe, German Rigau, Eneko Agirre, Aitor Ormazabal, Mikel Artetxe, and Aitor Soroa. 2024. Latxa: An open language model and evaluation suite for basque. *Preprint*, arXiv:2403.20266.

Kazuki Fujii, Taishi Nakamura, Mengsay Loem, Hiroki Iida, Masanari Ohi, Kakeru Hattori, Hirai Shota, Sakae Mizuki, Rio Yokota, and Naoaki Okazaki. 2024. Continual pre-training for cross-lingual llm adaptation: Enhancing japanese language capabilities. *Preprint*, arXiv:2404.17790.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, and Connor Leahy. 2020. The pile: An 800gb dataset of diverse text for language modeling. *Preprint*, arXiv:2101.00027.

Yasir Ghunaim, Adel Bibi, Kumail Alhamoud, Motasem Alfarra, Hasan Abed Al Kader Hammoud, Ameya Prabhu, Philip H. S. Torr, and Bernard Ghanem. 2023. Real-time evaluation in online continual learning: A new hope. *Preprint*, arXiv:2302.01047.

Evangelia Gogoulou, Timothée Lesort, Magnus Boman, and Joakim Nivre. 2024. Continual learning under language shift. *Preprint*, arXiv:2311.01200.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *Preprint*, arXiv:2106.09685.

Pavel Izmailov, Dmitrii Podoprikhin, Timur Garipov, Dmitry Vetrov, and Andrew Gordon Wilson. 2018. Averaging weights leads to wider optima and better generalization. *arXiv preprint arXiv:1803.05407*.

Shaoxiong Ji, Zihao Li, Indraneil Paul, Jaakko Paavola, Peiqin Lin, Pinzhen Chen, Dayyán O'Brien, Hengyu Luo, Hinrich Schütze, Jörg Tiedemann, and Barry Haddow. 2024. Emma-500: Enhancing massively multilingual adaptation of large language models. *Preprint*, arXiv:2409.17892.

Fajri Koto, Nurul Aisyah, Haonan Li, and Timothy Baldwin. 2023. Large language models only pass primary school exams in indonesia: A comprehensive test on indommlu. *Preprint*, arXiv:2310.04928.

"Fajri Koto, Haonan Li, Sara Shatanawi, Jad Doughman, Abdelrahman Boda Sadallah, Aisha Alraeesi, Khalid Almubarak, Zaid Alyafeai, Neha Sengupta, Shady Shehata, Nizar Habash, Preslav Nakov, and Timothy Baldwin". 2024. Arabicmmlu: Assessing massive multitask language understanding in arabic. In *Findings of the Association for Computational Linguistics: ACL 2024*.

Matthias De Lange, Gido van de Ven, and Tinne Tuytelaars. 2023. Continual evaluation for lifelong learning: Identifying the stability gap. *Preprint*, arXiv:2205.13452.

Risto Luukkonen, Ville Komulainen, Jouni Luoma, Anni Eskelinen, Jenna Kanerva, Hanna-Mari Kupari, Filip Ginter, Veronika Laippala, Niklas Muennighoff, Aleksandra Piktus, Thomas Wang, Nouamane Tazi, Teven Scao, Thomas Wolf, Osma Suominen, Samuli Sairanen, Mikko Merioksa, Jyrki Heinonen, Aija Vahtola, and 2 others. 2023. FinGPT: Large generative models for a small language. In *Proceedings of the*

2023 Conference on Empirical Methods in Natural Language Processing, pages 2710–2726, Singapore. Association for Computational Linguistics.

Daniel Morales-Brotons, Thijs Vogels, and Hadrien Hendrikx. 2024. Exponential moving average of weights in deep learning: Dynamics and benefits. *Transactions on Machine Learning Research*.

Thuat Nguyen, Chien Van Nguyen, Viet Dac Lai, Hieu Man, Nghia Trung Ngo, Franck Dernoncourt, Ryan A. Rossi, and Thien Huu Nguyen. 2023. Culturax: A cleaned, enormous, and multilingual dataset for large language models in 167 languages. *Preprint*, arXiv:2309.09400.

Evgenii Nikishin, Pavel Izmailov, Ben Athiwaratkun, Dmitrii Podoprikhin, Timur Garipov, Pavel Shvechikov, Dmitry Vetrov, and Andrew Gordon Wilson. 2018. Improving stability in deep reinforcement learning with weight averaging. In *Uncertainty in artificial intelligence workshop on uncertainty in Deep learning*.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. *Preprint*, arXiv:2203.02155.

German I Parisi, Ronald Kemker, Jose L Part, Christopher Kanan, and Stefan Wermter. 2019. Continual lifelong learning with neural networks: A review. *Neural networks*, 113:54–71.

Alexandre Ramé, Johan Ferret, Nino Vieillard, Robert Dadashi, Léonard Hussenot, Pierre-Louis Cedoz, Pier Giuseppe Sessa, Sertan Girgin, Arthur Douillard, and Olivier Bachem. 2024. Warp: On the benefits of weight averaged rewarded policies. *Preprint*, arXiv:2406.16768.

Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan Ferret, Peter Liu, Pouya Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, Anton Tsitsulin, and 178 others. 2024. Gemma 2: Improving open language models at a practical size. *Preprint*, arXiv:2408.00118.

Joshua Robinson, Christopher Michael Rytting, and David Wingate. 2023. Leveraging large language models for multiple choice question answering. *Preprint*, arXiv:2210.12353.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, Jonathan Tow, Alexander M. Rush, Stella Biderman, Albert Webson, Pawan Sasanka Ammanamanchi, Thomas Wang, Benoît Sagot, Niklas Muennighoff, Albert Villanova del Moral, and 373 others. 2023. Bloom: A 176b-parameter open-access multilingual language model. *Preprint*, arXiv:2211.05100.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *Preprint*, arXiv:1707.06347.

Oleh Shliazhko, Alena Fenogenova, Maria Tikhonova, Anastasia Kozlova, Vladislav Mikhailov, and Tatiana Shavrina. 2024. mGPT: Few-Shot Learners Go Multilingual. *Transactions of the Association for Computational Linguistics*, 12:58–79.

Zeerak Talat, Aurélie Névéol, Stella Biderman, Miruna Clinciu, Manan Dey, Shayne Longpre, Sasha Luccioni, Maraim Masoud, Margaret Mitchell, Dragomir Radev, Shanya Sharma, Arjun Subramonian, Jaesung Tae, Samson Tan, Deepak Tunuguntla, and Oskar Van Der Wal. 2022. You reap what you sow: On the challenges of bias evaluation under multilingual settings. In *Proceedings of BigScience Episode #5 – Workshop on Challenges & Perspectives in Creating Large Language Models*, pages 26–41, virtual+Dublin. Association for Computational Linguistics.

Franck Signe Talla, Herve Jegou, and Edouard Grave. 2024. Neutral residues: revisiting adapters for model extension. *Preprint*, arXiv:2410.02744.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, and 49 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

Mengzhou Xia, Mikel Artetxe, Chunting Zhou, Xi Victoria Lin, Ramakanth Pasunuru, Danqi Chen, Luke Zettlemoyer, and Ves Stoyanov. 2023. Training trajectories of language models across scales. *Preprint*, arXiv:2212.09803.

Guandao Yang, Tianyi Zhang, Polina Kirichenko, Junwen Bai, Andrew Gordon Wilson, and Chris De Sa. 2019. Swalp: Stochastic weight averaging in low precision training. In *International Conference on Machine Learning*, pages 7015–7024. PMLR.

Zheng-Xin Yong, Hailey Schoelkopf, Niklas Muennighoff, Alham Fikri Aji, David Ifeoluwa Adelani, Khalid Almubarak, M Saiful Bari, Lintang Sutawika, Jungo Kasai, Ahmed Baruwa, Genta Indra Winata, Stella Biderman, Edward Raff, Dragomir Radev, and Vassilina Nikoulina. 2023. Bloom+1: Adding language support to bloom for zero-shot prompting. *Preprint*, arXiv:2212.09535.

Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. 2024. Multilingual machine translation with large language models: Empirical results and analysis. *Preprint*, arXiv:2304.04675.

	LR	PPL	Dwn.	Cop.
Llama-2 (7B)	–	23.64	27.43	44.67
+ CPT (eu)	1e0-4	3.58	28.89	20.12
+ CPT (eu)	5e0-5	8.76	27.98	26.43
+ CPT (eu)	1e0-5	8.29	27.42	29.43

Table 5: **Preliminarily experiments** to determine whether lower learning rates can reduce the impact of forgetting of the ICL capabilities during CPT. We use the Llama-2-7B model and do CPT with Basque data only.

	PPL	Dwn.	Cop.
Llama-2 (7B)	23.64	27.43	**44.67**
+ CPT (eu+en)	**3.35**	**34.14**	43.43
+ CPT (eu)	3.58	28.89	20.12
+ LoRA (eu)	3.68	28.03	39.61

Table 6: **Results of CPT using LoRA** compared to full parameter CPT with and without English.

A Initial Experiments

In our initial experiments, we experimented with CPT of the Llama 2 7B model with Basque data only using smaller learning rates to reduce the impact of catastrophic forgetting. The results of these experiments are shown in Table 5. We find that reducing the learning rate up to a factor of 10 not only did not solve the problem, but it also hindered the learning of the new language.

B Continued Pre-training using LoRA

Table 6 shows the results of using LoRA in CPT compared to full parameter CPT with and without including English data.

C Multiple Choice Prompting Computation

In multiple choice prompting, the model is prompted with few-shot demonstrations c and a question q and the set of choices $A = \{A, B, C, D\}$. It generates a probability of the answer label $a \epsilon A$ conditioned on the prefix prompt given by Equation 1. The model's answer is then set to:

$$\underset{a \epsilon A}{\textbf{\textit{argmax}}}(P(a|c, q)) \quad (3)$$

Figure 6 shows examples from EusExams using multiple choice prompting.

D Detailed Downstream Performance

Table 7 reports detailed downstream results in the different subsets of Basque downstream tasks, while Table 8 does so for Arabic and Indonesian.

Galdera: LANEKO ARRISKUEN PREBENTZIOARI DAGOKIONEZ, KOLORE HAUEK AURKITU DITZAKEGU:
A: Gorria, berdea, urdina, horia edo hori-laranja
B: Gorria, horia edo hori-laranja
C: Gorria, zuria, beltza, urdina edo berdea
D: Gorria, urdina, beltza edo zuria
Erantzuna: A

Galdera: EUSKAL FUNTZIO PUBLIKOAREN LEGEAK EZARRITAKOARI JARRAIKIZ, NOIZ ESKA LEZAKE FUNTZIONARIO BATEK INTERES PARTIKULARRAGATIKO BORONDATEZKO ESZEDENTZIA BAT?
A: Aurreko bost urteetan edozein administrazio publikotan benetako zerbitzuak bete dituenean, eta eszedentzia horrek ezingo du segidako bi urte baino gutxiago iraun.
B: Aurreko hiru urteetan edozein administrazio publikotan benetako zerbitzuak bete dituenean, eta eszedentzia horrek ezingo du segidako bi urte baino gutxiago iraun.
C: Aurreko bi urteetan edozein administrazio publikotan benetako zerbitzuak bete dituenean, eta ez da zehazten eszedentzia horren gutxieneko iraupena.
D: Diziplina-espediente bat izan arren edo zigor bat betetzeke izan arren.
Erantzuna: A

Galdera: LEGE ORGANIKO BAT ONARTZEKO, BEHARREZKOA DA:
A: Kongresuaren eta Senatuaren gehiengo sinplea, baterako bilkuran.
B: Kongresuaren 2/3en gehiengoa eta Senatuaren gehiengo osoa proiektu osoari buruz.
C: Kongresuaren gehiengo osoa, proiektu osoari buruzko azken bozketan.
D: Kongresuaren gehiengo soila, baina oinarrizko eskubideekin eta askatasun publikoekin, Autonomia Estatutuen onespenarekin eta hauteskunde-araubide orokorrarekin zerikusia duten gai jakin batzuei buruzkoa denean baino ezin izango da onartu.
Erantzuna: C

Galdera: KALTEGARRITASUN-DEKLARAZIOA EZIN EGINGO DA ADMINISTRAZIO-EGINTZA EMAN ZENETIK DENBORA-TARTE HAU IGARO BADA:
A: 3 hilabete.
B: Urtebete.
C: 2 urte.
D: 4 urte
Erantzuna: D

Galdera: ADIERAZI BAIEZTAPEN HAUETATIK ZEIN DEN INFORMAZIO-LAGUNTZAKO DOKUMENTUEI BEREZ DAGOKIENA.
A: Sortzen diren bulegoan bertan suntsituko dira.
B: Erregistro Orokorrera bidaliko dira (artxiboko dokumentuetatik bereizita), modu seguru eta konfidentzialean suntsitzeko.
C: Unibertsitateko Dokumentu Ondarearen parte dira.
D: Oharra informazio-laguntzako dokumentutzat hartzen da.
Erantzuna: A

Galdera: ZERTAN OINARRITZEN DA KONSTITUZIOA?
A: Espainiako nazioaren batasunean.
B: Espainiako nazioa osatzen duten herri eta nazionalitate guztien batasunaren konpromisoan.
C: Espainiako nazioaren batasun ezin hautsizkoan.
D: Espainiako nazioa osatzen duten herrien pluraltasunean.
Erantzuna: C

Galdera: Idatzizko komunikazioan atzeraelikadura ez dela berehalakoa esaten dugunean, zer esan nahi dugu?:
A: Mezuaren igorleak nahi duenean gertatuko dela, ez hartzaileak nahi duenean; horregatik, beti egongo da erantzuna.
B: Mezuaren hartzaileak nahi duenean gertatuko dela, ez igorleak nahi duenean; horregatik, beti egongo da erantzuna.
C: Mezuaren hartzaileak nahi duenean gertatuko dela, ez igorleak nahi duenean; horregatik, batzuetan, gerta daiteke erantzunik ez egotea.
D: Mezuaren igorleak nahi duenean gertatuko dela, ez hartzaileak nahi duenean; horregatik, batzuetan, gerta daiteke erantzunik ez egotea.
Erantzuna: C

Galdera: Emakumea diskriminatzeko era guztiak deuseztatzeari buruzko 1979ko Konbentzioari (CEDAW) dagokionez:
A: Emakumeen eta gizonen berdintasunaren aldeko lehen aurrekari garrantzitsua da, eta Europako Kontseiluan onartu zen.
B: Europar Batasunean onartu zen gai horren inguruko lehen ituna da.
C: Mundu osoan onartutako printzipioak eta estatuek eta zenbait eragile pribatuk hartu beharreko neurri zehatzak aldarrikatzen ditu, juridikoki betebeharrekoak.
D: Mundu osoan onartutako printzipioak eta estatuek eta zenbait eragile pribatuk hartu beharreko neurri zehatzak aldarrikatzen ditu, baina ez dira juridikoki nahitaez betebeharrekoak.
Erantzuna: C

Galdera: Konstituzioaren I. tituluko 3. kapituluan jasotako politika sozial eta ekonomikoaren gidaoinarriak:
A: Eskubide subjektiboak dira, zentzu hertsian.
B: Botere publikoei emandako aginduak dira, haien jarduna bideratzeko.
C: Zuzenean aipa daitezke auzitegi arruntetan, garatuko dituen lege baten beharrik gabe.
D: Konstituzio Auzitegian aipatu ahal izango dira, konstituzio-babeserako errekurtsoaren bidez.
Erantzuna: B

Galdera: Autonomia Erkidegoko eta lurralde historikoetako erakundeen artean eskumen-gatazkarik sortuz gero, nork ebatziko du?:
A: Euskal Autonomia Erkidegoko Justizia Auzitegi Nagusiak.
B: Arbitraje Batzordeak.
C: Aholku Batzorde Juridikoak.
D: Auzitegi Goreneko eskumen-gatazken Salak.
Erantzuna: B

Galdera: Zein egintzaren aurka jar daiteke administrazio-errekurtsoa?:
A: Edozein izapide-egintzaren aurka.
B: Defentsa-gabezia sortzen duen izapide-egintza baten aurka.
C: Behin betiko egintza baten aurka bakarrik.
D: Administrazio-prozedurari amaiera ematen dion egintzaren aurka bakarrik.
Erantzuna: B

Galdera: Espainiako Konstituzioaren arabera, hauetako zein da ordenamendu juridikoaren balio goren bat?:
A: Nazioaren subiranotasuna.
B: Estatu soziala.
C: Berdintasuna.
D: Demokrazia.
Erantzuna: A

Figure 6: Example from EusExams using multiple choice prompting

Basque (eu)					
	EusProf	**EusExams**	**EusRead**	**EusTrivia**	**Average**
Random	25.00	25.00	25.83	26.55	25.59
Llama 2 (7B)	24.09	28.84	27.27	29.50	27.43
+ CPT (eu+en)	29.75	34.20	28.12	44.49	34.14
+ CPT (eu)	25.53	28.70	27.27	34.07	28.89
+ CPT (eu+en) (curr)	30.70	33.48	30.68	45.65	35.12
+ CPT w/ EMA (eu)	30.10	33.45	31.25	44.78	34.89
+ CPT w/ EMA (eu) (curr)	28.19	31.69	30.39	41.63	32.97
+ LoRA (eu)	25.82	27.90	28.49	29.93	28.03
Llama 2 (13B)	25.90	29.66	28.98	33.53	29.52
+ CPT (eu+en)	41.73	40.05	36.09	52.22	42.52
+ CPT (eu)	33.35	32.08	28.69	46.70	35.20
+ CPT (eu+en) (curr)	41.92	40.19	35.39	52.18	42.42
+ CPT w/ EMA (eu)	40.80	40.30	32.67	51.77	40.39
+ CPT w/ EMA (eu) (curr)	40.80	40.30	32.67	51.77	40.39
Llama 3.1 (8B)	32.52	48.01	43.03	45.70	42.31
+ CPT (eu+en)	53.34	54.55	60.47	54.67	55.75
+ CPT (eu)	52.54	53.41	59.07	53.33	54.84
Gemma 2 (9B)	37.19	25.56	52.24	53.88	42.22
+ CPT (eu+en)	47.19	29.10	59.21	62.08	49.39
+ CPT (eu)	43.75	26.66	56.31	57.08	45.95

Table 7: Detailed downstream results on Basque

	STEM	**Humanities**	**Language**	**Social Science**	**Local Culture**
Arabic (ar)					
Random	29.50	28.60	25.80	28.90	32.30
Llama 2 (7B)	33.70	32.65	28.40	32.80	34.70
+ CPT (ar+en)	35.02	35.23	32.36	33.82	35.31
+ CPT (ar)	34.40	35.18	28.24	31.73	33.79
+ CPT (ar+en) (curr)	37.23	34.51	32.35	31.78	36.76
+ CPT w/ EMA	34.48	33.22	31.52	31.34	36.23
+ CPT w/ EMA (curr)	33.21	31.95	29.02	30.13	34.45
Indonesian (id)					
Random	21.90	23.50	24.40	23.40	26.60
Llama 2 (7B)	26.57	26.03	28.47	25.76	26.19
+ CPT (id+en)	28.78	30.85	32.33	31.92	30.19
+ CPT (id)	25.17	26.72	28.23	26.19	27.86
+ CPT (id+en) (curr)	28.64	28.73	30.97	27.39	29.72
+ CPT w/ EMA	28.45	27.67	31.40	27.84	30.21
+ CPT w/ EMA (curr)	27.64	26.37	28.97	26.97	27.38

Table 8: Detailed downstream results in Arabic and Indonesian