Latxa: An Open Language Model and Evaluation Suite for Basque

Julen Etxaniz* **Oscar Sainz*** **Naiara Perez*** **Itziar Aldabe** **German Rigau**
Eneko Agirre **Aitor Ormazabal** **Mikel Artetxe** **Aitor Soroa**
HiTZ Center - Ixa, University of the Basque Country UPV/EHU
{julen.etxaniz,a.soroa}@ehu.eus

Abstract

We introduce Latxa, a family of large language models for Basque ranging from 7 to 70 billion parameters. Latxa is based on Llama 2, which we continue pretraining on a new Basque corpus comprising 4.3M documents and 4.2B tokens. Addressing the scarcity of high-quality benchmarks for Basque, we further introduce 4 multiple choice evaluation datasets: EusProficiency, comprising 5,169 questions from official language proficiency exams; EusReading, comprising 352 reading comprehension questions; EusTrivia, comprising 1,715 trivia questions from 5 knowledge areas; and EusExams, comprising 16,774 questions from public examinations. In our extensive evaluation, Latxa outperforms all previous open models we compare to by a large margin. In addition, it is competitive with GPT-4 Turbo in language proficiency and understanding, despite lagging behind in reading comprehension and knowledge-intensive tasks. Both the Latxa family of models, as well as our new pretraining corpora and evaluation datasets, are publicly available under open licenses.[1] Our suite enables reproducible research on methods to build LLMs for low-resource languages.

1 Introduction

Motivated by their increasing training cost and commercial interest, the development of Large Language Models (LLMs) has been led by close initiatives like GPT (OpenAI et al., 2023), Claude (Wu et al., 2023) and Gemini (Team, 2023). In recent times, a more open ecosystem has emerged following the release of various competitive models like Llama 2 (Touvron et al., 2023) and Mistral (Jiang et al., 2024). However, despite early efforts to build open multilingual models (Lin et al., 2022; Scao et al., 2023), the most competitive ones are notoriously English-centric. As shown in Table 1, all these open models perform poorly in low-resource languages like Basque, with most results marginally surpassing random chance.

In this work, we present Latxa, an open family of LLMs for Basque that substantially outperforms all these previous models. Basque is an agglutinative language written in Latin script and with no known relatives, although a significant part of the vocabulary is shared with contact languages like Spanish and French. Basque is the 52th language in Common Crawl, with 0.035% of the total content –for reference, English is the 1st language with 46% of the content and Spanish is the 5th with 4.6%. Our work builds on various open resources and models that we further expand to Basque, highlighting the importance of an open ecosystem for the development of language technology for low-resource languages. In particular, our models are based on Llama 2, which we continue training in Basque using a new corpus with 4.3M documents from 4 existing and 3 new sources. In addition, we release 4 diverse and challenging multiple-choice benchmarks comprising a total of 23,282 questions, covering language proficiency, reading comprehension, trivia questions, and public examinations.

As shown in Table 1, Latxa performs substantially better than all existing open models, with the 70B variant outperforming the previous best open model (Yi 34B) by 18.95 points in average. In addition, it also outperforms the Llama 2 model it is based on by 25.18 points, and it is also superior to GPT-3.5 Turbo in all datasets we evaluate on. Interestingly, our best model also outperforms GPT-4 Turbo in language proficiency exams (EusProf), despite lagging behind in reading comprehension and knowledge-intensive tasks. This suggests that the capabilities that an LLM exhibits in a given language are not determined by its linguistic competence in this particular language, opening the doors to further improvements in low-resource LLMs as

*Equal contribution.
[1] https://github.com/hitz-zentroa/latxa

Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 14952–14972

		XStory	Belebele	BasGLUE	[new] EusProf	[new] EusRead	[new] EusTrivia	[new] EusExams	Avg
Random		50.00	25.00	37.50	25.00	25.83	26.55	25.00	30.70
GPT-3.5 Turbo	n/a	–	57.33	48.62	31.24	36.65	46.71	42.42	–
GPT-4 Turbo	n/a	–	**90.67**	**66.54**	**56.70**	**75.85**	**73.12**	**70.22**	–
XGLM	7B	57.71	23.88	41.47	22.96	24.43	26.53	24.59	31.65
BLOOM	7B	57.18	27.00	40.17	25.34	28.41	27.17	25.07	32.91
Mistral	7B	51.09	**38.89**	39.22	25.01	**29.26**	34.58	32.15	35.74
Llama 2	7B	50.43	26.22	38.20	24.09	27.27	29.50	28.84	32.08
[new] Latxa	7B	**65.45**	37.33	**52.56**	**30.26**	25.00	**42.16**	**33.82**	**40.94**
mGPT	13B	55.39	25.00	37.56	25.00	24.15	27.17	25.73	31.43
Llama 2	13B	50.63	32.00	38.98	25.90	28.98	33.53	29.66	34.24
[new] Latxa	13B	**66.51**	**53.89**	**53.36**	**44.11**	**32.67**	**56.38**	**43.66**	**50.08**
Mixtral	8x7B	52.55	50.44	45.00	26.43	37.50	42.51	39.87	42.04
Yi	34B	52.22	54.56	43.90	27.30	34.66	42.57	39.68	42.13
Llama 2	70B	51.62	33.56	42.55	24.16	27.84	38.43	33.08	35.89
[new] Latxa	70B	**70.55**	**71.67**	**59.74**	**60.65**	**50.57**	**62.45**	**51.90**	**61.08**

Table 1: **Main results.** Best results in each compute class are in **bold**. Best overall results are underlined.

stronger English models become available.

This paper makes the following contributions: **(1)** We release a high-quality corpus for Basque, comprising 4.3M documents and 4.2B tokens. The corpus combines the EusCrawl v1.1, Egunkaria, Booktegi, Wikipedia, CulturaX, Colossal OSCAR and HPLT v1 datasets (the first 3 being new), which we carefully deduplicate and filter. **(2)** We release the Latxa family of Basque LLMs, comprising 3 models with 7B, 13B and 70B parameters. **(3)** We release 4 new multiple-choice benchmarks for Basque: EusProficiency (official language proficiency exams), EusReading (reading comprehension), EusTrivia (trivia questions from 5 knowledge areas), and EusExams (public examinations). **(4)** We present extensive experiments comparing Latxa to previous open and closed models. **(5)** We show that it is possible to train significantly stronger LLMs for low-resource languages building on the existing ecosystem of open models and resources. In a similar spirit to other open LLMs, such as Pythia (Biderman et al., 2023), LLM360 (Liu et al., 2023) and OLMO (Groeneveld et al., 2024), we release all the necessary data, code, weights and documentation to run and evaluate our models, facilitating similar efforts for other low-resource languages.

2 Training Corpora

Our training corpus combines various existing datasets, as well as some new ones that we release with this work. We have prioritized quality over quantity when constructing our corpus, prioritizing high-quality data sources and applying a thorough deduplication and filtering process. We next describe our data sources in §2.1, followed by our preprocessing pipeline in §2.2. Table 2 summarizes the statistics of the resulting dataset.

2.1 Data Sources

[new] EusCrawl v1.1. The original version of EusCrawl (Artetxe et al., 2022) was built using ad-hoc scrapers to extract text from 33 newswire websites, resulting in higher quality documents compared to general-purpose approaches. In this work, we release an updated version of EusCrawl, including new content up to November 2023. This increases the number of unique documents from 1.38 to 1.94 millions, for a total of 384 million words.

[new] Egunkaria. Euskaldunon Egunkaria was a daily newspaper written fully in the Basque language. The corpus includes approximately 176k news articles, editorials, and various types of reviews from the years 2001 to 2006, totalling 39 million words.

[new] Booktegi. The Booktegi platform hosts free content in Basque, such as books, interviews, and audio materials. The corpus comprises approximately 3 million words from 166 EPUB books covering essays, fiction, and poetry.

Wikipedia. We download and process a Basque Wikipedia dump,[2] obtaining nearly 550k documents and more than 54 million words. The plain text has been extracted with WikiExtractor.[3]

[2]The `20231101` dump corresponding to November 2023.
[3]`https://github.com/attardi/wikiextractor`

	Raw		Deduped		Filtered			
	Docs	Words	Docs	Words	Docs	Words	Toks	Source
CulturaX	1.60M	622M	1.33M	548M	1.31M	541M	1.84B	hf.co/uonlp/CulturaX
[new] EusCrawl v1.1	2.12M	411M	1.94M	384M	1.79M	359M	1.21B	Artetxe et al. (2022)
HPLT v1	2.29M	1.55B	1.56M	312M	0.37M	120M	421M	hplt-project.org
Colossal OSCAR	0.65M	283M	0.25M	111M	0.24M	105M	380M	hf.co/oscar-corpus
Wikipedia	0.55M	54M	0.55M	54M	0.41M	51M	182M	dumps.wikimedia.org
[new] Egunkaria	0.18M	40M	0.18M	39M	0.18M	39M	129M	n/a
[new] Booktegi	181	3M	177	3M	166	3M	8M	booktegi.eus
Total	7.39M	2.96B	5.80M	1.45B	4.30M	1.22B	4.17B	

Table 2: Data sources and statistics at each preprocessing stage. *"Toks"* are Llama 2 tokens.

CulturaX. CulturaX (Nguyen et al., 2023) is a large multilingual dataset resulting from the combination and processing of mC4 (Xue et al., 2021) and the four OSCAR releases `2019`, `21.09`, `22.01`, and `23.01` (Ortiz Suárez et al., 2019). These corpora originate, in turn, from 66 Common Crawl (CC) snapshots spanning from 2015 to 2022. Basque content constitutes 0.02% of CulturaX, encompassing nearly 1.60 million documents and 622 million words.

Colossal OSCAR. The largest release of the OSCAR project (Ortiz Suárez et al., 2019) to date, Colossal OSCAR 1.0, is based on 10 CC snapshots. Here, we use the two snapshots not covered by CulturaX, namely, `06-07-22` and `05-06-23`. Additionally, we have had access to an OSCAR-processed CC snapshot from April 2023. In total, we have obtained almost 650k documents in Basque from these datasets, totalling 282 million words.

HPLT v1. The High Performance Language Technologies project (HPLT; Aulamo et al., 2023) compiled another massive, multilingual dataset from the Internet Archive and CC. In this work, we use the first release, which contains 2.29 million documents (1.55 billion words) in Basque. It must be noted that, unlike the aforementioned web-based sources, the HPLT dataset was released without any deduplication or filtering. Consequently, our preprocessing approach has been particularly aggressive with this dataset (see §2.2).

2.2 Preprocessing

We used the Dolma toolkit (Soldaini et al., 2024) and Corpus Cleaner v2 (CCv2; Palomar-Giner et al., 2024) to normalize, deduplicate and filter the datasets. Since the majority of our data sources were intentionally selected, organized, and/or curated by their respective authors, our main focus has been on removing outliers and cross-dataset duplicates. This process is briefly outlined below, with further details available in Appendix A. The final size of the processed corpus is shown in Table 2. In total, it amounts to 1.22B words and 4.17B Llama 2 tokens. Each dataset is shuffled and then split separately into testing (1%), development (1%) and training (98%) example sets.

Normalization. CCv2 is first used to fix document encoding and whitespace normalization.

Deduplication. Cross-dataset document repetitions are identified and removed at both the URL and document content levels. Specifically, we conduct near-deduplication with Bloom filters (Bloom, 1970) as implemented in Dolma. To maximise corpus quality, we prioritized content from well-curated sources (Wikipedia, EusCrawl, Egunkaria and Booktegi) then from massive but comparatively cleaner sources (CulturaX and Colossal OSCAR) over HPLT (see Figure 2(a) in Appendix A). The latter undergoes additional deduplication at the paragraph level.

Filtering. Documents unlikely to contain quality content are identified and removed in two stages. First, a combination of heuristics from Gopher (Rae et al., 2022) and C4 (Raffel et al., 2020), with adaptations tailored to the Basque language is applied (e.g., regarding average word length). We also perform language identification with CLD2 through Dolma, which has predominantly impacted HPLT, with approximately one-third of this corpus being discarded at this stage. Finally, the corpora are processed with CCv2, which assigns an aggregated quality score per document based on a comprehensive set of taggers. Once again, HPLT has been affected most, resulting in a further 25% reduction in document counts for this dataset.

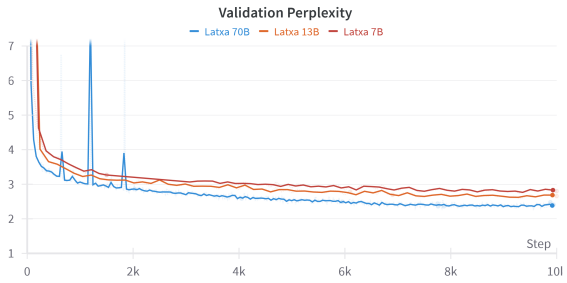


Figure 1: Validation perplexity throughout training.

Size	Time (GPU Hours)	Carbon Emitted (kg CO_2 eq)
7B	952.53h	124.47kg
13B	2,518.0h	329.06kg
70B	30,266.0h	3,955.17kg
Total	33,636.5h	4,408.7kg

Table 3: Carbon footprint of training different models

3 Latxa Models

We train 7B, 13B and 70B models following a continued pretraining approach. To that end, we use Llama 2 as the base model (Touvron et al., 2023), and continue training it using the corpus described in §2. To mitigate catastrophic forgetting from the original model[4], we also include English data in the continued pretraining stage. For that purpose, we use 500k random documents from The Pile (Gao et al., 2020), totaling 0.9B tokens.

3.1 Pretraining Details

The training of Latxa has been conducted using the GPT-Neox (Andonian et al., 2023) library. As infrastructure, we have leveraged the CINECA HPC Leonardo computing cluster located in Italy, which is powered by 3,456 nodes each containing 4x custom A100 64GB GPUs. The models are trained for 10k steps with a sequence length of 4,096 tokens and an effective batch size of 1M tokens, resulting in a total of 10B tokens. We use a cosine learning rate schedule, with a warm-up of 500 steps and decaying down to 3% of the peak learning rate. We set up the peak learning rate to be 1×10^{-4}. All other hyperparameters follow Touvron et al. (2023). Figure 1 shows the validation perplexity during training.

3.2 Carbon Emissions

Pretraining LLMs requires compute-expensive experiments, carrying a significant carbon footprint. The 7B, 13B and 70B models were trained on 32, 64 and 256 GPUs, respectively. We report the compute hours and power consumption involved in our experiments in Table 3. The carbon emitted was estimated using a GPU power consumption of $440W$ and a carbon efficiency of $0.297kg/kWh$ (carbon efficiency on Italy on February 9, 2024, according to ElectricityMaps[5]).

4 New Evaluation Datasets

To overcome the scarcity of Basque benchmarks that are suitable for evaluating base language models, we collect new evaluation data from various online games and tests. We have decided to take this approach instead of translating existing datasets to avoid translation artifacts (Artetxe et al., 2020). Most importantly, this allows to have localized datasets that test the models' knowledge about topics that are most relevant for Basque speakers. These tasks cover language proficiency (EusProficiency), reading comprehension (EusReading), trivia questions (EusTrivia), and exams of advanced professional level (EusExams). All the datasets consist of multiple-choice questions, making them suitable for few-shot learning akin to MMLU (Hendrycks et al., 2021) in English. We next describe each dataset in more detail, while Table 4 summarizes their statistics. For examples of each task, see Table 9 in Appendix C.

EusProficiency. EusProficiency comprises 5,169 exercises on different topics from past EGA exams, the official C1-level certificate of proficiency in Basque. We have collected the *atarikoa* exercises from EGA exams through the years 1998 to 2008. Atarikoa is the first qualifying test of EGA, which measures different aspects of language competency, such as reading comprehension, grammar, vocabulary, spelling, and writing. Each test generally has 85 multiple-choice questions, with 4 choices and a single correct answer. Currently, there is no comparable dataset available, nor could one be obtained by translating existing analogous datasets from other languages.

[4]Our preliminary experiments showed that adding English data was critical for strong few-shot performance. For instance, an early version of our 7B model without English data obtained 23.67 points on BeleBele (Bandarkar et al., 2023), 13 points less than the corresponding version with English data.

[5]https://www.electricitymaps.com/

EusReading. EusReading consists of 352 reading comprehension exercises (*irakurmena*) sourced from the same set of past EGA exams. Each test generally has 10 multiple-choice questions, with 4 choices and a single correct answer. These exercises are more challenging than Belebele (Bandarkar et al., 2023) due to the complexity and length of the input texts (see Table 4). As a result, EusReading is useful to measure long context understanding of models.

EusTrivia. EusTrivia consists of 1,715 trivia questions from multiple online sources. 56.3% of the questions are elementary level (grades 3-6), while the rest are considered challenging. A significant portion of the questions focuses specifically on the Basque Country, its language, and culture. Each multiple-choice question contains two, three or four choices (3.84 on average) and a single correct answer. Five areas of knowledge are covered:

- **Humanities and Natural Sciences** (27.8%): This category encompasses questions about history, geography, biology, ecology and other social and natural sciences.
- **Leisure and Art** (24.5%): This category includes questions on sports and athletes, performative and plastic arts and artists, architecture, cultural events, and related topics.
- **Music** (16.0%): Here are grouped all the questions about music and musicians, both classical and contemporary.
- **Language and Literature** (17.1%): This category is concerned with all kinds of literature productions and writers, as well as metalinguistic questions (e.g., definitions, synonyms, and word usage).
- **Mathematics and ICT** (14.5%): This category covers mathematical problems and questions about ICT, as well as questions about people known for their contributions to these fields of knowledge.

EusExams. EusExams is a collection of tests designed to prepare individuals for Public Service examinations conducted by several Basque institutions, including the public health system Osakidetza, the Basque Government, the City Councils of Bilbao and Gasteiz, and the University of the Basque Country (UPV/EHU). Within each of these groups, there are different exams for public positions, such as administrative and assistant roles. Each multiple-choice question contains 2 to 4 choices (3.90 on average) and one correct answer. The dataset is mostly parallel with 16k questions in Basque and 18k in Spanish, from which we only consider the Basque subset. It could be said to be similar to MMLU's professional-level questions, but focusing on knowledge relevant to the Basque community, with questions related to local public services and law.

		Input	Output	
	Items	chars	choices	chars
[new] **EusProf**	5,169	50	4	28
[new] **EusReading**	352	5,340	2-4	67
[new] **EusTrivia**	1,715	55	2-4	14
[new] **EusExams**	16,774	115	4	63
XStoryCloze	1,511	202	2	44
Belebele	900	584	4	28
BEC	1,302	97	3	
BHTCv1	1,854	265	12	
Korref	587	275	2	
QNLI$_{eu}$	238	158	2	
VaxxStance	312	209	3	
WiC$_{eu}$	1,400	375	2	

Table 4: Evaluation dataset statistics: number of examples, average input and output length in characters, and number of choices per example. BasqueGLUE tasks (lower section) do not have output length because they are classification tasks.

5 Experimental Setting

To assess the quality of Latxa models, we thoroughly evaluate them on a suite of diverse and challenging tasks against state-of-the-art models. In this section, we present the baseline models (§5.1) and tasks (§5.2), and describe the evaluation framework (§5.3). Results are analyzed and discussed in the next section (§6).

5.1 Baseline Models

We test the Llama 2 models (Touvron et al., 2023) to show the difference after our continued pretraining. We also evaluated other multilingual base models that do not intentionally include Basque in pretraining, namely, Mistral-7B (Jiang et al., 2023), Mixtral-8x7B (Jiang et al., 2024), and 01.AI's Yi-34B (AI et al., 2024).

Furthermore, we evaluate some of the leading open multilingual language models for Basque available to date, including BLOOM-7B (Scao et al., 2023), XGLM-7B (Lin et al., 2022), and mGPT-13B (Shliazhko et al., 2024). These models cover a broader range of languages than more re-

cent models, but are trained on fewer tokens and exhibit generally weaker performance.

Finally, we tested the latest GPT-3.5 Turbo (`gpt-3.5-turbo-0125`) and GPT-4 Turbo (`gpt-4-1106-preview` for BasqueGLUE tasks and `gpt-4-0125-preview` for the rest), as they are the leading commercial models for Basque.

5.2 Evaluation Datasets

In addition to the new evaluation datasets introduced in §4, the models have been evaluated on the following benchmarks:

- **Belebele** (Bandarkar et al., 2023): A multiple-choice reading comprehension dataset spanning 122 language variants.
- **XStoryCloze** (Lin et al., 2022): A professionally translated version of the StoryCloze dataset (Mostafazadeh et al., 2017) to 10 non-English languages. StoryCloze is a commonsense reasoning dataset that consists in choosing the correct ending to a four-sentence story.
- **BasqueGLUE** (Urbizu et al., 2022): A collection of 6 NLU datasets for Basque: sentiment analysis (BEC), stance detection (VaxxStance), topic classification (BTHCv2), coreference detection (EpecKorrefBin), question-answering NLI (QNLI_{eu}), and word-in-context (WiC_{eu}).

Collectively, these datasets allow us to evaluate the performance of the models on a wide range of competences including world knowledge, linguistic knowledge, reading comprehension, and common sense reasoning.

Following previous work (Brown et al., 2020; Touvron et al., 2023), we check for n-gram overlaps between these evaluation datasets and Latxa's training corpus, and find no evidence of wholesale or annotation contamination (Chowdhery et al., 2023; Sainz et al., 2023). Further information on our contamination study can be consulted in Appendix B.

5.3 Evaluation Framework

The models are evaluated using the LM Evaluation Harness library (Biderman et al., 2024) by Eleuther AI. To that end, we have implemented the new evaluation datasets following similar multiple-choice datasets that are already included in the library, such as Belebele. The specific prompts and examples for each task are reported in Table 9 in Appendix C. BasqueGLUE has also been implemented as a generative evaluation dataset (see Table 10).[6]

		Hum & Nat	Leis & Art	Music	Lang & Lit	Math & ICT
GPT-3.5 Turbo	n/a	50.10	48.93	43.64	43.00	44.18
GPT-4 Turbo	n/a	**75.47**	**75.30**	**61.82**	**66.89**	**84.74**
XGLM	7B	23.06	25.65	28.73	28.67	29.72
BLOOM	7B	22.01	25.42	26.91	32.42	**34.14**
Llama 2	7B	29.77	26.84	32.73	32.76	26.10
[new] Latxa	7B	**43.81**	**41.33**	**45.09**	**45.05**	33.73
mGPT	13B	22.22	26.13	26.91	32.76	32.13
Llama 2	13B	30.82	31.59	37.09	37.88	32.93
[new] Latxa	13B	**59.53**	**59.85**	**53.81**	**62.11**	**40.56**
Mixtral	8x7B	46.54	43.94	40.73	38.57	38.96
Yi	34B	40.88	42.99	40.73	41.30	**48.59**
Llama 2	70B	41.30	38.24	38.55	35.15	36.95
[new] Latxa	70B	**67.50**	**63.50**	**57.81**	**70.30**	46.58

Table 5: EusTrivia results by category (accuracy).

We use 5 in-context examples for all tasks but two: following common practice, XStoryCloze is evaluated in a 0-shot setting, and EusReading is evaluated in a 1-shot fashion, as more examples would not fit into the context of most models. In all cases, we compute the log probabilities of all candidates, and pick the one with the highest score as the models' final answer.

For GPT models, we have implemented the evaluation using the OpenAI API. We have kept the evaluation as similar as possible to allow a fair comparison with our models. As getting log probabilities of candidate XStoryCloze continuations from the API is not possible, we have decided not to evaluate GPT in that task. For few-shot tasks, we use the same prompts and provide few-shot examples as user and assistant messages. In addition, we use a system prompt in English to specify the set of candidate answers per task (see Appendix C).

6 Results

We report our main results in Table 1, while Table 5 and Table 6 report fine-grained results on the different subsets of EusTrivia and BasqueGLUE, respectively. In what follows, we summarize our main findings:

Effectiveness of continued pretraining. Latxa obtains the best results in each compute class, outperforming all previous open models by a large margin. As the only exception, Mistral 7B is better

[6] Where possible, we based our prompts on existing analogous datasets: QNLI for QNLI_{eu}, WiC for WiC_{eu}, and WSC for EpecKorrefBin.

		BEC *F1*	Vaxx *F1**	BHTC *F1*	Korref *acc*	QNLI *acc*	WiC *acc*	Avg
Random		33.33	33.33	8.33	50.00	50.00	50.00	37.50
BERTeus†	110M	69.43	59.30	**78.26**	**68.31**	**74.26**	70.71	70.04
ElhBERTeu†	110M	**69.89**	**63.81**	78.05	65.93	73.84	**71.71**	**70.54**
GPT-3.5 Turbo	n/a	59.52	38.17	42.66	50.09	50.00	51.29	48.62
GPT-4 Turbo	n/a	**67.90**	**57.10**	**52.21**	**88.25**	**71.85**	**61.93**	**66.54**
XGLM	7B	39.94	21.58	36.73	50.94	50.42	49.21	41.47
BLOOM	7B	37.94	20.72	39.10	48.21	47.48	47.57	40.17
Mistral	7B	40.63	21.40	24.81	48.04	50.84	49.57	39.22
Llama 2	7B	41.63	18.60	20.06	50.94	48.32	49.64	38.20
[new] Latxa	7B	**57.30**	**45.65**	**57.44**	**49.50**	**54.20**	**51.28**	**52.56**
mGPT	13B	35.41	17.54	23.73	47.53	50.84	50.29	37.56
Llama 2	13B	41.09	18.25	27.35	49.23	48.74	49.21	38.98
[new] Latxa	13B	**53.92**	**47.66**	**57.50**	**54.17**	**55.88**	**51.00**	**53.36**
Mixtral	8x7B	49.46	21.81	37.32	53.32	57.56	50.50	45.00
Yi	34B	47.08	29.33	30.69	54.68	49.58	52.00	43.90
Llama 2	70B	47.47	21.01	31.01	52.98	51.26	51.57	42.55
[new] Latxa	70B	**61.06**	**55.71**	**55.88**	**72.57**	**59.66**	**53.57**	**59.74**

Table 6: BasqueGLUE results by task. *VaxxStance is measured in terms of macro-average F1-score of the categories IN FAVOUR and AGAINST. †BERTeus and ElhBERTeu are fine-tuned encoders.

than Latxa 7B on Belebele and EusReading, but Latxa 7B wins in all the other 5 datasets. Our best model obtains 61.08 points on average, outperforming the previous best open model by 18.95 points. In addition, it outperforms the Llama 2 model it is based on by 25.18 points, confirming the effectiveness of continued pretraining to build language models for low-resource languages.

Open vs. closed models. With the exception of EusProficiency, the best results are obtained by GPT-4 Turbo, a closed commercial system. The difference between GPT-4 Turbo and the previous best open model (Yi) is abysmal. For instance, GPT-4 Turbo is 30.55 points better than Yi on EusTrivia, while the latter is only 16.02 points better than random chance. This can partly be attributed to previous open initiatives being primarily English-centric. While Latxa substantially narrows this gap, we believe that future research on open models should pay more attention to low-resource languages.

English-centric vs. multilingual models. The 3 multilingual models we evaluate (XGLM, BLOOM and mGPT) do better than all English-centric models on XStoryCloze, which requires linguistic competence in Basque and is evaluated in a zero-shot fashion. However, English-centric open models do generally better on other tasks, presumably due to their better in-context learning capabilities and general knowledge captured. Consistent with our previous point, this suggests that existing open models are either English-centric and struggle in low-resource languages like Basque, or are multilingual and significantly lag behind the English-centric models in language-agnostic capabilities.

Impact of scale. We find that larger Latxa models obtain substantially better results: the 70B model is 10.99 points better than the 13B model on average, which is 9.14 points better than the 7B model. This transcends conventional scaling laws, which establish that, when pretraining models from scratch in low-resource scenarios, the performance is bottlenecked by the amount of training data rather than the model size (Kaplan et al., 2020). However, our results show a different picture for continued pretraining, where bigger and stronger base models result in better performance despite the limited pretraining data in the target language. This suggests that even better results could be obtained through continued pretraining as stronger English-centric models become available, which is encouraging for low-resource languages.

General vs. language-specific knowledge. We find evidence that Latxa is particularly strong in tasks that test for proficiency in the Basque language. In particular, Latxa obtains the best results on Basque language proficiency exams (EusProficiency), despite lagging behind GPT-4 Turbo in the rest of the tasks. Similarly, Latxa outperforms

GPT-4 Turbo on the *Language & Literature* subset of EusTrivia, even if GPT-4 Turbo is superior in the rest of the categories (Table 5). This suggests that Latxa is more proficient than GPT-4 Turbo in Basque, but the latter does better in most tasks due to its stronger general capabilities. Another evidence of this is that Latxa is particularly weak in the *Maths & ICT* subset of EusTrivia, where it even lags behind Yi, while GPT-4 Turbo is particularly strong in this category. This is likely because most problems in this category can be understood with basic knowledge of Basque, but solving them may require more complex mathematical reasoning. This suggests that the general capabilities of language models are highly language-agnostic which, in line with our previous finding, suggests that stronger English-centric models can lead to stronger models for low-resource languages by using the same continued pretraining recipe.

Classical NLP tasks. Table 6 reports fine-grained results on BasqueGLUE, which comprises various classical NLP tasks like topic classification and coreference detection. In addition to our usual set of decoder-only models evaluated in a few-shot fashion, we report results for BERTeus (Agerri et al., 2020) and ElhBERTeu (Urbizu et al., 2023), which are encoder-only models that were fine-tuned specifically on these tasks. The best results are obtained by these specialized encoder-only models, which shows that the traditional pretraining/fine-tuning paradigm with BERT-style models is still competitive for classical NLP tasks. The only exception is *EpecKorrefBin*, where both GPT-4 Turbo and Latxa 70B perform substantially better than the fine-tuned encoder-only models. In future work, we would like to explore fine-tuning Latxa and other decoder-only models on these tasks.

7 Related Work

The amount of documents per language in Common Crawl[7] is helpful to organize the literature. There is no agreed-upon definition of low-resource language, so we set five arbitrary buckets of languages following a logarithmic distribution: *high* bucket with just English (rank 1, 46% of latest crawl), *high-medium* languages around Spanish (rank 5, 4.6%), *medium* around Danish (rank 24, 0.46%), *low* resource around Nepalese (rank 46, 0.044), and *very-low* around Somali (rank 81, 0.0046%). Basque would be low-resource (rank 52, 0.035%).

While there were some early efforts to build **open multilingual language models** like XGLM (Lin et al., 2022), BLOOM (Scao et al., 2023), mGPT (Shliazhko et al., 2024) and the translation oriented MADLAD-400 (Kudugunta et al., 2023), their performance significantly lags behind more recent English-centric models like Llama 2 (Touvron et al., 2023) or Mistral (Jiang et al., 2023). Concurrent to our work, Üstün et al. (2024) present Aya, a fine-tuned mT5 encoder-decoder (Xue et al., 2021), which has been instruction-tuned and supports 101 languages.

In the case of **high-medium-resource** languages, different teams have focused on building models from scratch. Ekgren et al. (2022) built a Swedish 20B model which was favorably evaluated on perplexity, followed by (Ekgren et al., 2024), which builds a 40B model for five Nordic languages, including low-resource Danish and Faroese, English and code. It was evaluated on perplexity for the three richest languages. Faysse et al. (2024) present a French-English bilingual model trained with the same number of tokens for each language, although the amount of text for French (376M documents) is half the English text. The authors stress the fact that the tokenizer should not be biased towards any of the two languages. The results on French (and English) LLM evaluation benchmarks show that their largest model (1.2B parameters) underperforms both 3B Llama 2 and Mistral models.

Regarding **medium-resource**, Luukkonen et al. (2023) focus on Finnish. Using a 19B word corpus they trained several models from scratch, ranging from 186M to 13B parameters. As an alternative, they also **continued pretraining** a multilingual model, the 176B BLOOM. The evaluation on FIN-bench, a version of Big-Bench (Srivastava et al., 2023), shows the 13B model underperforms the 7B model, while the continued pretraining model obtains the best results by a large margin. In follow-up work, Luukkonen et al. (2024) train a 40B bilingual model from scratch with a larger Finnish corpus, and obtain the best results in FIN-bench. They do not compare the results to commercial models. The results suggest that monolingual models trained from scratch saturate at relatively small sizes, and multilingual or continued pretraining would be the best option for these languages.

Focusing on **very-low-resource** languages,

[7] CC-MAIN-2024-22 crawl at https://commoncrawl.github.io/cc-crawl-statistics/plots/languages

Yong et al. (2023) compare different approaches to extend the 176B BLOOM model from 46 languages to new languages unseen at training, including Guarani (rank 116, 0.0006%) and seven larger languages for which they sample 100K documents at most. Model sizes range from 0.56B to 7.1B parameters. They report good results for adapters on zero-shot benchmarks. In concurrent work, Lin et al. (2024) present MaLA-500, a continued pretrained model using LORA based on Llama 2 that supports 500 languages. MaLA seems to improve Llama 2 on low to very-low resource languages, but degrades in some medium languages already covered in Llama 2. Evaluation is based on a variant of perplexity and text classification.

Previous work on **low-resource** languages has been done in the context of multilingual models (see above). In all cases, evaluation is based on perplexity and/or some readily available datasets, and does not include native benchmarks designed to evaluate base models.

8 Conclusion and Future Work

In this work, we present a new open framework for the development and evaluation of LLMs in Basque. The framework includes Latxa, a set of state-of-the-art generative LLMs that have been built by continuing to pretrain the Llama 2 7B, 13B and 70B models in Basque. The pretraining dataset is the largest public dataset available to date and includes data from carefully curated sources, as well as content derived from automatically filtered versions of Common Crawl. After preprocessing and deduplication, the released corpora comprise 1.22B words and 4.17B tokens. We also present 4 new evaluation datasets, collectively the largest evaluation benchmark for Basque that allows assessing the knowledge of the models about the Basque language and culture.

The Latxa models outperform all previous open models and GPT-3.5 Turbo, but they still lag behind GPT-4 Turbo in most benchmarks. Interestingly, Latxa 70B outperforms GPT-4 Turbo on EusProficiency and *Language & Literature* EusTrivia questions, suggesting that the capabilities of LLMs in a particular language are not determined by their linguistic competence in this language. This, along with the effectiveness of scale, suggests that applying the exact same continued pretraining recipe could lead to better Basque models as stronger English-only models become available.

In the future, we plan to extend the training dataset by gathering quality content from diverse Basque sources such as publishers or media, as well as building evaluation datasets to assess aspects such as truthfulness or hallucinations. We also plan to further tune Latxa to follow instructions, which should improve the overall capabilities of our models.

Limitations

To alleviate the potentially disturbing or harmful content, Latxa has been trained on carefully selected and processed data which comes mainly from local media, national/regional newspapers, encyclopedias and blogs. Still, the model is based on Llama 2 models and can potentially carry the same biases, risks and limitations.

Latxa models are pretrained LLMs without any task-specific or instruction fine-tuning. The model was not fine-tuned to follow instructions or to work as a chat assistant, therefore, this kind of usage is not tested nor recommended. That is, the model can either be prompted to perform a specific task or further fine-tuned for specific use cases.

Acknowledgements

This work has been partially supported by the Basque Government (Research group funding IT-1805-22 and IKER-GAITU project), the Spanish Ministry for Digital Transformation and of Civil Service, and the EU-funded NextGenerationEU Recovery, Transformation and Resilience Plan (ILENIA project, 2022/TL22/00215335). The models were trained on the Leonardo supercomputer at CINECA under the EuroHPC Joint Undertaking, project EHPC-EXT-2023E01-013. Julen Etxaniz and Oscar Sainz hold a PhD grant from the Basque Government (PRE_2023_2_0060 and PRE_2023_2_0137, respectively).

References

Rodrigo Agerri, Iñaki San Vicente, Jon Ander Campos, Ander Barrena, Xabier Saralegi, Aitor Soroa, and Eneko Agirre. 2020. Give your text representation models some love: the case for Basque. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4781–4788, Marseille, France. European Language Resources Association.

01. AI, :, Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Heng Li, Jiangcheng Zhu, Jianqun Chen, Jing Chang, Kaidong

Yu, Peng Liu, Qiang Liu, Shawn Yue, Senbin Yang, Shiming Yang, Tao Yu, Wen Xie, Wenhao Huang, Xiaohui Hu, Xiaoyi Ren, Xinyao Niu, Pengcheng Nie, Yuchi Xu, Yudong Liu, Yue Wang, Yuxuan Cai, Zhenyu Gu, Zhiyuan Liu, and Zonghong Dai. 2024. Yi: Open foundation models by 01.AI. *arXiv preprint arXiv:2403.04652*.

Alex Andonian, Quentin Anthony, Stella Biderman, Sid Black, Preetham Gali, Leo Gao, Eric Hallahan, Josh Levy-Kramer, Connor Leahy, Lucas Nestler, Kip Parker, Michael Pieler, Jason Phang, Shivanshu Purohit, Hailey Schoelkopf, Dashiell Stander, Tri Songz, Curt Tigges, Benjamin Thérien, Phil Wang, and Samuel Weinbach. 2023. GPT-NeoX: Large scale autoregressive language modeling in PyTorch.

Mikel Artetxe, Itziar Aldabe, Rodrigo Agerri, Olatz Perez-de Viñaspre, and Aitor Soroa. 2022. Does corpus quality really matter for low-resource languages? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7383–7390, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2020. Translation artifacts in cross-lingual transfer learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7674–7684, Online. Association for Computational Linguistics.

Mikko Aulamo, Nikolay Bogoychev, Shaoxiong Ji, Graeme Nail, Gema Ramírez-Sánchez, Jörg Tiedemann, Jelmer van der Linde, and Jaume Zaragoza. 2023. HPLT: High performance language technologies. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 517–518, Tampere, Finland. European Association for Machine Translation.

Lucas Bandarkar, Davis Liang, Benjamin Muller, Mikel Artetxe, Satya Narayan Shukla, Donald Husa, Naman Goyal, Abhinandan Krishnan, Luke Zettlemoyer, and Madian Khabsa. 2023. The Belebele benchmark: a parallel reading comprehension dataset in 122 language variants. *arXiv preprint arXiv:2308.16884*.

Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, et al. 2023. Pythia: A suite for analyzing large language models across training and scaling. In *International Conference on Machine Learning*, pages 2397–2430. PMLR.

Stella Biderman, Hailey Schoelkopf, Lintang Sutawika, Leo Gao, Jonathan Tow, Baber Abbasi, Alham Fikri Aji, Pawan Sasanka Ammanamanchi, Sidney Black, Jordan Clive, Anthony DiPofi, Julen Etxaniz, Benjamin Fattori, Jessica Zosa Forde, Charles Foster, Jeffrey Hsu, Mimansa Jaiswal, Wilson Y. Lee, Haonan Li, Charles Lovering, Niklas Muennighoff, Ellie Pavlick, Jason Phang, Aviya Skowron, Samson Tan, Xiangru Tang, Kevin A. Wang, Genta Indra Winata, François Yvon, and Andy Zou. 2024. Lessons from the trenches on reproducible evaluation of language models. *arXiv preprint arXiv:2405.14782*.

Burton H. Bloom. 1970. Space/time trade-offs in hash coding with allowable errors. *Commun. ACM*, 13(7):422–426.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2023. PaLM: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113.

Ariel Ekgren, Amaru Cuba Gyllensten, Evangelia Gogoulou, Alice Heiman, Severine Verlinden, Joey Öhman, Fredrik Carlsson, and Magnus Sahlgren. 2022. Lessons learned from GPT-SW3: Building the first large-scale generative language model for Swedish. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3509–3518, Marseille, France. European Language Resources Association.

Ariel Ekgren, Amaru Cuba Gyllensten, Felix Stollenwerk, Joey Öhman, Tim Isbister, Evangelia Gogoulou, Fredrik Carlsson, Judit Casademont, and Magnus Sahlgren. 2024. GPT-SW3: An autoregressive language model for the Scandinavian languages.

In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 7886–7900, Torino, Italia. ELRA and ICCL.

Manuel Faysse, Patrick Fernandes, Nuno M. Guerreiro, António Loison, Duarte M. Alves, Caio Corro, Nicolas Boizard, João Alves, Ricardo Rei, Pedro H. Martins, Antoni Bigata Casademunt, François Yvon, André F. T. Martins, Gautier Viaud, Céline Hudelot, and Pierre Colombo. 2024. CroissantLLM: A truly bilingual French-English language model. *arXiv preprint arXiv:2402.00786*.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, et al. 2020. The Pile: An 800GB dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*.

Dirk Groeneveld, Iz Beltagy, Pete Walsh, Akshita Bhagia, Rodney Kinney, Oyvind Tafjord, Ananya Harsh Jha, Hamish Ivison, Ian Magnusson, Yizhong Wang, et al. 2024. OLMo: Accelerating the science of language models. *arXiv preprint arXiv:2402.00838*.

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

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7B. *arXiv preprint arXiv:2310.06825*.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2024. Mixtral of experts. *arXiv preprint arXiv:2401.04088*.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.

Sneha Kudugunta, Isaac Caswell, Biao Zhang, Xavier Garcia, Christopher A Choquette-Choo, Katherine Lee, Derrick Xin, Aditya Kusupati, Romi Stella, Ankur Bapna, et al. 2023. MADLAD-400: A multilingual and document-level large audited dataset. *arXiv preprint arXiv:2309.04662*.

Peiqin Lin, Shaoxiong Ji, Jörg Tiedemann, André FT Martins, and Hinrich Schütze. 2024. MaLA-500: Massive language adaptation of large language models. *arXiv preprint arXiv:2401.13303*.

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 Diab, Veselin Stoyanov, and Xian Li. 2022. Few-shot learning with multilingual generative language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9019–9052, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Zhengzhong Liu, Aurick Qiao, Willie Neiswanger, Hongyi Wang, Bowen Tan, Tianhua Tao, Junbo Li, Yuqi Wang, Suqi Sun, Omkar Pangarkar, et al. 2023. LLM360: Towards fully transparent open-source LLMs. *arXiv preprint arXiv:2312.06550*.

Risto Luukkonen, Jonathan Burdge, Elaine Zosa, Aarne Talman, Ville Komulainen, Väinö Hatanpää, Peter Sarlin, and Sampo Pyysalo. 2024. Poro 34b and the blessing of multilinguality. *arXiv preprint arXiv:2404.01856*.

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, Samuel Antao, and Sampo Pyysalo. 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.

Nasrin Mostafazadeh, Michael Roth, Annie Louis, Nathanael Chambers, and James Allen. 2017. LSDSem 2017 shared task: The story cloze test. In *Proceedings of the 2nd Workshop on Linking Models of Lexical, Sentential and Discourse-level Semantics*, pages 46–51.

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. *arXiv preprint arXiv:2309.09400*.

OpenAI, :, Josh Achiam, Steven Adler, and Sandhini Agarwal et al. 2023. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*.

Pedro Javier Ortiz Suárez, Benoît Sagot, and Laurent Romary. 2019. Asynchronous pipelines for processing huge corpora on medium to low resource infrastructures. In *7th Workshop on the Challenges in*

the Management of Large Corpora (CMLC-7), Proceedings of the Workshop on Challenges in the Management of Large Corpora (CMLC-7) 2019. Cardiff, 22nd July 2019, pages 9 – 16, Mannheim. Leibniz-Institut für Deutsche Sprache.

Jorge Palomar-Giner, Jose Javier Saiz, Ferran Espuña, Mario Mina, Severino Da Dalt, Joan Llop, Malte Ostendorff, Pedro Ortiz Suarez, Georg Rehm, Aitor Gonzalez-Agirre, and Marta Villegas. 2024. A CURATEd CATalog: Rethinking the extraction of pretraining corpora for mid-resourced languages. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 335–349, Torino, Italia. ELRA and ICCL.

Jack W. Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, Eliza Rutherford, Tom Hennigan, Jacob Menick, Albin Cassirer, Richard Powell, George van den Driessche, Lisa Anne Hendricks, Maribeth Rauh, Po-Sen Huang, Amelia Glaese, Johannes Welbl, Sumanth Dathathri, Saffron Huang, Jonathan Uesato, John Mellor, Irina Higgins, Antonia Creswell, Nat McAleese, Amy Wu, Erich Elsen, Siddhant Jayakumar, Elena Buchatskaya, David Budden, Esme Sutherland, Karen Simonyan, Michela Paganini, Laurent Sifre, Lena Martens, Xiang Lorraine Li, Adhiguna Kuncoro, Aida Nematzadeh, Elena Gribovskaya, Domenic Donato, Angeliki Lazaridou, Arthur Mensch, Jean-Baptiste Lespiau, Maria Tsimpoukelli, Nikolai Grigorev, Doug Fritz, Thibault Sottiaux, Mantas Pajarskas, Toby Pohlen, Zhitao Gong, Daniel Toyama, Cyprien de Masson d'Autume, Yujia Li, Tayfun Terzi, Vladimir Mikulik, Igor Babuschkin, Aidan Clark, Diego de Las Casas, Aurelia Guy, Chris Jones, James Bradbury, Matthew Johnson, Blake Hechtman, Laura Weidinger, Iason Gabriel, William Isaac, Ed Lockhart, Simon Osindero, Laura Rimell, Chris Dyer, Oriol Vinyals, Kareem Ayoub, Jeff Stanway, Lorrayne Bennett, Demis Hassabis, Koray Kavukcuoglu, and Geoffrey Irving. 2022. Scaling language models: Methods, analysis & insights from training Gopher. *arXiv preprint arXiv:2112.11446*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(1):140:1–140:67.

Oscar Sainz, Jon Campos, Iker García-Ferrero, Julen Etxaniz, Oier Lopez de Lacalle, and Eneko Agirre. 2023. NLP evaluation in trouble: On the need to measure LLM data contamination for each benchmark. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10776–10787, Singapore. Association for Computational Linguistics.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. 2023. BLOOM: A 176B-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.

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.

Luca Soldaini, Rodney Kinney, Akshita Bhagia, Dustin Schwenk, David Atkinson, Russell Authur, Ben Bogin, Khyathi Chandu, Jennifer Dumas, Yanai Elazar, Valentin Hofmann, Ananya Harsh Jha, Sachin Kumar, Li Lucy, Xinxi Lyu, Nathan Lambert, Ian Magnusson, Jacob Morrison, Niklas Muennighoff, Aakanksha Naik, Crystal Nam, Matthew E. Peters, Abhilasha Ravichander, Kyle Richardson, Zejiang Shen, Emma Strubell, Nishant Subramani, Oyvind Tafjord, Pete Walsh, Luke Zettlemoyer, Noah A. Smith, Hannaneh Hajishirzi, Iz Beltagy, Dirk Groeneveld, Jesse Dodge, and Kyle Lo. 2024. Dolma: an open corpus of three trillion tokens for language model pretraining research. *arXiv preprint arXiv:2402.00159*.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. 2023. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615*.

Gemini Team. 2023. Gemini: A family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

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, Marseille, France. European Language Resources Association.

Gorka Urbizu, Iñaki San Vicente, Xabier Saralegi, and Ander Corral. 2023. Not enough data to pre-train your language model? MT to the rescue! In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 3826–3836, Toronto, Canada. Association for Computational Linguistics.

Sean Wu, Michael Koo, Lesley Blum, Andy Black, Liyo Kao, Fabien Scalzo, and Ira Kurtz. 2023. A comparative study of open-source large language models, GPT-4 and Claude 2: Multiple-choice test taking in nephrology. *arXiv preprint arXiv:2308.04709*.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

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

Ahmet Üstün, Viraat Aryabumi, Zheng-Xin Yong, Wei-Yin Ko, Daniel D'souza, Gbemileke Onilude, Neel Bhandari, Shivalika Singh, Hui-Lee Ooi, Amr Kayid, Freddie Vargus, Phil Blunsom, Shayne Longpre, Niklas Muennighoff, Marzieh Fadaee, Julia Kreutzer, and Sara Hooker. 2024. Aya model: An instruction finetuned open-access multilingual language model. *arXiv preprint arXiv:2402.07827*.

A Data Preprocessing Details

The mix of raw documents —7.39 million documents, 2.96 billion words— was deduplicated with Dolma (Soldaini et al., 2024) to discard both intra- and cross-dataset document repetitions. The number of words discarded per dataset at this stage is illustrated in Figure 2(a), where datasets are shown in order of preference (Egunkaria and Booktegi are omitted due to their size). In the case of EusCrawl v1.1, the duplicate documents arise mainly from an overlap between EusCrawl v1 and the updated content. For Colossal OSCAR, more than 60% of the documents are already present in the preferred datasets (i.e., EusCrawl, Egunkaria, Booktegi, EuWiki or CulturaX). HPLT was further deduplicated at the paragraph level due to its quality. After deduplication, the pretraining corpus amounts to 5.80 million documents and 1.45 billion words.

The deduplicated corpus was further filtered based on document-level features. Specifically, we applied Dolma's implementation of a set of heuristics from Gopher (Rae et al., 2022) and C4 (Raffel et al., 2020), and the Corpus Cleaner v2 (CCv2; Palomar-Giner et al., 2024). The document-level features are as follows:

- `eu`: percentage of the text that is written in Basque according to CLD2
- `# words`: number of words
- `word len`: mean word length
- `bullet`: fraction of lines starting with '*' or '-'
- `ellipsis`: fraction of lines ending with '...'
- `lorem ipsum`: whether it occurs in the text
- `brackets`: whether '{' occurs in the text
- `symbols`: '#' and '...' to word ratio
- `alpha`: fraction of words with at least one alphabetic character
- `CCv2`: aggregated score given by CCv2

The threshold applied to each feature and its impact on the datasets is shown in Table 7. We adapted Dolma's default thresholds regarding document and word length to better fit our data and the Basque language based on observed distributions of our corpora (see examples in Figure 2). Note that the filters are not mutually exclusive, that is, the same document might be flagged for removal by several filters.

As a result of this process, the least curated corpora, HPLT and Colossal OSCAR, were further reduced by 61% and 5% respectively, in terms of words. The 22.37% EuWiki documents flagged as being too short correspond to Wikipedia's redirect pages, which ultimately did not affect the word count significantly. The final size of the pretraining corpus is 4.30 million documents and 1.22 billion words, which equates to 4.17B Llama 2 tokens.

B Dataset Contamination

Following previous work on data contamination (Brown et al., 2020; Touvron et al., 2023), we check for token n-gram overlaps between test items and training data. To that end, we index training documents in Elasticsearch, applying the standard tokenizer to lowercase text, and removing stopwords (built-in Basque stopwords and all auxiliary verbs).

Given that test items vary greatly in length from one benchmark to another, we avoid establishing an arbitrary n-gram length threshold above which to consider a test item contaminated. Instead, we report statistics based on the longest n-grams matched, spanning our search from n-grams equal to each item's total length to just one word. For each n-gram size n we only considered test items of equal or bigger size when assessing contamination.

Results are summarized in Table 8, which reports the contamination percentages (*cont %*) across each benchmark with respect to a specific quartile of test questions or context lengths (n). Higher contamination values are to be expected at lower quartiles, as shorter n-grams (typically 1 to 5 words) tend to involve frequent word combinations and are thus more likely to overlap with the training data. We observe that contamination percentages tend to decrease to near zero after the first quartile, with the following exceptions.

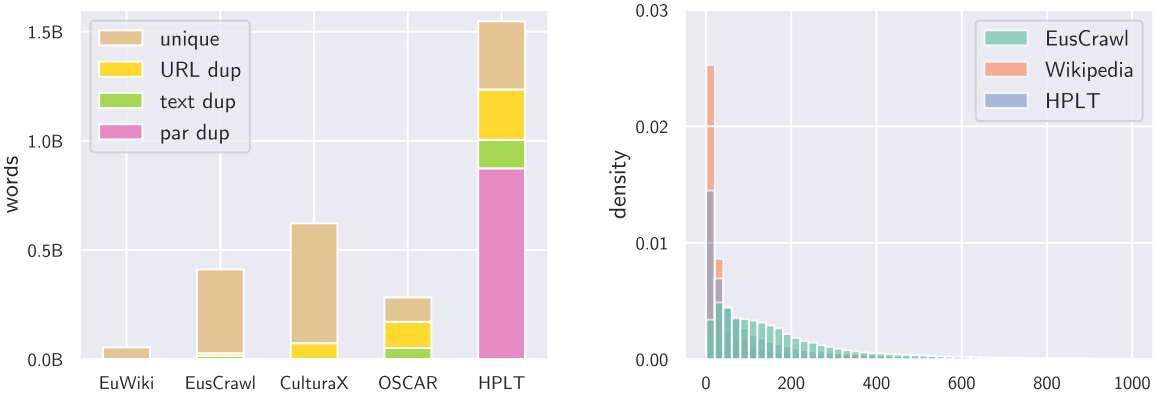
Notably, QNLI has substantial overlap even at higher n-gram lengths. This is explained by the fact that QNLI contexts were taken from Wikipedia. In line with the analysis of Chowdhery et al. (2023), we do not consider these items to be contaminated, as the questions and answers have not been found alongside the contexts in the training data.

In the case of EusProficiency, it must be noted it comprises particularly short test items, the median length of a question being 4 words. Upon manual analysis, we did not observe any annotation contamination.

As for EusExams, this evaluation benchmark consists of Public Service examinations and thus contains many references to national and regional laws, directives and plans, services, administration offices, etc. Such mentions can also be found on the

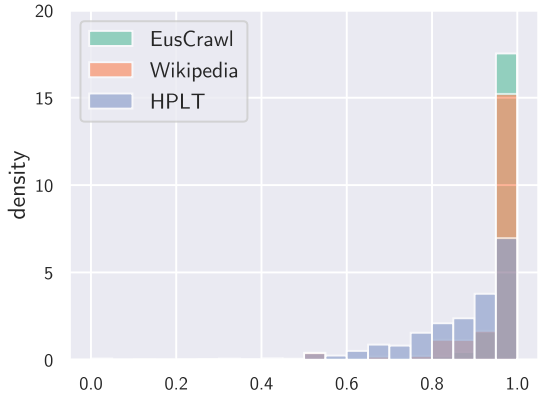
	CulturaX	EusCrawl	HPLT	OSCAR	EuWiki	Egunkaria	Booktegi
eu < 0.5	0.00	5.30	34.13	0.00	0.00	0.00	0.00
# words < 4	0.00	0.04	3.61	0.13	22.37	0.06	0.00
word len < 3	0.00	0.04	0.66	0.02	0.87	0.11	0.00
word len > 12	0.00	1.12	0.88	0.36	1.26	0.00	0.00
alpha < 0.8	0.11	1.21	23.93	19.14	4.19	0.84	0.00
symbols > 0.1	0.09	0.05	0.47	0.23	0.02	0.00	0.00
ellipsis > 0.3	0.74	0.09	0.25	3.47	0.00	0.01	0.00
bullets > 0.9	0.03	0.01	0.01	0.17	0.00	0.00	0.00
lorem ipsum	0.00	0.00	0.02	0.31	0.00	0.00	0.00
brackets	0.43	0.07	0.73	16.30	0.23	0.01	1.13
CCv2 < 0.55	0.04	0.56	25.12	0.11	1.62	1.19	5.08

Table 7: Percentage of documents dropped by each filter.

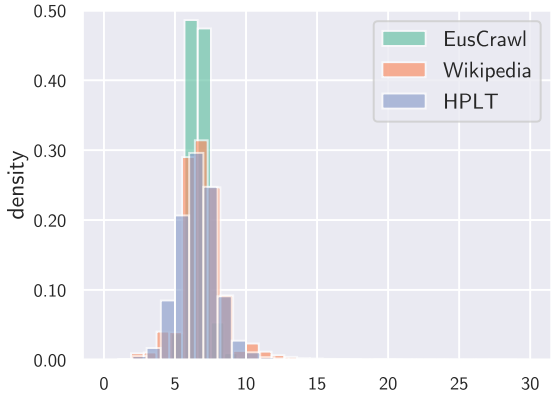


(a) Cross-dataset document repetitions

(b) # words: Words per document



(c) alpha: alphabetic characters to word length ratio



(d) word len: Medium characters per word

Figure 2: Basic corpus quality statistics before preprocessing

web, as this information is of public interest. Again, we did not observe annotation contamination.

C Task Examples and Prompts

Tables 9 and 10 contain, respectively, the prompt templates and examples of our new datasets (i.e., EusProficiency, EusReading, EusTrivia, and EusExams) and BasqueGLUE tasks. "*System*" refers to system prompts and applies only to GPT evaluations. Additionally, the number of shots and metrics used to measure the results are also specified per task.

D Model Card

We report Latxa's model card in Table 11.

E Detailed EusExams Results

We provide detailed results for EusExams by category in Table 12. Results are consistent across categories, our models outperform every model in the same size category by a large margin. Latxa 13B outperforms GPT-3.5 Turbo in most categories, but Latxa 70B is far from GPT-4 Turbo performance. This is expected as all categories in these exams require advanced knowledge. Health System is the most challenging category, followed by City Council. Public Office and University tests are easier for most models. For specific results of each test check Table 13.

	min		25%		50%		75%		max	
	n	**cont %**	**n**	**cont %**	**n**	**cont %**	**n**	**cont %**	**n**	**cont %**
Belebele	18	0.0	40	0.0	51	0.0	64	0.0	138	0.0
XStory Cloze	1	100.0	4	17.4	5	1.3	6	0.1	12	0.0
EusProficiency	1	99.7	3	34.1	4	6.4	6	0.6	21	0.0
EusReading	1	100.0	5	88.1	57	0.0	578	0.0	808	0.0
EusTrivia	1	100.0	4	7.1	5	1.3	7	0.0	21	0.0
EusExams	1	96.6	6	13.6	9	7.1	15	0.1	105	0.0
BEC	1	100.0	8	0.2	11	0.0	13	0.0	34	0.0
BHTCv1	1	99.8	19	2.8	25	0.4	34	0.1	147	0.0
Korref	8	0.0	21	0.0	25	0.0	33	0.0	66	0.0
QNLI	1	100.0	3	97.1	5	71.0	10	16.4	84	0.0
VaxxStance	2	99.4	15	0.3	22	0.0	29	0.0	39	0.0
WiC	2	100.0	12	2.6	18	0.4	26	0.0	42	0.0

Table 8: Data contamination results for all our evaluation datasets. The table shows the contamination percentage (*cont %*) considering different n-gram sizes (*n*) that depend on the length of each dataset's items.

EusProficiency *5-shot, accuracy*

System Prompt: Respond always with a single letter: A, B, C or D.

Prompt: `Galdera: {question}\nA. {opt[0]}\nB. {opt[1]} ... \nErantzuna:`

Example:
Galdera: Jatetxe batera sartu, eta bazkaltzen ari denari:
A. Gabon!
B. On egin diezazula!
C. Bejondeizula!
D. Agur t'erdi!
Erantzuna: B

Question: Upon entering a restaurant, to another diner: A. Good night! B. Enjoy! C. Bless you! D. Greetings! Answer: B

EusReading *1-shot, accuracy*

System Prompt: Respond always with a single letter: A, B, C or D.

Prompt: `Pasartea: {context}\n\nGaldera: {question}\nA. {opt[0]}\nB. {opt[1]} ... \nErantzuna:`

Example:
Pasartea: Ernest Hemingway, berak jakin barik, azkenekoz etorri da Bilbora, eta oro har, Penintsulara. Eta hori tamala, hilak 24 dituelarik Bilbon zezenak Ordoñez bere kutunari adarkada ederra sartu dio. Ez da ezer izan, zorionez. Biharamuneko El Correo Español egunkarian emandako argazkian ikusten den legez, idazleak bisita egin dion unean, toreatzailea hortxe dago, ondo bizirik, ohean. [...]

Galdera: 1960ko abuztuaren 24an
A. El Correo Español-eko C. Barrenarekin batera agertzen den Ordoñez toreatzailea harrapatu zuen zezen batek.
B. Ernest Hemingwayk Bilboko plazan adarkada jaso zuen Ordoñez toreatzaileari bisita egin zion.
C. El Correo-ko argazkian, zezen batek Ordoñez toreatzailea harrapatzen du.
D. Ernest Hemingway lehenengo eta azkeneko aldiz iritsi zen Bilbora.
Erantzuna: B

Passage: Ernest Hemingway, without his knowledge, came for the last time to Bilbao, and in general to the Peninsula. And indeed, on the 24th, at Bilbao, the bull gave his favourite Ordóñez a good goring. It was nothing, luckily. As can be seen from the photograph published in El Correo Español the next day, the bullfighter is there, alive, in bed, the moment the writer visits him. [...] Question: On August 24, 1960 A. The bullfighter Ordóñez, who appears next to C. Barrena of El Correo Español, was caught by a bull. B. Ernest Hemingway visited the bullfighter Ordóñez, who had received a goring in the square of Bilbao. C. In the photo of El Correo, a bull catches the bullfighter Ordóñez. D. Ernest Hemingway arrived in Bilbao for the first and last time. Answer: B

EusTrivia *5-shot, accuracy*

System Prompt: Respond always with a single letter: A, B, C or D.

Prompt: `Galdera: {question}\nA. {opt[0]}\nB. {opt[1]} ... \nErantzuna:`

Example:
Galdera: Zenbat kilo dauka tona batek?
A. 10.000 kilo
B. 1.000.000 kilo
C. 1.000 kilo
D. 100 kilo
Erantzuna: C

Question: How many kilograms are there in a tonne? A. 10,000 kilos B. 1,000,000 kilos C. 1,000 kilos D. 100 kilos Answer: C

EusExams *5-shot, accuracy*

System Prompt: Respond always with a single letter: A, B, C or D.

Prompt: `Galdera: {question}\nA. {opt[0]}\nB. {opt[1]} ... \nErantzuna:`

Example:
Galdera: UPV/EHUREN ONDAREA HAU DA:
A. UPV/EHUk jabetzan dituen ondasunak.
B. UPV/EHUk jabetzan dituen ondasun eta eskubideak.
C. UPV/EHUk jabetzan edo titularitatean dituen ondasun eta eskubideak, bai eta etorkizunean eskuratzen edo esleitzen zaizkion gainerako guztiak ere.
D. UPV/EHUk jabetzan dituen ondasunak, bai eta etorkizunean eskuratzen dituen gainerako guztiak ere.
Erantzuna: C

Question: UPV/EHU'S LEGACY IS: A. The property owned by UPV/EHU. B. The rights and property owned by the UPV/EHU. C. The rights and property of the UPV/EHU in ownership, as well as any other property acquired or assigned to it in the future. D. The property of the UPV/EHU in ownership, as well as any other property acquired or assigned to it in the future. Answer: C

Table 9: Prompt templates and examples of the new evaluation tasks. "*System*" refers to system prompts and only applies to GPT evaluations. Translations of the examples are given in *italics*.

BEC *5-shot, micro F1*

System: Respond always with one of these: negatiboa, neutrala, positiboa

Prompt: `Testua: {context}\nGaldera: Nolako jarrera agertzen du aurreko testuak?\nErantzuna:`

Example:
Testua: Eta Euskal Herrian noizko @eajpnv ? #URL
Galdera: Nolako jarrera agertzen du aurreko testuak?
Erantzuna: negatiboa

Text: And in the Basque Country when @eajpnv ? #URL Question: What sentiment does the previous text convey? Answer: negative

VaxxStance *5-shot, macro F1 (*N FAVOUR & AGAINST*)*

System: Respond always with one of these: bai, ez

Prompt: `Testua: {context}\nGaldera: Nolako jarrera agertzen du aurreko testuak txertoei buruz?\nErantzuna:`

Example:
Testua: 45 urtetik gorakoen txertaketa hasiko da igandean Israelenhttps://t.co/6opid1ULyd
Galdera: Nolako jarrera agertzen du aurreko testuak txertoei buruz?
Erantzuna: neutrala

Text: vaccination of over 45 years of age begins on Sunday in Israelhttps://t.co/6opid1ULyd Question: What stance does the previous text take on vaccines? Answer: neutral

BHTCv2 *5-shot, micro F1*

System: Respond always with one of these: Ekonomia, Euskal Herria, Euskara, Gizartea, Historia, Ingurumena, Iritzia, Komunikazioa, Kultura, Nazioartea, Politika, Zientzia

Prompt: `Testua: {context}\nGaldera: Zein da aurreko testuaren gaia?\nErantzuna:`

Example:
Testua: Eusko Jaurlaritza ari da prestatzen eta EAEko ikastetxe guztietara zabaltzeko asmoa du.
Galdera: Zein da aurreko testuaren gaia?
Erantzuna: Gizartea

Text: The Basque Government is preparing it and intends to extend it to all schools in the Basque Country. Question: What is the subject of the previous text? Answer: Society

EpecKorrefBin *5-shot, accuracy*

System: Respond always with one of these: ez, bai

Prompt: `Testua: {context}\nGaldera: Aurreko testuan, "*{w1}*" eta "*{w2}*" gauza bera dira?\nErantzuna:`

Example:
Testua: *Luis Uranga* harrituta azaldu da Portugalgo klubaren jokaerarekin. RICARDO SA PINTOK datorren denboraldian Lisboako Sportingen jokatu zuela pentsatzen genuen guztiok, baina une honetan Lisboarako bidea erabat zaildu zaio *Realeko aurrelariari*.
Galdera: Aurreko testuan, "*Luis Uranga*" eta "*Realeko aurrelariari*" gauza bera dira?
Erantzuna: ez

*Question: *Luis Uranga* was surprised by the way the Portuguese club acted. We all thought RICARDO SA PINTO would be playing next season at the Sporting of Lisbon, but right now the road to Lisbon has become very difficult for *the forward of La Real*. In the previous text, is "*Luis Uranga*" the same as "*Realeko aurrelariari*"? Answer: no*

QNLI$_{eu}$ *5-shot, accuracy*

System: Respond always with one of these: bai, ez

Prompt: `{question}\n{context}\nGaldera: aurreko galderari erantzuten al dio emandako testuak?\nErantzuna:`

Example:
Nortzuen lehen alaba izan zen Dua Lipa?
Liparen lehen hezkuntzako ikasketak musika klaseak eduki zituen, eta jotzen ikasi zuen lehen instrumentua biolontxeloa izan zen.
Galdera: aurreko galderari erantzuten al dio emandako testuak?
Erantzuna: ez

Whose first daugther was Dua Lipa? Lipa's primary education included music lessons, and the first instrument she learned to play was the cello. Question: does the text given answer the previous question? Answer: No.

WiC$_{eu}$ *5-shot, accuracy*

System: Respond always with one of these: bai, ez

Prompt: `1. esaldia: {sent[0]}\n2. esaldia: {sent[1]}\nGaldera: Aurreko bi esaldietan, "{word}" hitzak esanahi berdina du?\nErantzuna:`

Example:
1. esaldia: beste alde batetik, irakasleek materiala prestatzeko dituzten aukera informatikoak ere gero eta ugariagoak dira;
2. esaldia: Unitate horretan konturatuko zinen bezala, materialak aldakorrak dira: batzuk lurrindu egiten dira berotzen direnean, beste batzuk apurtu edo eraldatu, edo aldaketa kimikoak jasan ditzakete.
Galdera: Aurreko bi esaldietan, "material" hitzak esanahi berdina du?
Erantzuna: ez

Sentence 1: on the other hand, the computer possibilities for teachers to prepare materials are increasing; Sentence 2: As you may have noticed in that unit, materials are changeable: some evaporate when heated, others brake or transform, or they may undergo chemical changes. Question: In the two previous sentences, does the word "material" have the same meaning? Answer: no

Table 10: Prompt templates and examples of BasqueGLUE tasks. "*System*" refers to system prompts and only applies to GPT evaluations. Translations of the examples are given in *italics*.

Model Details	
Model Developers	(Anonymous upon publication)
Variations	Latxa comes in a range of parameter sizes: 7B, 13B, and 70B.
Input	Models input text only.
Output	Models generate text only.
Model Architecture	Latxa, similar to Llama 2, is an auto-regressive language model that uses an optimized transformer architecture.
Model Dates	Latxa was trained between October 2023 and February 2024.
Status	This is a static model trained on an offline dataset. Future versions of the model may include more updated data.
License	Latxa is based on Llama 2 models, and therefore, inherits their license. It is a custom commercial license available at: https://ai.meta.com/resources/models-and-libraries/llama-downloads/
Where to send comments	(Anonymous upon publication)
Intended Use	
Intended Use Cases	Latxa models are intended to be used with Basque data; for any other language, the performance is not guaranteed. Latxa inherits the Llama 2 License which allows for commercial and research use. Latxa family models are pretrained LLMs without any task-specific or instruction fine-tuning. That is, the model can either be prompted to perform a specific task or further fine-tuned for specific use cases.
Out-of-Scope Uses	The model was not fine-tuned to follow instructions or to work as a chat assistant, therefore, this kind of usage is not tested nor recommended.
Hardware and Software (Section 3)	
Training Factors	The training of Latxa was conducted using GPT-Neox library. As infrastructure, we leveraged the CINECA HPC Leonardo computing cluster located in Italy. At most, 256 custom A100 GPUs were used to train the models.
Carbon Footprint	Pretraining utilized a cumulative 34.7K GPU hours of computation on hardware of type A100 64Gb (TDP 440W). Estimated total emissions were 4.53tCO_2eq.
Training Data (Section 2)	
Overview	Latxa is trained on corpora from different sources. In general, quality was preferred over quantity, but content derived from automatically filtered versions of CommonCrawl was also included. After collecting the corpora, it was cleaned and deduplicated. Pretraining corpora includes: EusCrawl v1.1, Egunkaria, Booktegi, EuWiki, CulturaX, Colossal OSCAR, and, HLPT v1.
Data Freshness	The pretraining data has a cutoff of November 2023.
Evaluation	
	See Evaluation Data (Section 4), Experimental Setting (Section 5), and Results (Section 6)
Ethical Considerations and Limitations (Section 8)	
	To alleviate the potentially disturbing or harmful content, Latxa has been trained on carefully selected and processed data which comes mainly from local media, national/regional newspapers, encyclopedias and blogs. Still, the model is based on Llama 2 models and can potentially carry the same biases, risks and limitations.

Table 11: Model card for Latxa

		Public Office	University	City Council	Health System	Average
GPT-3.5 Turbo	n/a	47.29	43.43	41.61	40.50	42.42
GPT-4 Turbo	n/a	**76.64**	**76.22**	**69.63**	**64.61**	**70.22**
XGLM	7B	23.82	23.38	24.82	25.88	24.75
BLOOM	7B	24.08	24.02	25.56	24.92	24.57
Mistral	7B	34.72	33.38	27.84	28.37	30.76
Llama 2	7B	30.35	27.68	31.98	28.26	28.63
Latxa	7B	**37.22**	**37.29**	**35.24**	**29.39**	**33.32**
mGPT	13B	25.01	24.47	24.80	28.09	26.31
Llama 2	13B	31.99	30.52	26.81	26.52	28.54
Latxa	13B	**48.78**	**50.35**	**45.52**	**36.66**	**43.19**
Mixtral	8x7B	44.87	42.73	38.47	35.54	39.26
Yi	34B	41.13	41.76	36.61	36.18	38.62
Llama 2	70B	37.42	34.75	30.78	27.94	31.55
Latxa	70B	**58.16**	**56.26**	**50.04**	**43.78**	**50.07**

Table 12: Detailed accuracy results over EusExams categories. Best results in each compute class are in **bold**. Best overall results are underlined.

	GPT-3.5 Turbo	GPT-4 Turbo	XGLM 7B	BLOOM 7B	Mistral 7B	Llama 2 7B	Latxa 7B	mGPT 13B	Llama 2 13B	Latxa 13B	Mixtral 8x7B	Yi 34B	Llama 2 70B	Latxa 70B
Admin staff 2022	46.98	**72.70**	22.70	22.99	**34.05**	26.58	31.18	22.84	30.89	**43.53**	41.24	41.24	37.93	**56.03**
Support staff 2022	49.40	**82.13**	25.90	24.50	32.13	31.93	**43.57**	26.91	32.93	**52.61**	47.79	41.37	34.14	**59.84**
Admin assistant 2022	47.35	**77.62**	24.39	26.83	36.30	31.28	**38.16**	26.26	33.14	**49.35**	45.34	41.89	39.89	**57.53**
General questions 2022	45.41	**74.09**	22.27	21.98	**36.39**	31.59	35.95	24.02	31.00	**49.64**	45.12	40.03	37.70	**59.24**
Public Office	47.29	**76.64**	23.82	24.08	34.72	30.35	**37.22**	25.01	31.99	**48.78**	44.87	41.13	37.42	**58.16**
Bilbao Council 2022	42.38	**71.75**	22.70	26.67	33.17	30.63	**36.03**	24.60	28.89	**43.81**	40.00	36.83	32.38	**50.63**
Gasteiz Council 2021	40.83	**67.50**	26.94	24.44	22.50	33.33	**34.44**	25.00	24.72	**47.22**	36.94	36.39	29.17	**49.44**
City Council	41.61	**69.63**	24.82	25.56	27.84	31.98	**35.24**	24.80	26.81	**45.52**	38.47	36.61	30.78	**50.04**
Admin staff 2019	48.30	**79.56**	26.25	21.64	33.47	29.66	**37.68**	23.85	29.66	**54.91**	44.89	43.09	34.27	**60.12**
Admin assistant 2019	47.33	**81.11**	24.00	24.67	36.89	28.22	**42.44**	27.56	33.33	**58.89**	45.33	48.44	40.00	**64.44**
Library assistant 2019	45.41	**79.63**	22.87	26.04	**37.73**	25.04	32.22	25.38	29.72	**48.58**	42.24	43.91	35.06	**56.26**
Law 2019	36.14	**69.29**	21.71	24.00	31.71	28.71	**33.14**	22.29	30.14	**42.14**	38.86	36.57	30.00	**48.14**
Economics 2019	37.61	**73.22**	23.08	23.36	31.34	25.93	**31.91**	25.36	30.48	**44.73**	41.31	36.75	34.47	**52.99**
Business Admin 2019	43.21	**73.21**	19.64	23.21	29.29	27.86	**40.71**	23.93	30.36	**53.21**	42.50	39.64	32.86	**56.07**
Auxiliary staff 2019	45.25	**80.75**	26.50	24.25	35.25	27.50	**41.75**	24.75	32.25	**54.00**	46.00	47.00	39.00	**59.50**
Admin Tech. School (A) 2019	39.91	**73.39**	24.03	24.46	34.19	27.18	**34.05**	22.75	29.33	**46.78**	41.20	39.34	34.33	**53.08**
Admin Tech. School (B) 2019	47.75	**75.79**	22.37	24.54	**36.23**	29.05	34.89	24.37	29.38	**49.92**	42.24	41.07	32.72	**55.76**
University	43.43	**76.22**	23.38	24.02	33.38	27.68	**37.29**	24.47	30.52	**50.35**	42.73	41.76	34.75	**56.26**
Admin staff 2023	42.37	**61.44**	**28.81**	22.46	25.00	23.31	26.27	27.54	22.46	**36.86**	36.86	34.75	22.46	**43.64**
Health assistant 2023	34.73	**57.49**	22.16	27.54	17.37	**29.94**	21.56	29.94	22.16	**35.93**	31.74	33.53	20.36	**36.53**
Admin assistant 2023	37.58	**60.00**	25.45	22.42	26.67	27.27	**28.48**	29.09	23.03	**30.30**	31.52	**35.15**	23.03	32.73
Hospital porter 2023	33.13	**63.19**	22.70	25.77	26.38	28.22	**32.52**	33.13	25.15	**33.74**	35.58	30.06	26.99	**39.26**
Medical staff 2023	39.14	**60.59**	26.01	21.45	23.59	24.66	**25.74**	28.15	22.25	**36.19**	33.24	35.12	27.61	**39.68**
Service operator 2023	36.64	**58.02**	25.19	19.85	**29.01**	28.24	28.24	28.24	22.90	**32.06**	34.35	32.06	20.61	**42.75**
Superior technician 2023	40.19	**58.88**	**30.84**	24.92	25.86	28.35	26.17	28.04	26.48	**37.38**	33.33	33.33	27.41	**36.76**
Misc (cook, janitor, etc.) 2023	38.72	**61.65**	28.57	24.06	24.44	26.69	**29.70**	28.95	24.06	**38.72**	33.83	33.46	22.18	**39.47**
Admin assistant 2008	37.25	**57.96**	24.57	22.41	29.98	25.81	**28.13**	23.49	27.82	**31.68**	28.59	33.08	29.52	**44.05**
Admin staff 2008	36.63	**59.22**	24.87	24.06	**32.49**	28.48	30.35	24.47	28.61	**37.43**	35.29	35.56	31.95	**44.12**
Hospital porter 2008	42.08	**70.31**	26.96	28.05	34.24	**34.43**	32.24	26.23	29.33	**42.26**	40.98	41.71	35.34	**54.28**
Auxiliary nurse 2008	52.77	**81.69**	25.54	28.92	**35.23**	30.46	32.46	29.08	34.31	**44.46**	43.69	43.38	37.38	**56.00**
Nurse 2008	51.90	**81.00**	25.60	31.40	36.30	32.10	**36.50**	32.40	33.20	**43.60**	41.50	46.00	34.40	**57.30**
Family doctor 2008	43.83	**73.09**	25.01	25.52	30.65	27.69	**31.25**	24.50	29.54	**37.45**	37.08	39.34	31.85	**49.24**
Health System	40.50	**64.61**	25.88	24.92	28.37	28.26	**29.39**	28.09	26.52	**36.66**	35.54	36.18	27.94	**43.78**
Average	42.42	**70.22**	24.75	24.57	30.76	28.63	**33.32**	26.31	28.54	**43.19**	39.26	38.62	31.55	**50.07**

Table 13: Detailed results on EusExams tests and categories (in **bold**). Best results in each compute class are in **bold**. Best overall results are underlined.