Sub-1B Language Models for Low-Resource Languages: Training Strategies and Insights for Basque

Gorka Urbizu, Ander Corral, Xabier Saralegi and **Iñaki San Vicente**
Orai NLP Technologies
{g.urbizu,a.corral,x.saralegi,i.sanvicente}@orai.eus

Abstract

This work investigates the effectiveness of small autoregressive language models (SLMs) with up to one billion parameters (sub-1B) for natural language processing (NLP) tasks in low-resource languages, focusing on Basque. We analyze optimal training strategies by comparing training from scratch and continual pre-training using state-of-the-art SLM architectures. Our analysis considers factors such as model size and the extent of Basque presence in the pre-training corpus. To assess linguistic capabilities, models are evaluated on 12 NLP tasks using the Harness framework. We also conduct a manual evaluation of fine-tuned models on three downstream natural language generation (NLG) tasks: question answering (QA), summarization, and machine translation (MT). Our findings indicate that continual pre-training on a multilingual SLM substantially enhances linguistic performance compared to training from scratch, particularly in low-resource language settings where available corpora typically contain fewer than one billion words. Additionally, the presence of Basque during the pre-training and larger model sizes contribute positively to performance in NLG tasks.

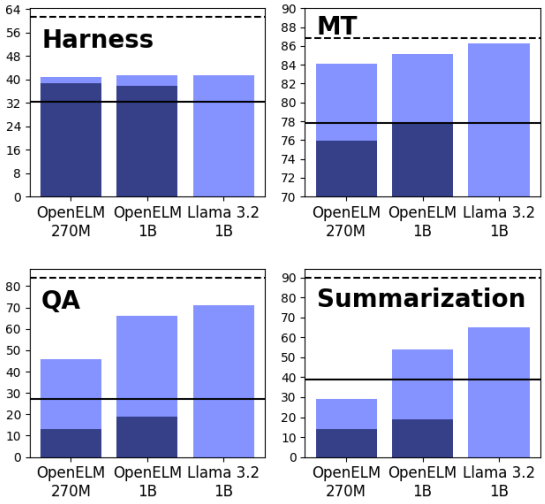

Figure 1: Comparison of from-scratch (dark blue) and continual-trained (light blue) models across 12 tasks in the Harness framework and NLG downstream tasks (QA, MT and summarization). Upperlines (dashed black) show Llama-eus-8B scores, and baselines (solid black) include random guessing for Harness, BART for QA and summarization, and a Transformer-based model for MT. Metrics: accuracy for Harness, correct answers for QA, correct/partially correct for summarisation, and COMET22 for MT.

1 Introduction

In recent years, we have witnessed a growing interest in small language models (SLMs) that can run efficiently on-device with low energy and memory consumption, as well as fast response times, such as MobiLlama (Thawakar et al., 2024), OpenELM (Mehta et al., 2024) or SmolLM2 (Allal et al., 2025). Leading research labs are also releasing smaller versions of their flagship models, namely Llama3.2 1B (Dubey et al., 2024), DeepSeek-R1 1.5B (DeepSeek-AI et al., 2025) and Qwen3-5 0.6B (Qwen-Team, 2025), to reach users and use cases with computational constraints.

This work focuses on SLMs with up to one billion parameters (sub-1B), specifically exploring their effectiveness for performing NLP tasks in low-resource languages, which struggle to collect over 1B word datasets[1], with Basque (see Appendix A for details about the language) as a primary case study. We aim to identify the most effective training strategy for these models. To address this, we investigate several key aspects of the training process using state-of-the-art SLM architectures (Mehta et al., 2024; Dubey et al., 2024): Is it more efficient to train models from scratch, or should we start with pre-trained models from other languages? Does prior exposure to the target language during

[1]FineWeb 2 (Penedo et al., 2024b) covers over 1,000 languages, though only 57 exceed one billion words.

Proceedings of the 5th Workshop on Multilingual Representation Learning (MRL 2025), pages 519–530

pre-training provide any advantages? And how does model size influence performance?

The results of our experiments show the following (see summary in Figure 1):

- When working with limited corpora (∼500M words), continuing training an SLM pre-trained in other languages allows for models with much better linguistic capabilities than those trained from scratch.
- When fine-tuned for Question & Answering (QA), summarization, and machine translation (MT) tasks, continual pre-trained sub-1B models perform notably better than robust baselines based on BART (Lewis, 2019) and Transformer-based sequence-to-sequence models. This advantage is even more remarked if the pre-trained base model has been exposed, even minimally, to Basque.
- The performance gap in NLG tasks between sub-1B continual pre-trained models and the state-of-the-art Llama-eus-8B (Corral et al., 2025) is smallest in MT, followed by QA, and is most pronounced in summarization—reflecting the increasing linguistic complexity required by each task.

As part of the experimentation, the first sub-1B models[2] for Basque have been created, along with fine-tuned versions for QA, summarization, and MT tasks. Furthermore, two new Basque datasets have been developed for QA and summarization tasks, namely CloseBookQA-eu[3] and SAMSUM-eu[4].

2 Model Training

We selected two competitive English-centric models: OpenELM (Mehta et al., 2024) and Llama3.2 (Dubey et al., 2024). Specifically, we included OpenELM-270M, OpenELM-1B, and Llama3.2-1B. This selection enables comparisons between two model sizes (270M vs 1B) and across models with varying levels of exposure to Basque during the pre-training phase (see Table 1).

To determine the most effective training strategy for sub-1B SLMs, we explored two approaches: training models from scratch (Liu et al., 2023; Tonja et al., 2024) and continual pre-training of multilingual models (Cui et al., 2023; Fujii et al., 2024; Kuulmets et al., 2024; Etxaniz et al., 2024; Corral et al., 2025).

From-scratch models —applied exclusively to the OpenELM architecture— were pre-trained on ZelaiHandi (San Vicente et al., 2024), the largest freely available Basque text corpus, comprising 521 million words. Continual models were trained using an 80-20 mix of ZelaiHandi and a FineWeb (Penedo et al., 2024a) subset, following prior works (Fujii et al., 2024; Kuulmets et al., 2024; Corral et al., 2025) to avoid catastrophic forgetting, as English results reported in Appendix F indicate. For full pre-training detail see Appendix C.

Continual pre-trained models retained their base model's tokenizer, while from-scratch models used a native 32K Llama3 tokenizer trained on ZelaiHandi, resulting in a 50% reduction in the tokens-per-word ratio and a tokenization scheme more closely aligned with the morphological structure of Basque (for a more in-depth analysis of the tokenizers see Section 4.3).

3 Analysis of Language Priors in SLMs

In this section, we examine the extent to which the models described in Section 2 were exposed to Basque during pre-training. To this end, we compare the base and continual 1B versions of Llama-3.2-1B and OpenELM-1B.

Following previous work (Wang et al., 2024), we sampled 512 generations per model using only the beginning-of-sequence token as input. To examine how language priors shift with stronger language cues, we also assessed model generations when provided with partial prompts of varying lengths in Basque[5]. Each generation had a maximum length of 300 tokens and was produced with a temperature of 1.0. The primary language of each generated sequence was determined automatically using FastText (Joulin et al., 2016).

Table 1 presents the results of our analysis of language priors in SLMs. Our findings show that both of the base versions of Llama-3.2-1B and OpenELM-1B exhibit a strong bias toward English in their zero-word cue generations, with English accounting for over 90% of outputs. When given a Basque word as a cue, these models still generate predominantly English text, with only a slight increase in Basque output in the

[2] hf.co/collections/orai-nlp/slms-for-basque
[3] hf.co/datasets/orai-nlp/ClosedBookQA-eu
[4] hf.co/datasets/orai-nlp/SAMSUM-eu

[5] Document beginnings from the ZelaiHandi validation set up to 2 words.

Model	EU Cue	EN	EU	Oth
Llama 1B base	0-words	**94.1**	0.0	5.9
	1-words	**87.3**	4.7	8.0
	2-words	**62.9**	21.7	15.4
OpenELM 1B base	0-words	**90.2**	0.0	9.8
	1-words	**94.9**	0.0	5.1
	2-words	**79.1**	1.4	19.5
Llama 1B cont.	0-words	13.5	**85.5**	1.0
	1-words	7.7	**89.6**	2.7
	2-words	5.5	**91.4**	3.1
OpenELM 1B cont.	0-words	8.6	**91.0**	0.4
	1-words	2.9	**96.3**	0.8
	2-words	2.9	**94.9**	2.1

Table 1: Analysis of language priors in SLMs (base and continual), showing the percentage of generations classified as English (EN), Basque (EU), or others.

case of Llama-3.2-1B. As more words are added to the prompt, Basque output increases, though OpenELM-1B remains notably less responsive to Basque cues than Llama-3.2-1B. These results highlight that OpenELM-1B has been exposed to less Basque data during pre-training, which likely contributes to its lower responsiveness to Basque cues. This suggests that, in theory, Llama-3.2-1B is a more suitable candidate for continual pre-training, as its initial exposure to Basque provides a stronger foundation for further adaptation.

In contrast, the continually pre-trained models exhibit a strong bias (over 85%) toward Basque in the zero-word cue generations, which further increases when Basque cue words are provided.

4 Evaluation

We conducted an intrinsic evaluation of the linguistic competences of both from-scratch and continual pre-trained models and compared them to the original model using the Harness evaluation framework. While Harness offers an automatic and cost-effective method for assessing the potential linguistic performance of SLMs, it does not fully reflect real-world performance, as the scores are based on the system selecting the most appropriate answers from multiple-choice questions. To better evaluate the models in realistic settings, we also fine-tuned and manually evaluated them on downstream NLG tasks. In addition, we explore how native tokenizers contribute to more efficient and linguistically aligned from-scratch models.

4.1 Intrinsic Evaluation of Linguistic Abilities

To evaluate models' linguistic competences in Basque, we employed a variety of existing benchmarks including language proficiency, reading comprehension, general knowledge and commonsense reasoning tasks: ARC_eu, Winogrande_eu, MMLU_eu and HellaSwag_eu (Corral et al., 2025); BL2MP (Urbizu et al., 2024); BasqueGLUE (Urbizu et al., 2022); Belebele (Bandarkar et al., 2024); X-StoryCloze (Lin et al., 2021a); EusProficiency, EusReading, EusExams, and EusTrivia (Etxaniz et al., 2024). Evaluations were carried out with the LM Evaluation Harness framework (Gao et al., 2024), following an in-context few-shot setup as in previous work (Etxaniz et al., 2024; Corral et al., 2025). Results are shown in Table 2.

The OpenELM base models perform below random chance, likely due to limited exposure to Basque data during pretraining. The Llama-3.2-1B base model performs slightly better than random, indicating that its exposure to Basque data, though minimal, offers some advantage (see Section 3).

When trained from scratch, all OpenELM models outperform their base counterparts. However, these from-scratch models often perform at random levels across many tasks. Notably, the OpenELM-270M from-scratch model achieves the highest overall performance, which might indicate that a 1B model could struggle to generalize effectively with a modest 521M-word Basque training dataset due to its larger parameter size.

Substantial improvements are observed with continual pre-training across all models, with the 1B-parameter models performing comparably, while the 270M model lags behind. Continual pre-training consistently outperforms from-scratch pre-training, especially in the 1B model, suggesting that the available Basque training data —similar to other low-resource languages— is insufficient for from-scratch pre-training and leveraging multilingual pre-training through continual training proves to be more effective.

While there remains a performance gap of approximately 20 points between the continual pre-trained variants and Llama-eus-8B, the results are consistent with expectations from scaling laws (Hoffmann et al., 2022), highlighting the strong capabilities of smaller models given their size.

Model		Bl2mp	Arc	WnGr.	Mmlu	HSwg	Beleb.	XStrC.	Exams	Prof.	Read.	Trivia	BGlue	Avg.
Random		*50.0*	*25.0*	*50.0*	*25.0*	*25.0*	*25.0*	*50.0*	*25.0*	*25.0*	*25.8*	*26.6*	*37.5*	*32.5*
OpenELM 270M	base	44.7†	26.0	47.6†	25.6	28.0	26.0	50.1	25.3	25.0	22.4†	26.2†	36.2†	31.9†
	scratch	88.1	32.0	47.2†	27.8	40.0	27.9	55.7	24.9†	24.0†	25.3†	27.4	38.6	38.2
	continual	89.9	33.6	53.2	23.3†	45.2	27.9	55.3	25.0	24.7†	**30.4**	27.1	41.0	39.7
OpenELM 1B	base	46.2†	24.4†	41.2†	25.2	27.6	**28.1**	49.8†	25.7	24.8†	23.3†	26.4†	37.8	31.7†
	scratch	87.2	28.8	47.6†	25.2	40.4	25.4	54.1	24.5†	24.1†	24.2†	26.2†	37.9	37.1
	continual	**90.4**	**42.0**	55.6	25.9	**48.0**	26.3	**60.4**	26.2†	24.4†	28.7	26.2	**42.7**	**41.4**
Llama 1B	base	49.1†	29.6	52.0	26.7	24.4†	27.9	50.0	26.5	23.8†	25.3†	**28.6**	38.4	33.5
	continual	88.9	**42.0**	**56.8**	**28.5**	46.4	27.9	60.2	**27.1**	**25.5**	23.3†	28.2	41.4	**41.4**
Llama-eus 8B		*89.2*	*55.2*	*67.2*	*53.3*	*63.6*	*73.4*	*65.7*	*52.5*	*48.4*	*54.6*	*56.2*	*55.3*	*61.2*

Table 2: Results from the intrinsic evaluation of linguistic abilities, conducted using 5 in-context examples for most tasks, except for HellaSwag (10-shot), ARC (25-shot), BL2MP (0-shot), X-StoryCloze (0-shot), and EusReading (1-shot). The best-performing model is highlighted in bold, and † denotes models performing below random guess.

4.2 Downstream NLG Tasks Evaluation

We further assessed our models by fine-tuning them on three downstream NLG tasks of varying difficulty—ordered from most to least challenging: summarization, QA (including both hard and factoid questions), and English-to-Basque MT. Fine-tuning details are provided in Appendix D.

To address the lack of task-specific Basque datasets, we constructed training data for the QA and summarization tasks. For QA, we constructed CloseBookQA-eu based on the Belebele-eus MCQA dataset (Bandarkar et al., 2024), and enriched it with translated examples from the MCTest MCQA dataset (Richardson et al., 2013) as well as semi-automatically generated factoid questions derived from news articles. For summarization, we automatically translated the SAMSUM dataset (Gliwa et al., 2019). In the case of the MT task, we compiled a 2M-sentence English-Basque parallel dataset from OPUS (Tiedemann, 2009). Appendix D.1 offers further details on dataset creation.

Evaluation methodologies varied by task. For QA-easy (factoid questions), QA-hard and summarization, a native Basque speaker from our team manually evaluated a random test set of 100 examples per task. QA responses were deemed correct or incorrect, while summarization outputs were rated correct, partially correct, or incorrect. For MT, evaluation was performed by computing the COMET22 (Rei et al., 2022) metric on the Flores-200 benchmark (Team et al., 2024).

Regarding baselines, we fine-tuned a monolingual BART for the QA and summarization tasks, and trained a Transformer-based model for MT (see Appendix B for further details). Additionally, we fine-tuned Llama-eus-8B (Corral et al., 2025) on downstream tasks to establish the upper bound performance for each task.

Model		QA	Sum	MT
Baseline		42|12	19 (39)	77.8
OpenELM 270M	scratch	17|09	06 (14)	75.9
	continual	64|28	14 (29)	84.1
OpenELM 1B	scratch	31|07	06 (19)	77.8
	continual	84|48	37 (54)	85.1
Llama 1B	continual	**88|54**	**39 (65)**	**86.3**
Llama-eus 8B		*95|73*	*60 (90)*	*86.8*

Table 3: Results on the downstream NLG tasks of QA, Summarization and MT). The QA task is formed by two datasets of different difficulty (QA-easy|QA-hard). Scores in parentheses for the summarization task indicate the sum of correct and partially correct outputs.

Table 3 presents the results of the fine-tuned models on downstream tasks. For the continual pre-trained models, performance differences across architectures and sizes align with the level of language understanding required by each task, with larger gaps observed in more complex tasks—ordered from most to least complex: QA-hard, summarization, QA-easy, and machine translation. Notably, Llama 1B consistently outperforms OpenELM 1B, highlighting the benefits of Basque-specific priors discussed in Section 3. In line with scaling laws (Hoffmann et al., 2022), performance improves with model size, with 1B models outperforming their 270M counterparts and the 8B model achieving the highest overall gains.

Following the trend in Section 4.1, from-scratch models fail to surpass continual pre-trained ones, reinforcing the importance of leveraging prior linguistic knowledge through continual pre-training. Despite the potential benefits of a

Tokenizer	Vocab.	TPW	Morph.
OpenELM	32K	3.23	0.12
Llama3	128K	2.95	0.20
Native	32K	1.60	0.41

Table 4: Vocabulary size, tokens-per-word (TPW) ratio and morphological alignment score of different tokenizers used by our models.

Tokenizer	Vocab.	TPW	Morph.	Harness Avg.
Original	128K	2.90	0.20	35.49
Native	128K	1.43	0.56	35.04

Table 5: Vocabulary size, tokens-per-word (TPW) ratio and morphological alignment score for each tokenizer of equal size, with average results for from-scratch llama3.2 1B models on the Harness evaluation.

native Basque tokenizer, the results indicate it does not offer a significant advantage in making from-scratch models competitive. Section 4.3 analyzes the impact of native tokenizers and shows that, although they improve the tokens-per-word ratio and better align with Basque morphology, they do not lead to a significant improvement in linguistic performance.

4.3 Impact of Native Tokenizers

One potential advantage of training a model from scratch is the ability to use a native tokenizer fully adapted to the target language. This results in a lower tokens-per-word ratio, which implies shorter sequence lengths to process the same word sequence, leading to faster and more memory-efficient models.

As stated in Section 2, continually pre-trained models retain their base's tokenizer. In contrast, our from-scratch models use a new native 32K Llama3 tokenizer trained on ZelaiHandi (San Vicente et al., 2024), resulting in a 50% reduction in the tokens-per-word ratio[6], as shown in Table 4.

Furthermore, a native tokenizer is expected to align more closely with morpheme boundaries, which might be beneficial for morphologically rich languages like Basque. To evaluate this morphological alignment, we compare the tokenized subwords with the expected lemma-morpheme boundaries. A score of 1 is assigned if the first subword matches the lemma.

We conduct this evaluation using a dataset of 100K sentences (over 1M words), which have been automatically annotated[7] with lemma and morpheme boundaries—e.g., *Brasil da aurten|**go** herrialde gonbidatu|**a***—extracted from the 5M-word Basque corpus defined by Urbizu et al. (2024). This corpus comprises news articles and Wikipedia articles, offering a representative sample of real-world Basque usage.

As shown in Table 4, native tokenizers achieve higher morphological alignment scores. However, results from Sections 4.1 and 4.2 indicate that this alignment advantage does not yield sufficient performance gains for scratch-trained models to match those continually pre-trained with suboptimal multilingual tokenizers. The performance gap is especially pronounced in downstream NLG tasks.

To more precisely assess the impact of using a native tonenizer versus the English-centric tokenizer, we trained two additional Llama 3.2-1B models from scratch: one using the original 128k-token vocabulary and the other using a native tokenizer of equivalent size[8].

Table 5 shows the Harness evaluation results for models trained from scratch. Although the native tokenizer provides better morphological alignment and achieves greater compression—as evidenced by a lower tokens-per-word ratio—it does not lead to improved linguistic performance compared to the original tokenizer. This suggests that, for Basque and with a training corpus of around 500 million words, a native tokenizer does not necessarily enhance the model's linguistic competence. This finding holds despite Basque's morphological complexity, particularly its rich system of case endings, and is consistent with the results reported by Urbizu et al. (2024).

5 Conclusions

This work examines the effectiveness of SLMs with up to 1B parameters for NLP tasks in low-resource languages, focusing on Basque. Our findings show that continual pre-training of a multilingual SLM notably improves performance compared to training from scratch, with larger model sizes and the presence of Basque during pre-training further enhancing the performance on NLG tasks.

[6]Calculated on the ZelaiHandi validation set.

[7]Using an Apertium-based custom implementation.

[8]With same training procedure of the native 32K tokenizer.

Limitations

Basque has been chosen as a case study, as it is an isolated language with complex morphology, and a corpus of 521 million words has been used for training. We consider this scenario to be representative of a significant number of low-resource languages. However, extending these conclusions to other languages may require additional experiments that account for their specific linguistic characteristics and level of digital development.

For the construction of the SLMs, we explored training strategies both from scratch and based on continual pre-training. In some languages, developing SLMs using knowledge distillation strategies could be of interest, and we leave this analysis for future work.

The evaluation of SLMs on downstream tasks has been limited to three representative tasks: QA, summarization, and MT. Training for these tasks was conducted using datasets of a fixed size. In future work, we aim to extend this study to additional downstream tasks and analyze the impact of dataset size on the fine-tuning process for each task.

Ethical Concerns

The outputs of the SLMs trained for this work may show undesired biases and produce offensive language. Although the Basque text sources gathered to pre-train the SLMs were selected by hand, they contain bad words from fictional sources and social biases that were not handled here. These aspects must be analyzed and treated before building applications that interact with final users.

Acknowledgments

This work has been partially funded by the Basque Government (ICL4LANG project, grant no. KK-2023/00094) and the European Union (EFA 104/01-LINGUATEC IA project, INTERREG POCTEFA 2021-2027 program). Pre-training and fine-tuning of SLMs were conducted using the Hyperion system at the Donostia International Physics Center (DIPC). We also acknowledge the support of Google's TFRC program for pre-training the BART baseline on TPUs. Finally, we thank Idoia Davila Uzkudun for her contributions to manual data curation and evaluation.

References

Loubna Ben Allal, Anton Lozhkov, Elie Bakouch, Gabriel Martín Blázquez, Guilherme Penedo, Lewis Tunstall, Andrés Marafioti, Hynek Kydlíček, Agustín Piqueres Lajarín, Vaibhav Srivastav, et al. 2025. Smollm2: When smol goes big–data-centric training of a small language model. *arXiv preprint arXiv:2502.02737*.

Lucas Bandarkar, Davis Liang, Benjamin Muller, Mikel Artetxe, Satya Narayan Shukla, Donald Husa, Naman Goyal, Abhinandan Krishnan, Luke Zettlemoyer, and Madian Khabsa. 2024. The belebele benchmark: a parallel reading comprehension dataset in 122 language variants. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 749–775, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

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

Ander Corral, Ixak Sarasua Antero, and Xabier Saralegi. 2025. Pipeline analysis for developing instruct LLMs in low-resource languages: A case study on Basque. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 12636–12655, Albuquerque, New Mexico. Association for Computational Linguistics.

Yiming Cui, Ziqing Yang, and Xin Yao. 2023. Efficient and effective text encoding for chinese llama and alpaca. *arXiv preprint arXiv:2304.08177*.

Tri Dao, Daniel Y. Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. 2022. Flashattention: Fast and memory-efficient exact attention with io-awareness. *Preprint*, arXiv:2205.14135.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Meng Li,

Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Shengfeng Ye, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wanjia Zhao, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanhong Xu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv 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. *arXiv preprint arXiv:2404.17790*.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. 2024. A framework for few-shot language model evaluation.

Bogdan Gliwa, Iwona Mochol, Maciej Biesek, and Aleksander Wawer. 2019. Samsum corpus: A human-annotated dialogue dataset for abstractive summarization. *EMNLP-IJCNLP 2019*, page 70.

Sylvain Gugger, Lysandre Debut, Thomas Wolf, Philipp Schmid, Zachary Mueller, Sourab Mangrulkar, Marc Sun, and Benjamin Bossan. 2022. Accelerate: Training and inference at scale made simple, efficient and adaptable. https://github.com/huggingface/accelerate.

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

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. 2022. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*.

Hele-Andra Kuulmets, Taido Purason, Agnes Luhtaru, and Mark Fishel. 2024. Teaching llama a new language through cross-lingual knowledge transfer. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3309–3325, Mexico City, Mexico. Association for Computational Linguistics.

Alexandre Lacoste, Alexandra Luccioni, Victor Schmidt, and Thomas Dandres. 2019. Quantifying the carbon emissions of machine learning. *arXiv preprint arXiv:1910.09700*.

Mike Lewis. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, Ramakanth Pasunuru, Sam Shleifer, Punit Singh Koura, Vishrav Chaudhary, Brian O'Horo, Jeff Wang, Luke Zettlemoyer, Zornitsa Kozareva, Mona T. Diab, Veselin Stoyanov, and Xian Li. 2021a. Few-shot learning with multilingual language models. *CoRR*, abs/2112.10668.

Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, Ramakanth Pasunuru, Sam Shleifer, Punit Singh Koura, Vishrav Chaudhary, Brian O'Horo, Jeff Wang, Luke Zettlemoyer, Zornitsa Kozareva, Mona T. Diab, Veselin Stoyanov, and Xian Li. 2021b. Few-shot learning with multilingual language models. *CoRR*, abs/2112.10668.

Peng Liu, Lemei Zhang, Terje Farup, Even W Lauvrak, Jon Espen Ingvaldsen, Simen Eide, Jon Atle Gulla, and Zhirong Yang. 2023. Nlebench+ norglm: A comprehensive empirical analysis and benchmark dataset for generative language models in norwegian. *arXiv preprint arXiv:2312.01314*.

Sachin Mehta, Mohammad Hossein Sekhavat, Qingqing Cao, Maxwell Horton, Yanzi Jin, Chenfan Sun, Seyed Iman Mirzadeh, Mahyar Najibi, Dmitry Belenko, Peter Zatloukal, et al. 2024. Openelm: An efficient language model family with open training and inference framework. In *Workshop on Efficient Systems for Foundation Models II@ ICML2024*.

Christopher Moseley, editor. 2010. *Atlas of the World's Languages in Danger*, 3rd edition. UNESCO Publishing, Paris.

OpenAI. 2024. Gpt-4o: Openai's new flagship model. Accessed: 2025-05-19.

Guilherme Penedo, Hynek Kydlíček, Loubna Ben Allal, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro Von Werra, and Thomas Wolf. 2024a. The fineweb datasets: Decanting the web for the finest text data at scale. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Guilherme Penedo, Hynek Kydlíček, Vinko Sabolčec, Bettina Messmer, Negar Foroutan, Martin Jaggi, Leandro von Werra, and Thomas Wolf. 2024b. Fineweb2: A sparkling update with 1000s of languages.

Qwen-Team. 2025. Qwen3.

Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase, and Yuxiong He. 2020. Zero: Memory optimizations toward training trillion parameter models. In *SC20: International Conference for High Performance Computing, Networking, Storage and Analysis*, pages 1–16. IEEE.

Gema Ramírez-Sánchez, Jaume Zaragoza-Bernabeu, Marta Bañón, and Sergio Ortiz-Rojas. 2020. Bifixer and bicleaner: two open-source tools to clean your parallel data. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 291–298, Lisboa, Portugal. European Association for Machine Translation.

Ricardo Rei, José GC De Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André FT Martins. 2022. Comet-22: Unbabel-ist 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585.

Matthew Richardson, Christopher J.C. Burges, and Erin Renshaw. 2013. MCTest: A challenge dataset for the open-domain machine comprehension of text. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 193–203, Seattle, Washington, USA. Association for Computational Linguistics.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2020. Winogrande: An adversarial winograd schema challenge at scale. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 8732–8740. AAAI Press.

Iñaki San Vicente, Gorka Urbizu, Ander Corral, Zuhaitz Beloki, and Xabier Saralegi. 2024. Zelaihandi: A large collection of basque texts.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

NLLB Team et al. 2024. Scaling neural machine translation to 200 languages. *Nature*, 630(8018):841.

Omkar Thawakar, Ashmal Vayani, Salman Khan, Hisham Cholakal, Rao M Anwer, Michael Felsberg, Tim Baldwin, Eric P Xing, and Fahad Shahbaz Khan. 2024. Mobillama: Towards accurate and lightweight fully transparent gpt. *arXiv preprint arXiv:2402.16840*.

Jörg Tiedemann. 2009. News from opus - a collection of multilingual parallel corpora with tools and interfaces. In N. Nicolov, K. Bontcheva, G. Angelova, and R. Mitkov, editors, *Recent Advances in Natural Language Processing*, volume V, pages 237–248. John Benjamins Publishing Company.

Atnafu Lambebo Tonja, Bonaventure FP Dossou, Jessica Ojo, Jenalea Rajab, Fadel Thior, Eric Peter Wairagala, Aremu Anuoluwapo, Pelonomi Moiloa, Jade Abbott, Vukosi Marivate, et al. 2024. Inkubalm: A small language model for low-resource african languages. *arXiv preprint arXiv:2408.17024*.

Gorka Urbizu, Iñaki San Vicente, Xabier Saralegi, Rodrigo Agerri, and Aitor Soroa. 2022. Basqueglue: A natural language understanding benchmark for basque. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1603–1612.

Gorka Urbizu, Muitze Zulaika, Xabier Saralegi, and Ander Corral. 2024. How well can BERT learn the grammar of an agglutinative and flexible-order language? the case of Basque. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 8334–8348, Torino, Italia. ELRA and ICCL.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6000–6010, Red Hook, NY, USA. Curran Associates Inc.

Jiayi Wang, Yao Lu, Maurice Weber, Max Ryabinin, Yihong Chen, Raphael Tang, and Pontus Stenetorp. 2024. Multilingual pretraining using a large corpus machine-translated from a single source language. *arXiv preprint arXiv:2410.23956*.

Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R. Bowman. 2020. BLiMP: The benchmark of linguistic minimal pairs for English. *Transactions of the Association for Computational Linguistics*, 8:377–392.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

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

A Basque Language

Basque is a language with roughly 810K fluent speakers in the region of the Basque Country, spanning northern Spain and southwestern France. It is currently classified as vulnerable according to The UNESCO Atlas of the World's Languages in Danger (Moseley, 2010). Basque is a language isolate (unrelated to any other known languages) and uses the Latin script. It is a morphologically rich language, with a flexible word order and follows an ergative–absolutive syntactic alignment. Despite being low-resource in terms of corpora (< 1B words), Basque does have annotated datasets for a number of NLU and NLG tasks, thanks to the effort of a strong local NLP community.

B Baselines

B.1 BART

The monolingual BART base model (139M parameters), used as a baseline for question answering (QA) and summarization tasks, was pre-trained on the ZelaiHandi corpus (San Vicente et al., 2024). It employs a Byte-Pair Encoding (BPE) tokenizer with a 50K token vocabulary, which was also trained on ZelaiHandi.

The model was trained for 154 epochs (equivalent to 1,460K steps) with a batch size of 32, a learning rate of 1e-4, and a sequence length of 512 tokens. The final checkpoint was retained as it achieved the best performance based on validation loss. We used the Flax implementation of BART from the Hugging Face Transformers library (Wolf et al., 2020) and pre-trained the model on a single TPUv3-8 node for one week.

B.2 MT Baseline

The Baseline MT system was trained using the sequence-to-sequence Transformer architecture (Vaswani et al., 2017) as implemented in the Eole Toolkit[9] with the default configuration (6 layers, 1024 size vectors). We apply BPE tokenization (Sennrich et al., 2016) learned on 32,000 merge operations on the joint training parallel data. The training corpus comprises of 2.2M parallel sentences gathered from various sources from the Opus collection (Tiedemann, 2009). The model was trained for 230K steps (early stopping after 10 validation steps, validating each 10K steps). Validation is done over 8K parallel sentences composed of the Flores benchmark validation dataset and 5K sentences excluded from the training data. Training was carried out on a single Nvidia RTX A5000 GPU.

C Pre-Training Details

From-scratch models were pre-trained for up to 25 epochs, while continual models were further pre-trained for up to 5 epochs. In both cases, we selected the best-performing checkpoint according to validation loss for the final model. Our OpenELM and Llama models, based on the architectures of OpenELM[10] (Mehta et al., 2024) and Llama3.2[11] (Dubey et al., 2024), have a maximum sequence length of 2048 and 4096, respectively. OpenELM models were pre-trained with a cosine learning rate of 3e-5 and an effective batch size of around 4M following the configuration of OpenELM-270M (Mehta et al.,

[9] https://eole-nlp.github.io/eole/

[10] Licensed under Apple Sample Code License

[11] Licensed under Llama3.2 Community License Agreement

Model	Size	GPU time	kgCO_2eq
OE270M-s	270M	282h	30.46
OE270M-c	270M	138h	14.90
OE1B-s	1.1B	571h	61.67
OE1B-c	1.1B	290h	31.32
LL1B-c	1B	122h	13.18

Table 6: Carbon footprint of pre-training our models. Llama3-1B is more efficient and emitted less CO_2 due to the available flash attention implementation. OE = OpenELM, LL = LLama3. s = scratch. c = continual.

2024). Llama3-1B was further trained with a cosine learning rate of 1e-4 and an effective batch size of around 2M following the configuration of LLama-eus-8B (Corral et al., 2025).

Pre-training LMs involves computationally intensive experiments that contribute significantly to carbon emissions. For efficient large-scale pre-training, we opted for the Hugging Face Transformers (Wolf et al., 2020) library, alongside DeepSpeed ZeRO (Rajbhandari et al., 2020) and Accelerate (Gugger et al., 2022). Flash Attention (Dao et al., 2022) was only available for Llama3 models since OpenELM does not have it implemented on Transformers.

The training was conducted on NVIDIA A100 80GB GPUs (1-8). We provide details on model size, compute hours, and carbon emissions for our experiments in Table 6. Carbon emissions were estimated using the Machine Learning Impact calculator[12] (Lacoste et al., 2019).

D Fine-Tuning Details

Each foundational model was fine-tuned for up to five epochs, independently on each task (QA, summarization, and MT). To ensure optimal performance, we selected the checkpoint with the lowest validation loss.

We used a batch size of 32 and a learning rate of 3e-5. However, for certain models where the validation loss curve showed instability—collapsing before completing the first epoch—, we reran fine-tunings, reducing the learning rate until achieving a run with a stable validation loss trajectory.

We fine-tuned the BART model on QA and summarization with the same batch size, learning rate and epochs as the rest of the models, selecting the best-performing checkpoint on validation loss. The transformer baseline on MT was trained from scratch (see Appendix B.2).

D.1 Downstream Datasets

ClosedBookQA. For question answering (QA), we constructed ClosedBookQA-eu, a closed-book QA dataset derived from three sources: the MCQA Belebele-eus dataset[13] (Bandarkar et al., 2024), the MCTest dataset[14] (Richardson et al., 2013), and semi-automatically generated examples based on news content.

Belebele is a multiple-choice QA (MCQA) dataset that includes a passage (context), a question, and four possible answers. Although a Basque version of Belebele is available, it only provides a test set of 900 examples. To adapt it for a generative QA setting, we extracted passage–question–answer triplets and discarded examples that were unanswerable[15] without the full set of answer choices. After filtering, we retained 573 usable examples, which we split into 423 for training, 50 for validation, and 100 for the QA-hard test set.

To further expand the training data, we incorporated MCTest, which contains 2,000 MCQA examples. These were translated into Basque using a proprietary document-level machine translation system based on Llama-eus-8B (Corral et al., 2025). After manually filtering out translation errors, 1,962 examples were retained. The final training set thus comprised 2,385 examples.

In addition to the QA-hard test set derived from Belebele, we created a complementary QA-easy test set of 100 simpler factoid questions. This set was generated using GPT-4o (OpenAI, 2024) in a two-step process: first, selecting passages from 100 Basque news articles not included in ZelaiHandi, and second, generating corresponding questions and answers. All examples were manually reviewed, corrected, and refined by a native Basque speaker to ensure both linguistic quality and appropriate difficulty.

Summarization. For summarization, there is no publicly available summarization dataset in Basque. To address this, we automatically translated SAMSum[16] (Gliwa et al., 2019), a

[12] https://mlco2.github.io/impact#compute

[13] Licensed under CC-BY-SA 4.0

[14] Licensed under Microsoft Research License

[15] E.g., "Which of these is true?" or "Which option is not mentioned?"

[16] Licensed under CC-BY-NC-ND 4.0

human-annotated dialogue dataset for abstractive summarization, using a proprietary document-level MT system based on Llama-eus-8B. We then filtered out examples with incomplete translations or non-Basque outputs.

The translated test set was further refined by a native speaker to obtain 100 high-quality, manually curated test examples. In total, we obtained 11,313 training examples, 636 validation examples, and 100 manually curated test examples for evaluation.

Machine translation. In the case of the MT task, we compiled an English-Basque dataset gathered from various sources in OPUS[17] (Tiedemann, 2009). The final corpus contains a 2.2M parallel sentences, obtained after applying rule-based cleaning, and used BiCleaner (Ramírez-Sánchez et al., 2020) with a threshold of 0.9.

E Harness Benchmarks for Basque

To assess our model's performance in Basque, we utilized a range of existing benchmarks:

- **ARC_HT_eu_sample** (Corral et al., 2025): A subset of 250 samples manually translated to Basque from the ARC dataset (Clark et al., 2018). The ARC dataset consists of genuine grade-school level, multiple-choice science questions.
- **Winogrande_HT_eu_sample** (Corral et al., 2025): A subset of 250 samples manually translated to Basque from the WinoGrande dataset (Sakaguchi et al., 2020). WinoGrande is a dataset of 44k problems specifically designed to test commonsense reasoning.
- **MMLU_HT_eu_sample** (Corral et al., 2025): A subset of 270 samples manually translated to Basque from the MMLU dataset (Hendrycks et al., 2021). The MMLU dataset is a massive multitask test consisting of multiple-choice questions from various branches of knowledge. The test spans subjects in the humanities, social sciences, hard sciences, and other areas that are important for some people to learn.
- **HellaSwag_HT_eu_sample** (Corral et al., 2025): A subset of 250 samples manually translated to Basque from the HellaSwag dataset (Zellers et al., 2019). The HellaSwag dataset commonsense NLI evaluation benchmark.
- **BL2MP** (Urbizu et al., 2024): The BL2MP test set is designed to assess the grammatical knowledge of language models in the Basque language, inspired by the BLiMP (Warstadt et al., 2020) benchmark.
- **BasqueGLUE** (Urbizu et al., 2022): BasqueGLUE is an NLU benchmark for Basque, which has been elaborated from previously existing datasets and following similar criteria to those used for the construction of GLUE and SuperGLUE.
- **Belebele** (Bandarkar et al., 2024): Belebele is a multiple-choice machine reading comprehension dataset spanning 122 language variants.
- **X-StoryCloze** (Lin et al., 2021b): XStoryCloze consists of the professionally translated version of the English StoryCloze dataset to 10 non-English languages. It is a commonsense reasoning framework for evaluating story understanding, story generation, and script learning.
- **EusProficiency**, **EusReading**, **EusExams**, and **EusTrivia** (Etxaniz et al., 2024): Basque-specific benchmarks covering proficiency tests based on past EGA exams (C1 level Basque), reading comprehension, public service exam preparation, and trivia questions, respectively.

This comprehensive evaluation approach enables us to measure the model's capabilities across various tasks, providing a thorough understanding of its formal and functional competencies in Basque.

F English Results on Harness

The continual models were trained using an 80-20 mix of ZelaiHandi and a FineWeb (Penedo et al., 2024a) subset, following prior works (Fujii et al., 2024; Kuulmets et al., 2024; Corral et al., 2025) to avoid catastrophic forgetting. Thus, they are expected to retain some English knowledge from the pretraining. To measure English linguistic abilities of the continual models and see if they are kept from the base model, we evaluated base and continual versions of OpenELM-1B and Llama-1B

[17] Includes data licensed under various open licenses.

Model		Arc	WnGr.	Mmlu	HSwg	Beleb.	XStrC.	Avg.
Random		*25.0*	*50.0*	*25.0*	*25.0*	*25.0*	*50.0*	*32.5*
OpenELM-1B	base	53.2	**70.8**	**30.7**	**67.6**	27.7	**72.1**	**53.2**
	continual	45.6	58.0	24.4	56.8	27.2	68.6	46.8
Llama 1B	base	52.8	68.4	28.1	45.2	**34.7**	71.3	50.5
	continual	**53.6**	66.4	23.7	64.8	30.4	71.1	51.7

Table 7: Results from the intrinsic evaluation of linguistic abilities on the English counterparts of the datasets used for Basque, conducted using 5 in-context examples for most tasks, except for HellaSwag (10-shot), ARC (25-shot) and X-StoryCloze (0-shot). The best-performing model is highlighted in bold.

in the English versions of the subsets of several NLU tasks used to evaluate the models in Basque in Section 4.1, described in Appendix E.

The results for English are shown in Table 7. It shows that the results of the base models hold after continual training for Basque, with a few exceptions, proving that models kept most of their English linguistic abilities, and the 80-20 corpora approach is successful at avoiding catastrophic forgetting. When we compare both architectures, while Llama 1B holds its results in overall, there is a small drop in the case of OpenELM, which might be caused by the lack of prior exposure to Basque.