LatamGPT 70B Technical Report

Omar U. Florez[1], Carlos Aspillaga[1], Alexandra García[1], Eugenio Herrera-Berg[1], Gonzalo Fuentes[1], Manuel Cifuentes[1], Tomás Couso[1], Sebastián Cifuentes[1], Rubén Quintupil[1], Alexandra Arriagada[1], Clemente Henríquez, Gabriela Arriagada[1,2], Alexandra Davidoff[1], Romina Torres[1,3], Sebastián Donoso[1], Tomás Gómez[1], Alejandro Antilao[4], Andrés Cerda[4], Mauricio Leiva[4], Álvaro Paredes[4], Jose Navarro[3], Alvaro Soto[1,2]

LatamGPT Team (CENIA)[1]

[1]Chilean National Center for AI
[2]Pontificia Universidad Católica de Chile
[3]Universidad Adolfo Ibañez
[4]Data Observatory

Date: June 2026 **Website:** https://www.latamgpt.org/

Abstract

We present LatamGPT 70 B, a fully open foundation language model continually pre-trained and post-trained on data collected primarily from 20 Latin American countries in Spanish and Portuguese. LatamGPT is designed both to strengthen regional capacity for developing large language models and to improve the representation of Latin American factual, linguistic, and cultural knowledge in open models. We make four main technical contributions: (1) The LatamGPT-Corpus, a 296.5 B-token pre-training corpus covering 20 Latin American countries, including 58.8 B tokens contributed directly by 79 academic, governmental, and civil society institutions; (2) A reproducible pipeline for collecting, governing, filtering, and processing regionally grounded multilingual data; (3) A continual pre-training methodology at 70 B scale that addresses the tradeoff between regional adaptation and the preservation of the base model's general capabilities; and (4) A post-training recipe combining supervised fine-tuning, safety tuning, and multiple-choice question fine-tuning. We evaluate LatamGPT on standard benchmarks and on Choclo and Trueque, two benchmarks designed to assess Latin American factual and cultural knowledge. Alongside model weights, we release the corpus, data-processing pipeline, and training recipes to support further development of AI in Latin America.

1 Introduction

1.1 Motivation and Scope

Latin America represents 6.6% of global GDP and 8.8% of the world's population, yet receives only 1.12% of global AI investment [Centro Nacional de Inteligencia Artificial, 2024, Soto et al., 2026]. This investment gap is reflected in the training and data foundations of current large language models (LLMs). In effect, LLMs development has concentrated in regions with abundant compute infrastructure and digitized text corpora, leaving regions such as Latin America at the margins of model development. Studies of Common Crawl, the primary web-scale source used for pre-training LLMs, further identify a structural bias toward English-speaking countries with strong digital infrastructure, leaving Latin American Spanish and Portuguese underrepresented relative to

the region's economic and demographic weight [Baack, 2024]. This asymmetry propagates into the capabilities of widely deployed models: while they perform well on global benchmarks, they exhibit systematic gaps in Latin American factual knowledge, culturally specific content, and dialectal variation within Spanish and Portuguese.

LatamGPT 70 B is a fully open LLM built primarily on data from 20 Latin American countries. Its development is motivated by two complementary objectives. First, it seeks to strengthen regional technical capacity for developing, training, and evaluating LLMs in Latin America and the Caribbean. Second, it aims to address gaps in the representation of Latin American factual, linguistic, historical, and cultural knowledge in current open models. In this sense, LatamGPT is conceived not only as a language model, but also as a public technological asset designed to reduce regional AI gaps and empower Latin American societies through infrastructure grounded in their own cultural and historical contexts.

Three successive waves of open LLM development frame the context for its design. The first wave, exemplified by BERT [Devlin et al., 2019] and GPT-2 [Radford et al., 2019], established the potential of pretrained language models and demonstrated the role of openly released model weights in advancing the state of the art in AI. The second wave, initiated around 2020, was marked by large-scale open models such as BLOOM (176 B; BigScience Workshop et al. 2022), which advanced transparency through extensive documentation of multilingual training data and provided an influential example of participatory, multidisciplinary model development. The third wave, beginning around 2023, is characterized by explicit regional strategies: OLMo [Groeneveld et al., 2024] pioneered full data and pipeline openness in the United States; Falcon [Almazrouei et al., 2023] demonstrated the development of competitive open-weight models in the Gulf region; Qwen [Bai et al., 2023] optimized for Chinese and English at scale in Asia; and SEA-LION [Ng et al., 2025] targeted Southeast Asian language ecosystems. Across these initiatives, a consistent pattern emerges: performance on regional tasks depends not only on model scale, but also on data quality, representativeness, and cultural grounding.

LatamGPT 70 B represents a Latin American initiative within this third wave. To our knowledge, it is the largest open LLM continually pre-trained on a corpus specifically designed to represent the languages, knowledge, and cultural contexts of Latin America. The model is initialized from Llama 3.1 70 B [Dubey et al., 2024], a strong multilingual open-weight foundation model, and continually pre-trained on the LatamGPT-Corpus, a 296.5 B-token corpus assembled through large-scale public-data acquisition and formal data-sharing collaborations with 79 institutions across 20 Latin American countries. Unlike multilingual models that rely predominantly on web-scale sources such as Common Crawl, LatamGPT incorporates 58.8 B tokens contributed directly by academic, governmental, and civil society institutions across the region. These contributions provide access to locally relevant sources that are difficult to capture through public web crawls alone.

We position LatamGPT not merely as a model intended to improve performance on regional benchmarks, but as fully open infrastructure for Latin American AI development. In contrast to open-weight releases that provide only the trained model, we release the model weights, the LatamGPT-Corpus, and the complete data-collection and preprocessing pipeline, together with reproducible recipes for continual pre-training (CPT), supervised fine-tuning (SFT), and safety alignment. By making each stage of the development process available, LatamGPT provides a foundation on which institutions and other resource-constrained initiatives in Latin America and beyond can build, adapt, and extend regionally grounded language technologies.

1.2 Model Overview

LatamGPT 70 B is developed through a three-stage pipeline. First, we construct the Latam Corpus, a 296.5 B-token pre-training corpus compiled from web, institutional, and academic sources. Second, we perform continual pre-training (CPT) of Llama 3.1 70 B using the Latam Corpus and 35 B replay tokens from Llama's original pre-training distribution. Third, we apply supervised fine-tuning (SFT) on a multilingual instruction mixture. Table 1 summarizes the model's key properties.

Table 1: LatamGPT 70B model properties.

Property	**Value**
Base model	Llama 3.1 70B
Parameters	70.6B
Context length	8,192 tokens
CPT corpus (total)	296.5 B tokens
Common Crawl (LatAm)	220.4 B tokens (74.3%)
Partnerships	58.8 B tokens (19.8%)
Synthetic	17.3 B tokens (5.8%)
Countries covered	20
Languages	Spanish, Portuguese
Precision	Mixed (FP8 forward/backward, bf16 optimizer)

1.3 Key Contributions

- The LatamGPT-Corpus, a 296.5 B-token pre-training corpus predominantly in Spanish and Portuguese, with data sourced primarily from 20 Latin American countries, including 58.8,B tokens contributed directly by 79 academic, governmental, and civil society institutions
- A reproducible pipeline for collecting, governing, filtering, and processing web and institutional data across Latin America, supported by documented governance procedures and a three-tier data licensing framework
- A continual pre-training methodology at 70 B-parameter scale that addresses the tradeoff between regional adaptation and preservation of the base model's general capabilities, with ablations over learning rates, replay-token ratios, and training-phase composition
- A post-training recipe combining supervised fine-tuning, safety tuning, and multiple-choice question fine-tuning
- An empirical evaluation on standard benchmarks and on Choclo and Trueque, two benchmarks designed to assess Latin American factual and cultural knowledge, comparing LatamGPT with open models of comparable scale

2 Data

2.1 Pre-training Data Construction

This section describes how the Latam Corpus is constructed, from raw data collection to the sequences consumed by the training framework. We first present the collection and filtering pipeline that

produces the final 296.5 B-token corpus. We then describe the distributed tokenization system that converts the corpus into fixed-length sequences, followed by the assembly stage that selects and assigns the final training mixture to different phases.

2.1.1 Data Pipeline

The pre-training corpus combines large-scale data collection from public sources with datasets contributed through regional partnerships. Public data was acquired through web scraping and direct downloads from repositories such as Hugging Face and GitHub, while partner organizations contributed their datasets through shared Google Cloud buckets. This approach combines widely available web content with locally relevant sources that are not publicly available on the web. All sources enter a common pipeline for validation, filtering, anonymization, and preprocessing. Additional details on data governance, licensing, and transparency will be provided in a separate technical report, containing the LatamGPT data catalog.

The corpus draws from four complementary collection streams. **(1) Targeted web scraping** covers regionally curated sources, including open-access academic repositories, government- and non-governmental organization-backed cultural archives, and more than one thousand validated regional news outlets and digital media sites. **(2) Data donations and partnerships** incorporate corpora contributed by regional institutions, including previously unpublished content reviewed for privacy, copyright, and licensing compliance. **(3) Targeted Common Crawl extraction** filters public Common Crawl snapshots for documents that originate from or substantively discuss Latin American countries and regions. **(4) A synthetic data pipeline** supplements countries and domains that remain underrepresented in naturally collected web content by generating targeted documents that improve geographic and thematic balance.

The data pipeline produces three distinct data volumes, each marking a stage of progressive refinement. Initial collection yields approximately 2.8 trillion tokens. Preprocessing, which includes dataset splitting, metadata generation, language filtering, deduplication, heuristic filtering, and anonymization, reduces this volume to 1.7 trillion tokens. Two subsequent filtering phases produce the final 296.5 B-token pre-training corpus:

$$2.8\,\text{T collected tokens} \longrightarrow 1.7\,\text{T preprocessed tokens} \longrightarrow 296.5\,\text{B final corpus tokens.}$$

1. **Phase 1 filters** reduce the 1.7 T-token preprocessed corpus to 429.9 B tokens, a 74.9% reduction. These filters evaluate Latin American relevance, license compliance, toxicity, and semantic quality. Only documents in Spanish, Portuguese, and English are retained. Country and relevance labels are assigned by custom XLM-RoBERTa models [Conneau et al., 2020]. Topic classification uses a custom FastText model, while toxicity filtering relies on an XLM-RoBERTa model trained through teacher-student distillation. These filters apply exclusively to Common Crawl documents. Partnership-contributed data is retained without country-level filtering because its Latin American provenance has already been established by the contributing institutions.
2. **Phase 2 filters** reduce the corpus from 429.9 B to 296.5 B tokens, a further 31% reduction. This phase removes documents with low syntactic coherence that remain after Phase 1, concentrating the final corpus on higher-quality and more informative examples.

The annotation stage follows a classifier-based strategy similar to FineWeb-Edu [Penedo et al., 2024]. Model-based classifiers approximate relevance, quality, and safety judgments at web-corpus scale. The resulting metadata supports monitoring and rebalancing by country, topic, language, and source.

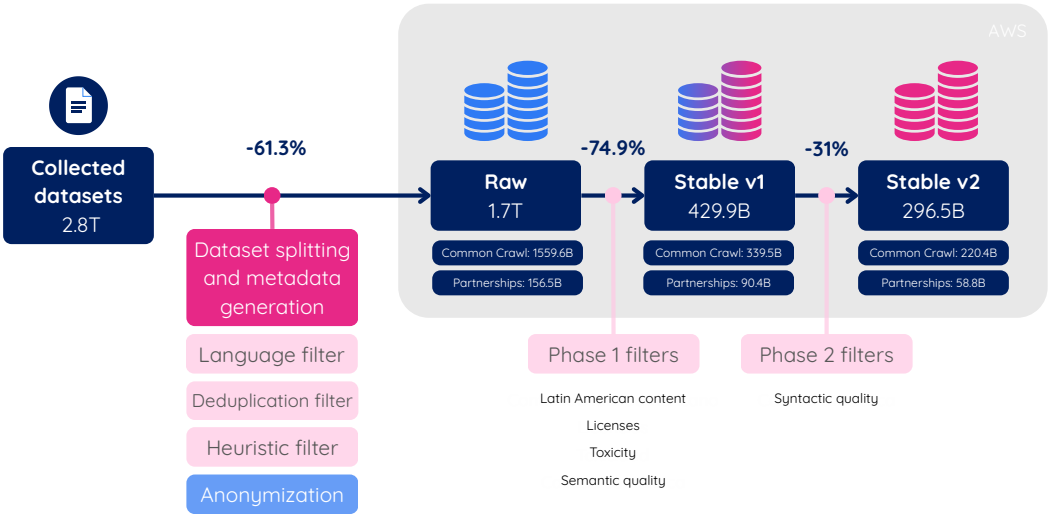


Figure 1: Diagram of the data collection, labeling, and preprocessing pipeline.

Table 2: Custom models used in the metadata annotation stage of the processing pipeline.

Component	Architecture	Target Classes	Output
Language ID	FastText (fine-tuned)	ES, PT, EN, Other	Language label
Country Detection	XLM-RoBERTa (fine-tuned)	20 LatAm countries	Multi-label presence
Topic Classification	FastText (fine-tuned)	10-topic ontology	Multi-label topic
Toxicity / Safety	XLM-RoBERTa (teacher-student)	Safe, Toxic, Explicit	Safety label
LatAm Relevance	XLM-RoBERTa (fine-tuned)	Binary (Yes / No)	Relevance score

Figure 1 summarizes the collection, labeling, preprocessing, and quality-control stages. The main operational risks are excessive filtering, tokenization and packing bottlenecks, and overly aggressive deduplication. We monitor these risks using per-stage token counts, redaction and deduplication rates, and processing latency. Unexpected losses at any stage trigger manual review.

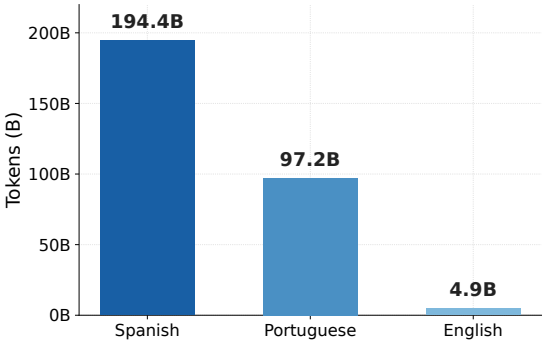
The final corpus contains approximately 187 million documents and 296.5 B tokens, drawn from sources across 20 countries. It includes web content, books, code, news, and domain-specific collections in Spanish, Portuguese, and English. The mixture strengthens the representation of Latin American Spanish and Brazilian Portuguese while retaining sufficient general-domain content and source diversity for broad pre-training.

Figure 2 shows the linguistic and geographic composition of the corpus. Spanish accounts for 194.4 B tokens, Portuguese for 97.2 B, and English for 4.9 B. At the country level, the largest contributors are Argentina with 59.2 tokens, Mexico with 56.4 B, Colombia with 32.5 B, Chile with 31.6 B, and

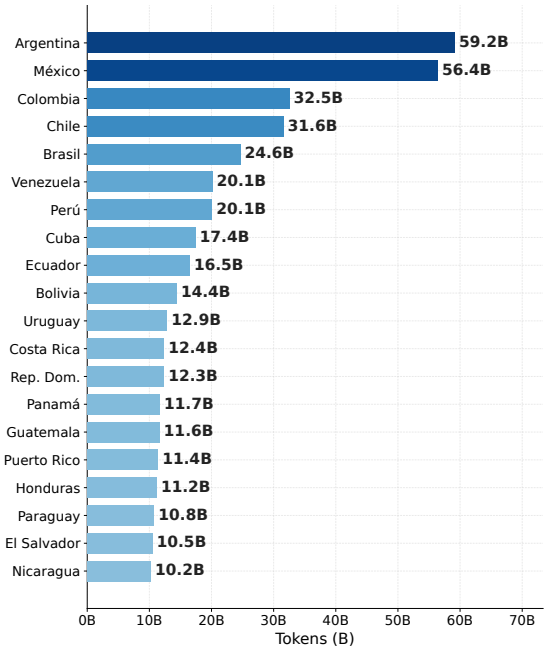
Table 3: Comparison of LatamGPT-Corpus against major multilingual pre-training corpora.

Corpus	Tokens	LatAm Focus	Languages	Country of origen
mC4 [Xue et al., 2021]	400B+	No	100+	No
CulturaX [Nguyen et al., 2024]	6.3T	No	167	No
ROOTS [Laurençon et al., 2022]	1.6T	No	46	No
LatamGPT-Corpus	**296.5B**	**Yes**	**ES, EN, PT**	Yes†

† Country of origin is detected via a classifier.



(a) Distribution by language.



(b) Multilabel token coverage by country.

Figure 2: Overall distribution in the pre-training corpus.

Brazil with 24.6 B. Peru contributes 20.1 B tokens, followed by Venezuela with 20.1 B, Cuba with 17.4 B, and Ecuador with 16.5 B. Each of the 20 countries represented in the corpus contributes at least 10 B tokens. Note that a document may be attributed to more than one country.

This distribution reflects deliberate sampling rather than each country's raw presence on the internet, which would disproportionately favor the region's largest and most digitally represented countries. The collection strategy instead seeks to ensure meaningful representation for both mid-sized and smaller countries. Token totals are calculated after language and country inference, personally identifiable information (PII) redaction, and duplicate filtering.

Figure 3 summarizes the corpus by topic. Political content accounts for 89.2 B tokens, or 30.1% of the effective corpus, while humanities-related content contributes 73.2 B tokens, or 24.7%. Economy and education each account for 33.7 B tokens, or 11.4%. Communication contributes 27.7 B tokens, or 9.4%, while science contributes 27.1 B tokens, or 9.1%. Arts, medicine, sports, and content related to Indigenous peoples account for smaller but still meaningful shares.

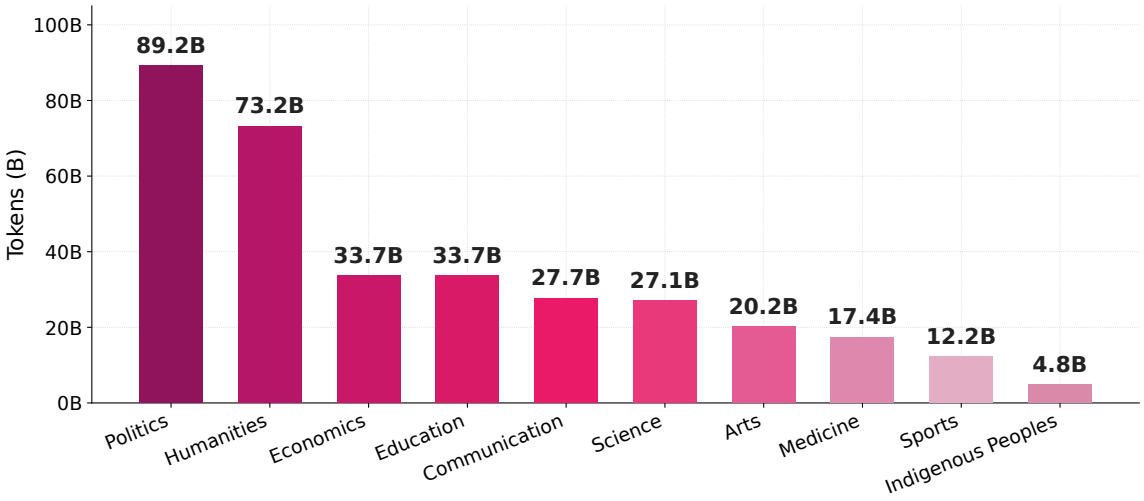


Figure 3: Multilabel token coverage by topic domain.

Because topic classification is multi-label, a document may contribute to more than one category. The percentages therefore measure topic coverage rather than mutually exclusive shares of the corpus. The distribution provides broad coverage of socially and culturally relevant domains while also identifying areas for further rebalancing, particularly medicine, sports, and content related to Indigenous peoples.

The pipeline scores each document along two dimensions: Latin American cultural relevance and syntactic quality. These scores exhibit a moderate negative correlation. Documents with high regional density often contain colloquialisms, non-standard spelling, or conversational structures that receive lower scores under strict syntactic-coherence criteria. The pipeline therefore applies adaptive quality thresholds, relaxing syntactic requirements when regionally grounded data is scarce. This approach preserves culturally informative content that a single global threshold would otherwise remove.

2.1.2 Tokenization and Padding Pipeline

The tokenization pipeline transforms the filtered and deduplicated corpus into fixed-length sequences that serve as input to pre-training. Figure 4 illustrates the two-stage tokenization process, implemented as a distributed process deployed on Kubernetes. A distributed design is necessary because the corpus contains hundreds of billions of tokens spread across millions of documents and thousands of input files. Processing this volume on a single machine would result in prohibitive runtimes and substantial input/output bottlenecks. By partitioning the corpus across many pods, the pipeline parallelizes document streaming, tokenization, and Apache Arrow serialization while keeping memory use bounded and allowing failed shards to be retried independently.

This architecture also enables us to automate the creation of multiple tokenized versions of the pre-training dataset. We can therefore test hypotheses rapidly across alternative corpus compositions, data mixtures, sampling strategies, and memory-replay configurations. The role of these configurations in the final training mixture is discussed in greater detail in Section 2.1.3 and Section 3.

The first stage runs as a 64-pod Indexed Job. Through round-robin partitioning, each pod receives

a disjoint subset of the input Parquet files. Each pod then streams documents from Amazon S3, tokenizes them in memory, and writes the output directly to Apache Arrow. By avoiding intermediate disk writes, this design keeps peak memory use bounded independently of input file size. The second stage also uses 64 pods to aggregate statistics and metadata across the output shards, with pod 0 acting as the coordinator. The resulting Arrow files can then be loaded and accessed as a single dataset.

We use the Llama 3.1 tokenizer [Dubey et al., 2024], a byte-pair encoding model with a vocabulary of 128,256 tokens. All sequences are padded or truncated to a fixed context length of 8,192 tokens. The output schema contains `input_ids`, `attention_mask`, and `labels`, with the labels replicating the input identifiers for causal language-model training. At training time, Apache Arrow provides memory-mapped, zero-copy access, reducing input/output overhead across distributed GPU nodes.

More than 75% of web documents contain fewer than 1,024 tokens. Padding each document independently to the 8,192-token context window would therefore introduce substantial overhead. Without packing, the frequency-weighted effective padding across all length ranges reaches 86.59%. This means that, on average, more than five-sixths of each sequence would consist of padding.

Sequence packing addresses this inefficiency by concatenating multiple documents within each 8,192-token context window and inserting a separator token at every document boundary. As shown in Table 4 and Figure 5, packing reduces the effective padding overhead from 86.59% to 12.74%, a $6.8\times$ reduction. Most of the remaining overhead is concentrated in the 7k–8k token bin, where long documents leave only a small portion of the context window unused.

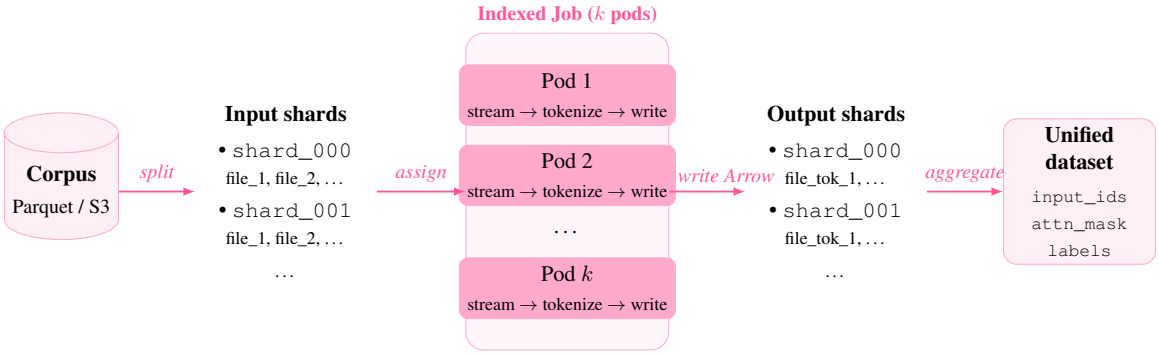


Figure 4: Overview of the distributed tokenization pipeline. The corpus is partitioned into input shards, each assigned to one pod in a k-pod Kubernetes Indexed Job. Each pod streams documents from S3, tokenizes them in memory, and writes the result to Apache Arrow format. A subsequent aggregation step merges per-shard statistics into a unified dataset manifest with fields `input_ids`, `attention_mask`, and `labels`.

2.1.3 Dataset Assembly

We initially constructed and began training with the full 296.5,B-token pre-training corpus. Subsequent experiments showed that a reduced training set achieved comparable downstream performance, so we adopted a reduced mixture of 203.2 B tokens. This reduced mixture was obtained primarily by applying more aggressive filtering to publicly sourced documents, particularly content likely to

Table 4: Sequence Packing Impact on Effective Padding Overhead

Length Range	Mean Pad %	Frequency % (No Packing)	Effective Pad % (No Packing)	Frequency % (With Packing)	Effective Pad % (With Packing)
0–1023 tokens	93.76	75.57	70.86	0.72	0.67
1024–2047 tokens	81.26	14.78	12.01	1.35	1.10
2048–3071 tokens	68.76	3.46	2.38	1.72	1.18
3072–4095 tokens	56.26	1.37	0.77	2.22	1.25
4096–5119 tokens	43.76	0.77	0.34	3.09	1.35
5120–6143 tokens	31.26	0.50	0.16	5.04	1.58
6144–7167 tokens	18.76	0.37	0.07	12.08	2.26
7168–8191 tokens	6.26	0.30	0.02	53.57	3.35
8192 tokens	0.00	2.88	0.00	20.23	0.00
Total Padding Overhead			**86.59%**		**12.74%**

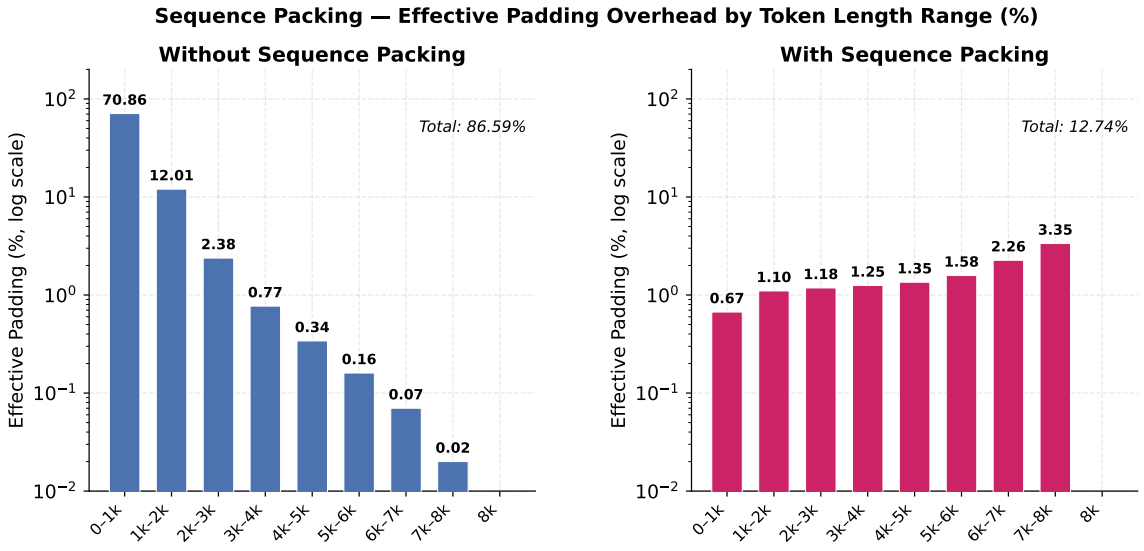


Figure 5: Effective padding overhead by token-length range, without (left) and with (right) sequence packing.

overlap with the original pre-training distribution of the base model. By prioritizing non-web data, this scheme preserves higher-quality alliance-contributed data that is more closely aligned with the regional objectives of the model. The resulting dataset provides a more compute-efficient training mixture without a measurable loss in downstream performance.

The configuration described below corresponds to the 203.2 B-token version used to produce the final training budget reported in Table 5. The mixture consists of four components that were assembled through a pipeline of four PySpark jobs deployed on AWS Glue, with one job for each mixture component.

i. **Latam (Quality-Filtered).** The Latam component is drawn from the filtered pre-training corpus, with alliance-contributed and publicly sourced documents processed under different selection criteria. Alliance-contributed documents are retained because their Latin American

provenance and institutional origin have already been established. Publicly sourced documents are evaluated using the same syntactic-quality filters applied during Phase 1 of corpus construction, but with stricter threshold values. All alliance-contributed documents are then combined with a random sample of the qualifying publicly sourced documents.

ii. **Upsampling.** The Upsampling component targets high-value institutional sources that remain underrepresented in the broader corpus. Among others, these sources include academic repositories, legal and parliamentary archives, and regional news outlets. Sampling fractions are tuned independently for each source, with most institutional collections included in full.

iii. **Replay.** The Replay component comes from a dedicated memory-replay corpus: a sample drawn from the original pre-training distribution of the base model on which we perform continual pre-training (CPT). Reintroducing this data during training mitigates catastrophic forgetting of the model's existing capabilities. Multilingual prose and Wikipedia content pass through syntactic-quality filters, while code and mathematics corpora are included without additional filtering. Documents are stratified by content type, with higher sampling fractions assigned to science, technology, engineering, and mathematics data, the domains most vulnerable to degradation during adaptation.

iv. **Synthetic.** The Synthetic component is drawn from the synthetic pipeline described in Section 2. It includes region-specific slang, general synthetic documents, and question-answer pairs. No additional filtering is applied during dataset assembly.

Phase assignment and sequence preparation. The four components described above are combined, and each document receives a `training_phase` label. Latam and Replay documents are stochastically divided between Phase 0 and Phase 1, while Synthetic and Upsampling documents are assigned entirely to Phase 1.

The two phases serve complementary roles. Phase 0 provides broad distributional coverage through Latam and Replay data. Phase 1 is an enriched continuation mixture that concentrates high-value institutional sources, synthetic regional content, selected Latam data, and a smaller Replay component. This design places the most informative and underrepresented examples near the end of training, when smaller parameter updates can refine regional knowledge while limiting general-domain forgetting.

Before tokenization, the unified dataset passes through three steps:

1. **Split:** documents exceeding the context window are divided into fixed-length chunks.
2. **Pack:** chunks are bin-packed into full-length sequences separated by `<|end_of_text|>` tokens. Phase 0 and Phase 1 content are never mixed within a single sequence.
3. **Train/validation split:** packed sequences are randomly partitioned into training and validation sets.

Table 5 summarizes the final training budget. Phase 0 contributes 128.0 B tokens, or 62.98% of the 203.2 B-token total. It consists of 104.0 B Latam tokens and 24.0 B Replay tokens.

Phase 1 contributes 75.2 B tokens, or 37.02% of the total. Upsampling accounts for 37.3 B tokens, nearly half of the phase, followed by 26.0 B Latam tokens, 6.0 B Synthetic tokens, and 6.0 B Replay tokens. This composition makes Phase 1 more selective than Phase 0 while retaining enough Replay data to preserve general capabilities. The rationale for the phased schedule and Replay fractions is discussed in Section 3.

Table 5: Token Counts and Budget Distribution Across Training Phases

Phase	Component	Volume	Intra-Phase %	Global % (203.2 B)
Phase 0	Phase 0 Total	128.0 B	—	62.98%
	Latam	104.0 B	81.26%	51.18%
	Replay	24.0 B	18.74%	11.80%
	STEM	16.8 B	13.14%	8.27%
	Gen. Knowledge	3.6 B	2.82%	1.77%
	Multilingual	3.6 B	2.78%	1.75%
Phase 1	Phase 1 Total	75.2 B	—	37.02%
	Upsampling	37.3 B	49.58%	18.35%
	Latam	26.0 B	34.49%	12.77%
	Synthetic	6.0 B	7.97%	2.95%
	Replay	6.0 B	7.95%	2.95%
	STEM	4.2 B	5.57%	2.06%
	Gen. Knowledge	0.9 B	1.19%	0.44%
	Multilingual	0.9 B	1.19%	0.44%
Global	**Total Training**	**203.2 B**	**—**	**100.00%**

2.2 Post-training Data Mixture

The post-training dataset contains 70 B tokens, assembled from selected data slices drawn from eight sources suitable for supervised fine-tuning (SFT) of LLMs. We describe each source below and summarize the resulting composition at the end of the subsection.

i. **Open datasets from prior work.** We use slices of two reference SFT corpora from AI2: Tülu 3 [Lambert et al., 2024] and OLMO 3 [Team Olmo et al., 2025]. Both release data, code, and weights, giving us a stable baseline against which to evaluate each candidate addition. The slices we use are `latam-gpt/tulu-3-sft-mixture-no-identity` and `latam-gpt/Dolci-Instruct-SFT-No-Tools-No-Hardcoded-Sample-200k`.

ii. **Translated open datasets.** We translated several English SFT corpora into Spanish to test whether off-the-shelf instruction data could be transferred without further curation: `es-smoltalk`, `es-ultrachat`, and a translated Tülu 3 / OLMo 2 mixture. Pilot runs on these slices did not improve performance on Trueque or Choclo over the untranslated Tülu 3 baseline, and we excluded them from the final SFT mixture. Safety data was the only exception, as translation preserved sufficient utility for inclusion. We therefore retained `gretel-safety-alignment-es-v1`.

iii. **Copuchat.** Copuchat is a public chat platform that we operate, providing users with free access to a strong proprietary chat model in exchange for the optional donation of their conversations. Donated conversations are anonymized before release, following the WildChat data-collection model [Zhao et al., 2024]. Because the platform remains active, successive releases incorporate newly collected data and revised selection criteria. The snapshots evaluated are `copuchatV2`, `copuchat-sft`, and `copuchat-conversations`. These conversations capture regional usage patterns that are absent from existing public SFT corpora.

iv. **Document-to-QA.** Open SFT mixtures contain limited Latin American cultural content, while Copuchat conversations, although regionally grounded, are not concentrated on cultural topics. To

address this gap, we generated synthetic question-answer pairs from Spanish Wikipedia using a four-steps pipeline.

First, we traverse the category graph of the Spanish Wikipedia dump (`eswiki-20260101`, snapshot of January 1st, 2026) for 19 Latin American countries and nine cultural domains: culture, cuisine, music, folklore, history, literature, art, festivals, and mythology. We retain articles belonging to any descendant category and remove wikitext markup, including templates, references, and inline formatting. This stage yields 524,336 articles.

Second, for each article, we generate between one and five QA pairs using Qwen3-30B-A3B [Yang et al., 2025] with JSON-guided decoding. The number of pairs is scaled according to article length, producing 1,353,539 pairs in total.

Third, a second pass using the same model classifies each pair according to its Ibero-American relevance and refines its country assignment. Spain, Portugal, and the meta-labels *Latinoamérica* and *Iberoamérica* are also permitted at this stage. Pairs that satisfy the category-based criteria but are classified as generic are discarded. After this filtering step, 1,062,687 pairs (78.5%) remain.

Fourth, we sample the remaining pairs to improve geographic and thematic balance. Each country contributes at most 5,000 pairs, and each domain is subsequently increased to at least 6,795 pairs, except for mythology, whose source pool is smaller. This process yields 104,625 pairs. Brief answers in the sampled set are then expanded by re-prompting Qwen3-32B with the source article as context, but only when the article contains additional relevant information.

The previous 4-steps pipeline produces two datasets: `wikipedia-qa-iberoamerica`, containing the 104,625 balanced and expanded conversations, and `wikipedia-qa-iberoamerica-1M`, containing the 1,062,687 pairs retained after relevance filtering, without balanced sampling or answer expansion. Figure 6 summarizes the pipeline.

v. **Long-context (LongMagpie).** Magpie [Xu et al., 2025] exploits the autoregressive structure of aligned chat models. Chat templates delimit each turn with role-marker tokens (`<|system|>`, `<|user|>`, and `<|assistant|>`). Magpie provides the model only with the template prefix up to and including the `<|user|>` marker, without any user content, prompting the model to generate a plausible user query. The same model then answers the generated query, producing an instruction-response pair without seed prompts or manual prompt engineering. LongMagpie [Gao et al., 2025] extends this approach to long-context data by prepending a document to the partial template, thereby grounding the generated query in the document.

The standard recipe did not work reliably with long documents. Instead of switching to a user-query format after the `<|user|>` marker, the model often continued the document with declarative sentences and rarely generated a question. We observed the same behavior with both Llama-3.1-70B-Instruct [Dubey et al., 2024] and Qwen2.5-72B-Instruct [Yang et al., 2024] as source models, suggesting that the failure mode is not specific to Spanish or to a single model.

To induce question generation, we prefixed the user-turn slot in two ways and retained both variants to increase diversity in the resulting question distribution: a meta-phrase and a single one. Both prefixes reliably elicited a question. However, the model frequently generated the answer within the same turn, so we retained only the text spanning from the first to the corresponding question mark before treating the generated text as the user query. The remainder of the LongMagpie pipeline followed the published procedure.

We applied this recipe to documents from `red-pajama-hq-es`, producing `red-pajama-hq-qa-es` and `red-pajama-hq-qa-es-chat`.

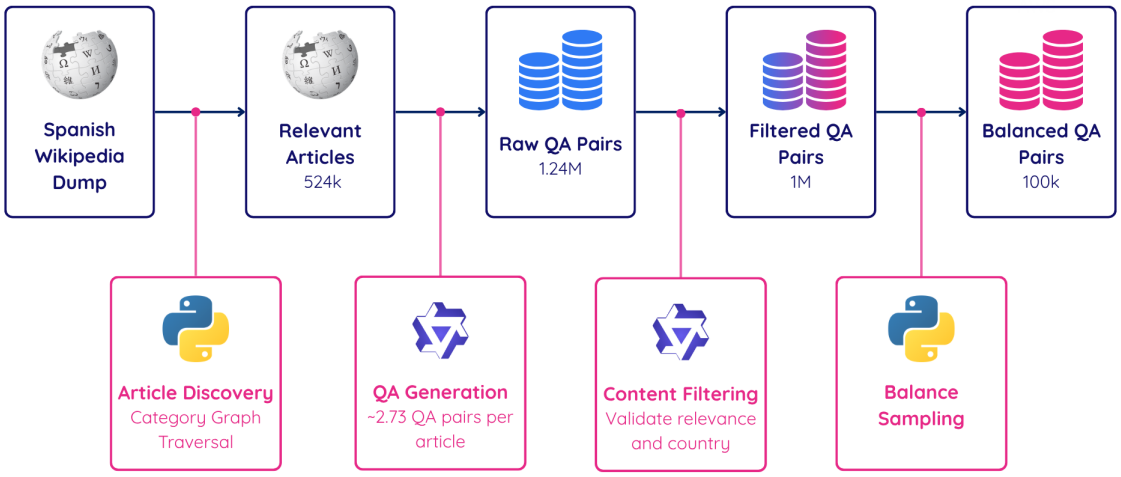


Figure 6: Document-to-QA pipeline. The Spanish Wikipedia dump is filtered to retain relevant articles through category-graph traversal, converted into QA pairs using Qwen3-30B-A3B, refined through a relevance-classification pass with the same model, and balanced using per-country and per-domain caps. Counts are rounded; exact values are provided in the main text.

vi. **Persona-conditioned generation.** Real user queries reflect users' expertise, interests, and life circumstances, distributions that translated or imported SFT data may not adequately capture. Following Ge et al. [2024], we diversify instruction generation by first constructing a roster of synthetic personas and then conditioning each generated instruction on one of them. Conditioning generation on a persona steers the model toward the types of questions and responses that would be plausible for that user profile. A diverse set of personas can therefore broaden the range of user needs and perspectives represented in the generated instructions without requiring these profiles to be specified manually.

The pipeline consists of three stages. *(i) Text-to-Persona.* For each document in a sample from `red-pajama-hq-es`, we prompt Qwen3-30B-A3B-Instruct [Yang et al., 2025] to generate a single, specific persona description. The prompt excludes proper names, ages, and hedging language. Each document is reused with several stance variants, allowing the same source text to elicit multiple distinct personas.

(ii) Deduplication. The resulting roster is filtered using MinHash, followed by a multilingual embedding-similarity pass. This configuration follows Ge et al. [2024], except that we replace the English-language embedder used in the original work with a multilingual sentence encoder. The filtered roster is released as `latam-gpt/personas`.

(iii) Persona-to-Instruction. Each persona is provided to Qwen3-30B-A3B-Instruct with a prompt asking what request that persona might make, subject to public-knowledge constraints and variation in surface form. Each resulting instruction is then answered independently by Qwen3-30B-A3B-Instruct and Llama-3.3-70B-Instruct [Dubey et al., 2024], yielding two candidate responses per instruction for potential preference-based downstream use. The resulting dataset is released as

`latam-gpt/personas-instruct`.

In pilot SFT mixtures, `latam-gpt/personas-instruct` did not improve over the `wikipedia-qa-iberoamerica` baseline and was excluded from the final 70 B mixture. One possible reason is that we did not implement the Persona-to-Persona expansion stage proposed by Ge et al. [2024], which expands the roster by traversing interpersonal relationships and is intended to recover personas that are underrepresented in web text. Because our pipeline relies only on Text-to-Persona generation, the resulting roster may retain systematic coverage gaps.

vii. **Safety.** The safety mixture follows the safety-oriented data recipe used in OLMo 3 [Team Olmo et al., 2025]. It draws on the training partitions of three datasets: CoCoNot, which targets contextual noncompliance and nuanced refusal behavior for requests that should not be directly fulfilled [Brahman et al., 2024]; WildGuardMix, a moderation-oriented mixture containing harmful and benign prompts, adversarial examples, refusal labels, and fine-grained safety categories [Han et al., 2024]; and WildJailbreak, a large-scale safety dataset containing both standard and adversarial jailbreak-style prompts, together with harmful and benign contrastive examples designed to improve robustness while reducing over-refusal [Jiang et al., 2024]. For all three datasets, we use only the released training partitions and exclude all validation, evaluation, and test data.

We downloaded the complete training partitions of the three resources and constructed a 50,000-example mixture for preliminary testing using category-aware sampling. Whenever possible, examples were sampled to approximate a uniform distribution across the categories defined within each source dataset, while maintaining broad coverage of benign, harmful, adversarial, and refusal-oriented examples in the overall mixture. Table 6 reports the resulting composition.

Since these resources were originally released in English, we translated the selected examples into Spanish as part of the safety-data construction process. Translation was performed using `dolphin-2.9.1-llama-3-70b`, an instruction-tuned Llama-3-70B derivative described by its authors as uncensored and highly compliant [Hartford and Cognitive Computations, 2024]. This choice was methodological, because the datasets contain harmful, adversarial, and ethically sensitive material. Therefore, using a highly compliant model reduces the likelihood of refusals, omissions, or safety-driven reformulations during translation, thereby helping preserve the semantic content and intended category of the original examples. To assess translation quality, we manually reviewed 100 randomly sampled translations for semantic fidelity to the original English text, preservation of the intended safety category, and formatting consistency.

viii. **Identity.** LatamGPT is initialized from Llama 3.1 70B and therefore inherits its self-identification behavior. To align the model with its own identity, we include conversations in which users ask about the model's name, origin, or provenance and the assistant responds as LatamGPT. We evaluate three candidate datasets: `synthetic-identity`, `self-identity-v2`, and `self-identity-v3`. The final mixture retains `synthetic-identity`.

ix. **MCQ Benchmark Training Data.** Multiple-choice benchmarks require models to select among a fixed set of answer options, a format that can expose systematic selection biases. Models trained predominantly on open-ended instruction data may favor particular answer positions or superficial response patterns rather than reasoning consistently over the available options. To improve the model's ability to handle this format and produce better-calibrated answer distributions, we assembled `sft_mcqa_unified`, a unified collection of SFT-formatted examples drawn exclusively from the *official training splits* of ten established benchmarks. Each example is converted into an instruction-following format consisting of a system prompt, a question, and a set of answer choices, with no

Table 6: Composition of the safety data mixture.

Dataset	Subcategory	Count	Reference	Dataset total
CoCoNot	copyright violations	82	[Brahman et al., 2024]	531
	dangerous or sensitive topics	75		
	misinformation	120		
	privacy violations	101		
	triggers for offensive language	60		
	wildchats	93		
WildGuardMix	benign	2,828	[Han et al., 2024]	5,028
	causing material harm by disseminating misinformation	86		
	copyright violations	144		
	cyberattack	105		
	defamation encouraging unethical or unsafe actions	135		
	disseminating false or misleading information encouraging disinformation campaigns	219		
	fraud assisting illegal activities	133		
	mental health over-reliance crisis	121		
	others	17		
	private information individual	179		
	sensitive information organization government	210		
	sexual content	126		
	social stereotypes and unfair discrimination	374		
	toxic language hate speech	180		
	violence and physical harm	171		
WildJailbreak	adversarial benign	13,377	[Jiang et al., 2024]	44,441
	adversarial harmful	14,056		
	vanilla benign	8,504		
	vanilla harmful	8,504		
			Total	**50,000**

exposure to the evaluation partitions used for benchmarking. Table 7 summarizes the resulting composition, comprising 370,684 examples in total.

Table 7: Composition of `sft_mcqa_unified`. Each entry is an SFT-formatted version of the official *training* split; no evaluation partitions are included.

Dataset	Reference	Examples
`medmcqa`	[Pal et al., 2022]	120,765
`mmlu_auxiliary_train`	[Hendrycks et al., 2020]	89,479
`synthetic_latam_mcq`	(this work)	50,390
`winogrande`	[Sakaguchi et al., 2021]	40,398
`hellaswag`	[Zellers et al., 2019]	39,905
`sciq`	[Welbl et al., 2017]	11,679
`commonsenseqa`	[Talmor et al., 2019]	9,741
`openbookqa`	[Mihaylov et al., 2018]	4,957
`arc_easy`	[Clark et al., 2018]	2,251
`arc_challenge`	[Clark et al., 2018]	1,119
Total		**370,684**

`synthetic_latam_mcq` (50,390 examples) is a dataset created for this work by generating multiple-choice questions synthetically from documents in the pre-training corpus using an LLM,

targeting Latin American cultural and factual knowledge not covered by existing benchmarks.

Final mixture. The 70 B-token SFT mixture follows the **Base-2** configuration selected from the 8 B ablations described in Section 4.1. It comprises `tulu-3-sft-mixture-no-identity`, `wikipedia-qa-iberoamerica-1M`, `copuchat` (February 2025 snapshot), `synthetic-identity`, `sft_mcqa_unified` (MCQ Unified; see Table 7), and the safety slice `safety-sft-50k-sample`. The translated safety dataset `gretel-safety-alignment-es-v1` is reserved for the dedicated safety fine-tuning stage described in Section 4.3.

3 Pre-training

3.1 Model Architecture

LatamGPT 70 B is initialized from Llama 3.1 70 B [Dubey et al., 2024], a decoder-only Transformer trained using over 15 trillion tokens of multilingual text. Initializing LatamGPT from a strong open base model is motivated by both computational efficiency and transfer learning. The Chinchilla scaling law [Hoffmann et al., 2022] prescribes approximately $6N$ training tokens for an N-parameter model at optimal compute allocation, placing the corresponding token budget for a 70 B model near 420 B tokens, a regime already exceeded by the original pre-training of Llama 3.1. Domain-specific CPT therefore functions as a targeted adaptation, steering the base model towards Latin American knowledge rather than building from scratch.

The base architecture is a standard grouped-query attention (GQA) Transformer [Ainslie et al., 2023] with rotary positional embeddings (RoPE, Su et al. 2021) and pre-normalization through RMSNorm. We do not modify the architecture; all adaptation occurs through parameter updates during CPT. Table 8 reports the main architectural hyperparameters.

Table 8: LatamGPT 70 B architecture.

Hyperparameter	**Value**
Transformer layers	80
Hidden dimension	8,192
FFN intermediate dim (SwiGLU)	28,672
Attention heads	64
KV heads (GQA)	8
Head dimension	128
Vocabulary size	128,256
Context length (CPT)	8,192 tokens
Positional encoding	RoPE ($\theta = 500{,}000$)
Normalization	RMSNorm (pre-norm)
Activation	SwiGLU

3.2 Training Infrastructure and Parallelism Strategy

All CPT experiments run on an AWS SageMaker HyperPod cluster comprising 16 `p5en.48xlarge` nodes, each hosting 8 NVIDIA H200 SXM5 GPUs, for a total of 128 H200 GPUs. Training uses NeMo 2.0 with Megatron-Core as the parallel-computation backend.

We use tensor parallelism with TP=8, Fully Sharded Data Parallelism with FSDP=16, and no pipeline parallelism (PP=1), yielding an effective data-parallel degree of 16 ranks.

The choice of TP=8 follows the intra-node interconnect topology. Each `p5en.48xlarge` node provides 8 H200 GPUs connected through NVLink and NVSwitch at approximately 900 GB/s of bandwidth per GPU. Tensor parallelism requires an all-reduce after every attention and feed-forward sublayer in both the forward and backward passes. These operations are frequent and latency-sensitive. Confining tensor parallelism to 8 GPUs within a single NVLink domain ensures that these all-reduces do not cross node boundaries.

Extending tensor parallelism beyond 8 GPUs would required cross-node tensor-parallel communications through the AWS Elastic Fabric Adapter (EFA). Its effective point-to-point bandwidth, approximately $\sim$200 GB/s, is approximately $4.5\times$ lower than the intra-node NVLink bandwidth and introduces substantial tensor-parallel overhead relative to computation.

FSDP uses the `hybrid_shard` strategy. Model parameters are fully sharded within each intra-node NVLink domain, while synchronization across data-parallel ranks uses EFA for inter-node communication. Because this communication occurs at a lower frequency than the collectives required by tensor parallelism, EFA is better suited to this role.

Activation checkpointing is enabled to reduce peak memory use, while activation offloading is disabled to avoid CPU-memory bandwidth bottlenecks. Storage is provided by FSx for Lustre, with NCCL configured to use EFA for collective communication. We set the NCCL timeout to 14,400-seconds to accommodate pauses during checkpoint saving.

3.3 Training Recipe

Precision. We train using FP8 mixed precision through the SageMaker Model Parallel library. Matrix multiplications in the forward and backward passes are performed in FP8, while optimizer states, gradient accumulation, and layer-normalization operations remain in BF16 to preserve numerical stability. Because FP8 has a limited dynamic range, each tensor requires a scaling factor that maps its BF16 values into the representable FP8 range before computation and rescales the resulting values afterward. These scaling factors are derived from the running maximum absolute value (*amax*) of each tensor.

We set `fp8_amax_history_len` $= 1{,}024$ and `fp8_amax_compute_algo` $=$ `max`, which selects the largest observed *amax* value within the recorded history. Maintaining a history of 1,024 values, rather than estimating the scale from only the current step, produces a more stable estimate of tensor magnitude. This reduces oscillations in the scaling factors and helps prevent gradient spikes when the data distribution changes.

Batch size and throughput. We evaluate per-GPU micro-batch sizes of $\{4, 8, 10, 18, 24\}$ at a fixed sequence length of 8,192 tokens using sequence packing. This coarse search identifies the configuration that maximizes sustained training throughput.

The global batch size is the product of the per-GPU micro-batch size, the number of data-parallel ranks, and the number of gradient accumulation steps. With tensor parallelism of 8, each data-parallel rank spans eight GPUs and processes the same micro-batch on each device. The 128-GPU configuration therefore contains 16 data-parallel ranks.

With four gradient accumulation steps, a micro-batch size of 18 yields

$$18 \times 16 \times 4 = 1{,}152$$

sequences per optimizer step, equivalent to approximately 9.4 million tokens. This configuration sustains approximately 141.6 thousand tokens per second across all 128 GPUs.

We report sustained throughput in tokens per second per GPU, as shown in Figure 7. This metric measures the effective rate at which each device processes training data and allows throughput to be compared across micro-batch configurations.

We select a micro-batch size of 18 based on two observations. First, larger micro-batches use GPU memory more effectively, increase arithmetic intensity on the H200 Tensor Cores, and shift the primary bottleneck from memory access toward computation. As the micro-batch size increases from 4 to 18, per-GPU throughput rises from 819 to 1,106 tokens per second.

Second, larger micro-batches improve the overlap between computation and FSDP communication. With an FSDP degree of 16, parameters and gradients are transferred through fixed-size flattened buckets using all-gather and reduce-scatter operations. A larger micro-batch lengthens the computation window of each micro-step, allowing a greater fraction of this communication to be hidden behind matrix multiplications and reducing exposed communication time.

The trend reverses beyond a micro-batch size of 18. A micro-batch size of 24 exceeds the available memory budget for the 70B model at a sequence length of 8,192 tokens, even with activation checkpointing enabled. Supporting this configuration requires additional micro-batch splitting, which shortens the computation window, exposes more communication, and reduces throughput to 983 tokens per second per GPU.

A micro-batch size of 18 therefore achieves the highest sustained throughput on `p5en.48xlarge` nodes, reaching 1,106 tokens per second per GPU.

Learning-rate schedule. We use cosine annealing with no linear warmup, following the approach of several recent CPT studies [Ibrahim et al., 2024, Dubey et al., 2024]. Phase 0 begins with $\mathrm{lr}_{\max}^{(0)} = 10^{-6}$ and follows a global cosine schedule toward $\mathrm{lr}_{\min}^{(0)} = 10^{-7}$ until the Phase 0 boundary.

The global cosine schedule is parameterized as

$$\mathrm{lr}_t = \mathrm{lr}_{\min} + \frac{1}{2}\left(\mathrm{lr}_{\max} - \mathrm{lr}_{\min}\right)\left[1 + \cos\left(\pi \frac{t}{T_{\mathrm{decay}}}\right)\right], \tag{1}$$

where T_{decay} denotes the total scheduled training budget across both phases and t is the cumulative training position measured in optimizer steps or processed tokens.

The initial learning rate for Phase 1 is derived by evaluating the Phase 0 cosine schedule at the phase boundary. Because Phase 0 accounts for 62.98% of the total scheduled token budget, this boundary occurs at $t = 0.6298\, T_{\mathrm{decay}}$:

$$\mathrm{lr}_{\mathrm{init}}^{(1)} = \mathrm{lr}_{\min}^{(0)} + \frac{1}{2}\left(\mathrm{lr}_{\max}^{(0)} - \mathrm{lr}_{\min}^{(0)}\right)\left[1 + \cos\left(\pi \cdot 0.6298\right)\right]. \tag{2}$$

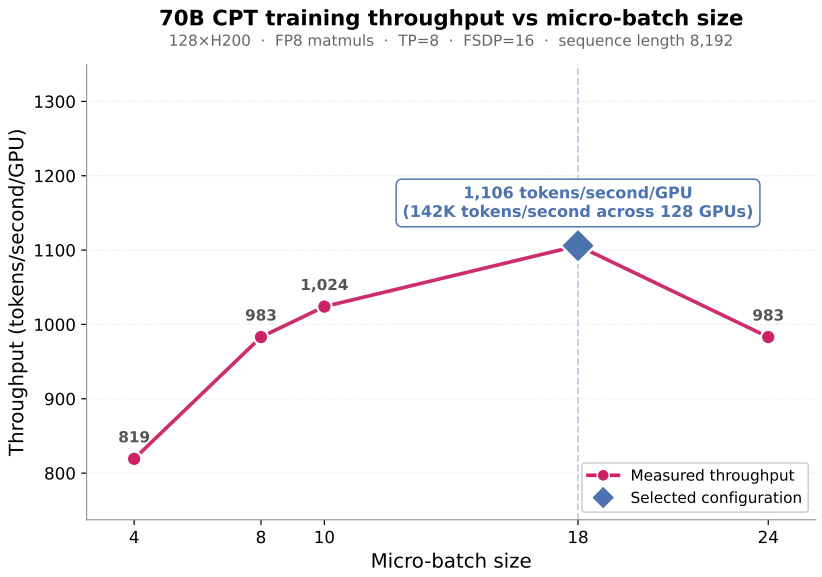


Figure 7: Training throughput in tokens per second versus per-GPU micro-batch size for 70B CPT on 128 H200 GPUs, using TP=8, FSDP=16, and a sequence length of 8,192 tokens.

Phase 1 starts from $1.2 \times \mathrm{lr}_{\mathrm{init}}^{(1)}$, corresponding to a 20% increase over the learning rate at the Phase 0 boundary, and then decays to $\mathrm{lr}_{\mathrm{min}}^{(1)} = 0$ over the remaining token budget.

This controlled increase provides additional plasticity for adapting to the more selective Phase 1 mixture without restarting from the original peak learning rate. The resulting trajectory introduces a small, deliberate increase at the phase boundary rather than a fully continuous schedule. This avoids the much larger discontinuity that would result from restarting Phase 1 at the original peak learning rate after the model has already adapted to the Phase 0 distribution. Table 9 summarizes the main training hyperparameters.

3.4 Multi-Scale Continued Pre-training Recipe Selection

Training is organized into two phases. **Phase 0** uses the broad pre-training mixture, providing general-domain coverage across 20 Latin American countries. **Phase 1** continues from the Phase 0 checkpoint using a more selective mixture that emphasizes high-quality institutional sources, alliance-contributed data, synthetic regional content, and a smaller memory-replay component. The learning rate is adjusted at the phase boundary to strengthen Latin America-specific adaptation while limiting the loss of capabilities inherited from the base model.

Selecting the final 70 B training recipe requires balancing adaptation to Latin American data against forgetting of general-domain knowledge. Because an extensive search at 70 B would be prohibitively expensive, we conduct most ablations at the 1 B and 8 B parameter scales and reserve a smaller set of targeted runs for validation at 70 B.

The following subsections first formalize the adaptation-forgetting criteria used to compare candidate recipes, then examine how the useful learning-rate range changes with model size, and finally evaluate whether perplexity-based filtering can reduce the training data without discarding useful regional signal. Together, these analyses determine the continued pre-training configuration applied

Table 9: CPT training hyperparameters. Phase 0 and Phase 1 share the same infrastructure configuration, while the learning-rate trajectory differs between phases.

Hyperparameter	**Phase 0**	**Phase 1**
Framework	NeMo 2.0 + Megatron-Core	
Precision	FP8 / BF16 mixed	
FP8 amax history length	1,024	
FP8 amax algorithm	`max`	
Nodes $\times$ GPUs	16×8 H200	
Tensor parallelism (TP)	8	
FSDP degree	16	
Sharding strategy	`hybrid_shard`	
Sequence length	8,192 tokens	
Sequence packing	Yes	
Per-GPU micro-batch size	18	
Gradient accumulation	4	
Global batch size	1,152 sequences	
Activation checkpointing	Yes	
Activation offload	No	
Optimizer	AdamW	
Gradient clipping	1.0	
LR schedule	Cosine annealing	

to LatamGPT 70 B.

3.4.1 Adaptation-Forgetting Framework

Each continued pre-training recipe defines a point in a two-dimensional space. Adaptation is measured through improvement on regional benchmarks, while forgetting is measured through degradation on general-domain benchmarks. We formalize this trade-off and the selection procedure used to identify the 70B candidate from the multi-scale ablations.

Metrics. We evaluate each candidate on two general-domain benchmarks and two regional benchmarks. Massive Multitask Language Understanding (MMLU) measures broad academic and factual knowledge across multiple disciplines, while HellaSwag evaluates commonsense reasoning and sentence-completion ability. We use both to quantify retention of the capabilities inherited from the base model.

Choclo measures Latin America-specific knowledge and language understanding, while Trueque HardTruth evaluates factual knowledge grounded in Latin American countries and institutions. We use these regional benchmarks to measure adaptation to the target data distribution. Section 5.1 describes the benchmarks, evaluation protocols, and results in greater detail.

Let $M(c)$, $H(c)$, $C(c)$, and $T(c)$ denote the MMLU, HellaSwag, Choclo, and Trueque HardTruth scores of candidate c. Let M_{base}, H_{base}, C_{base}, and T_{base} denote the scores of the corresponding Llama base model at the same scale.

Forgetting and adaptation are expressed as signed percentage changes relative to the base model:

$$\mathrm{MMLU}_{\mathrm{forget}}(c) = \frac{M_{\mathrm{base}} - M(c)}{M_{\mathrm{base}}} \times 100\%. \tag{3}$$

$$\mathrm{HellaSwag}_{\mathrm{forget}}(c) = \frac{H_{\mathrm{base}} - H(c)}{H_{\mathrm{base}}} \times 100\%. \tag{4}$$

$$\mathrm{Choclo}_{\mathrm{adapt}}(c) = \frac{C(c) - C_{\mathrm{base}}}{C_{\mathrm{base}}} \times 100\%. \tag{5}$$

$$\mathrm{Trueque}_{\mathrm{adapt}}(c) = \frac{T(c) - T_{\mathrm{base}}}{T_{\mathrm{base}}} \times 100\%. \tag{6}$$

Positive adaptation values indicate improvement over the corresponding base model, whereas positive forgetting values indicate degradation.

Adaptable set. Let $\mathbb{C}_s$ denote the set of candidate models at scale s. The adaptable set $\mathbb{F}_s$ contains candidates that do not regress on Choclo and remain within five percentage points of the MMLU and HellaSwag base scores:

$$\mathbb{F}_s = \left\{ c \in \mathbb{C}_s \;\middle|\; \mathrm{Choclo}_{\mathrm{adapt}}(c) \geq 0,\ \mathrm{MMLU}_{\mathrm{forget}}(c) \leq 5,\ \mathrm{HellaSwag}_{\mathrm{forget}}(c) \leq 5 \right\}. \tag{7}$$

This constraint defines the adaptation region in which regional gains are achieved without excessive degradation of general-domain capabilities. Candidates outside the adaptation region may fail because of regional regression, excessive general-domain forgetting, or both. Importantly, being outside the adaptation region does not automatically imply catastrophic forgetting.

Learning-rate limit. The learning-rate limit lr_s^* at scale s is the highest learning rate for which at least one candidate remains in the adaptable set:

$$\mathrm{lr}_s^* = \max \left\{ \mathrm{lr}(c) \mid c \in \mathbb{F}_s \right\}. \tag{8}$$

Learning rates above this limit fall outside the adaptation region. Rates moderately above the limit may produce regional regression or excessive forgetting, whereas substantially higher rates may enter the catastrophic-forgetting regime.

Selected candidate. At scale s, the selected candidate c_s^* maximizes the macro-average Trueque HardTruth score across four evaluation configurations: prompted and unprompted evaluation, each with and without few-shot examples:

$$c_s^* = \arg\max_{c \in \mathbb{F}_s} \bar{T}(c), \qquad \bar{T}(c) = \frac{1}{4} \left[T_{\mathrm{nsr}}(c) + T_{\mathrm{wsr}}(c) + T_{\mathrm{nsp}}(c) + T_{\mathrm{wsp}}(c) \right]. \tag{9}$$

The subscripts denote no-few-shot unprompted (`nsr`), with-few-shot unprompted (`wsr`), no-few-shot prompted (`nsp`), and with-few-shot prompted (`wsp`). Averaging across these settings reduces sensitivity to any single prompting or few-shot configuration.

Learning-rate sensitivity. Peak learning rate is the primary variable governing the adaptation-forgetting trade-off. Learning rates that are too high may accelerate adaptation but overwrite capabilities inherited from the base model. Rates that are too low preserve those capabilities but produce limited regional adaptation.

The appropriate operating range also changes with model size, making direct transfer of a learning rate from a smaller model unreliable. The following subsection characterizes this scale dependence using experiments at 1 B and 8 B, followed by targeted validation at 70 B.

3.4.2 Learning-Rate Scaling Across Model Sizes

Before committing the full 70 B compute budget, we characterize the adaptation-forgetting frontier at the 1 B and 8 B scales using Llama 3.2 1 B and Llama 3.1 8 B, respectively, and then validate four targeted configurations using Llama 3.1 70 B.

We use Llama 3.2 1 B as the small-scale exploration anchor because it belongs to the same model family and was derived from larger Llama models through pruning and knowledge distillation. This provides greater architectural and behavioral continuity with Llama 3.1 8 B and 70 B than an independently trained 1 B model would provide, although the three scales are not directly equivalent.

Each scale serves a complementary role. The 1 B experiments map the qualitative shape of the adaptation-forgetting frontier at low compute cost and eliminate clearly catastrophic or under-adapted learning rates. The 8 B experiments refine these boundaries, characterize the adaptation region more precisely, and test consistency across ten independently trained candidates.

Finally, the four 70 B candidates test whether the smaller-scale trends transfer and identify the strongest configuration within the inferred 70 B adaptation region. We therefore use the 1 B and 8 B results to narrow the search space, not to predict the 70 B optimum directly.

Adaptation region at each scale. At each scale, we define two operational boundaries. The **ceiling** is the highest measured learning rate that remains inside the adaptation region. It is the largest tested value for which MMLU and HellaSwag forgetting remain at or below 5% and Choclo does not regress.

The **floor** is the lowest measured learning rate that still produces a Trueque improvement over the corresponding base model.

The floors and ceilings at 1 B and 8 B are measured directly. At 70 B, the ceiling is measured at $\mathrm{lr} = 10^{-6}$. The floor is not measured directly. Applying the observed $1/N$ scaling rule from the 8 B floor of 3×10^{-7} gives

$$3 \times 10^{-7} \cdot \frac{8.03}{70.6} \approx 3.4 \times 10^{-8}.$$

We therefore use 3×10^{-8} as the approximate inferred 70 B floor.

The sub-floor markers shown for 1 B and 8 B in Figure 8 are extrapolated from the nearest candidate inside the adaptation region, while the sub-floor regime at 70 B remains uncharacterized.

The measured ceilings are 3×10^{-5} at 1 B, 1×10^{-5} at 8 B, and 1×10^{-6} at 70 B. Within the adaptation region, we select the ceiling at each scale: the highest learning rate that satisfies all

adaptation and preservation constraints. This criterion favors faster adaptation per training step while remaining consistent across model sizes.

Cross-scale trends and established priors. Using $N_{1B} = 1.23 \times 10^9$, $N_{8B} = 8.03 \times 10^9$, and $N_{70B} = 70.6 \times 10^9$, the measured ceiling tightens by $3.0\times$ from 1 B to 8 B and by $10\times$ from 8 B to 70 B.

The 1 B-to-8 B ratio is consistent with the $1/\sqrt{N}$ learning-rate scaling reported for hyperparameter transfer across model sizes [Yang et al., 2022], which predicts a $2.6\times$ reduction between these scales.

The 8 B-to-70 B ratio is closer to $1/N$ scaling, which predicts an $8.8\times$ reduction and is consistent with the scale-dependent learning-rate decline observed in pre-training [Hoffmann et al., 2022] and continued pre-training [Ibrahim et al., 2024]. The observed 8 B-to-70 B reduction is approximately $1.1\times$ steeper than the $1/N$ prediction. We attribute this additional tightening to the greater sensitivity of the 70 B base model to parameter updates after extensive original pre-training.

Benchmark-level results provide further evidence of this behavior. At 70 B, no continued pre-training candidate exceeds the Llama 3.1 70 B Choclo HardTruth score of 26%. At 8 B, the best continued pre-training candidate improves over the Llama 3.1 8 B base by only 0.6% on Choclo, across independently trained candidates whose scores span only 1.0%.

These observations indicate that the qualitative cross-scale trend is transferable, but the specific learning-rate values are not.

70B candidates. We train four 70 B configurations: two inside the measured adaptation region and two above its ceiling.

Phase 0 alone at $\mathrm{lr} = 10^{-6}$ tests whether the broad Latin American mixture is sufficient without the Phase 1 continuation. Phase 1 at $\mathrm{lr} = 10^{-6}$ with annealing to zero evaluates the configuration suggested by the 70 B adaptation region. Phase 1 at $\mathrm{lr} = 10^{-5}$ tests whether the ceiling identified at 8 B transfers directly to 70 B. Finally, Phase 0+1 at $\mathrm{lr} = 10^{-4}$ provides a substantially higher-rate configuration for measuring catastrophic forgetting.

The two configurations above the measured ceiling degrade performance to different degrees. Relative to the Llama 3.1 70 B MMLU baseline of 76.4%, performance falls to 71.9% at $\mathrm{lr} = 10^{-5}$ and to 53.1% at $\mathrm{lr} = 10^{-4}$.

The $\mathrm{lr} = 10^{-5}$ candidate remains within the 5% MMLU forgetting budget but falls outside the adaptation region because it regresses on Choclo. This result illustrates that being outside the adaptation region does not automatically imply catastrophic forgetting.

The $\mathrm{lr} = 10^{-4}$ configuration produces much broader degradation, reducing MMLU by 23.3%, HellaSwag by 12.0%, and Choclo by 10.0%. We therefore classify this configuration as catastrophic forgetting. The distinction is important: rates moderately above the ceiling may violate one adaptation-region constraint, whereas substantially higher rates can degrade both regional and general-domain capabilities.

In contrast, Phase 1 at $\mathrm{lr} = 10^{-6}$ with annealing to zero preserves MMLU at 76.6% and HellaSwag at 85.8%, effectively matching the base model, while achieving the highest macro-averaged Trueque score among the in-region candidates. We therefore select $\mathrm{lr} = 10^{-6}$ as the 70 B operating point that

best balances Latin American adaptation with retention of pre-trained knowledge. This configuration becomes the pre-trained base model for LatamGPT 70 B.

Compute cost. The 1 B and 8 B exploration requires approximately two orders of magnitude less compute than full training at 70 B. The framework reduces 70 B recipe selection to four targeted validation runs rather than an open-ended search at the largest scale.

The additional cost consists of one exploration campaign at 1 B and one at 8 B. Both campaigns also provide useful recipes and scaling evidence for future continued pre-training work at those model sizes.

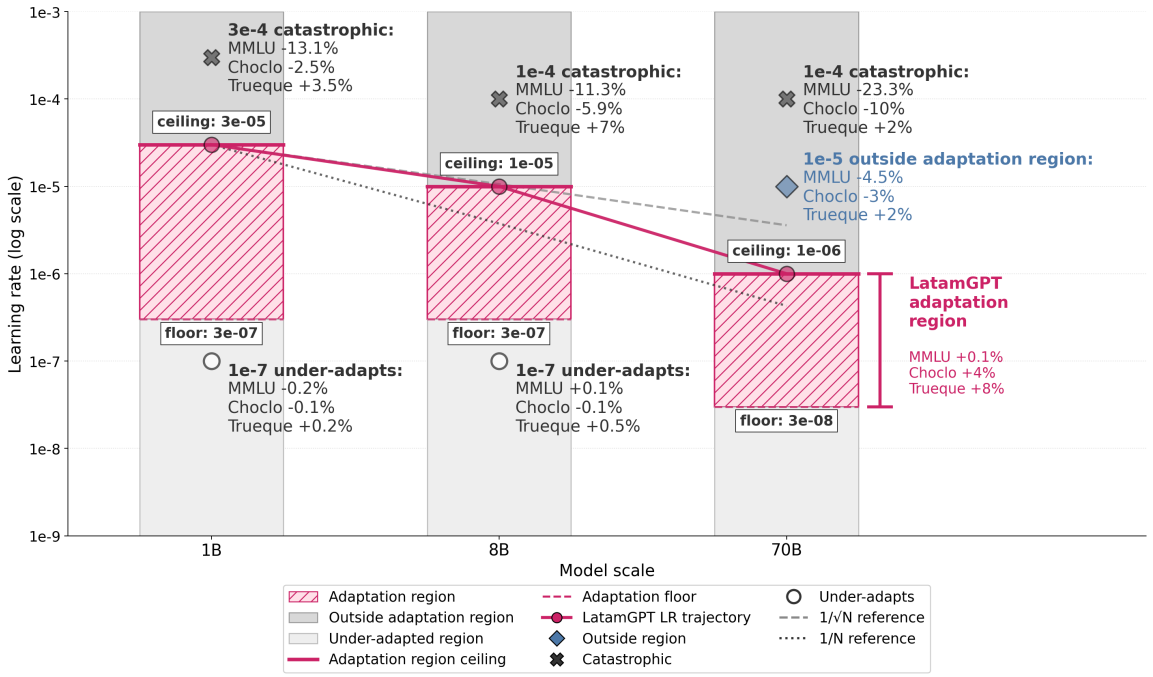


Figure 8: Adaptation regions across the 1 B, 8 B, and 70 B scales. The shaded band at each scale marks the adaptation region between the floor and ceiling learning rates: $[3 \times 10^{-7}, 3 \times 10^{-5}]$ at 1 B, $[3 \times 10^{-7}, 1 \times 10^{-5}]$ at 8 B, and $[3 \times 10^{-8}, 1 \times 10^{-6}]$ at 70 B. The 1 B and 8 B boundaries are measured from the candidate runs. The 70 B floor is estimated by applying the $1/N$ scaling rule to the measured 8 B floor of 3×10^{-7}, yielding approximately 3×10^{-8}. The gray region above each ceiling contains learning rates outside the adaptation region. Rates moderately above the ceiling may produce regional regression or forgetting, whereas substantially higher rates enter the catastrophic-forgetting regime. The region below each floor marks under-adaptation. Annotations report their corresponding MMLU, Choclo, and Trueque changes relative to the same-scale base model.

Table 10 summarizes all models trained for this analysis. The selected candidate at each scale is shown in bold. The final 70 B configuration is CPT Phase 1 with $\mathrm{lr} = 10^{-6}$ and annealing to zero.

Figure 9 visualizes the adaptation-forgetting trade-off among the 8 B candidates. The shaded band marks the adaptable set $\mathbb{F}_s$, which allows at most 5% forgetting on MMLU and HellaSwag while requiring non-negative Choclo adaptation. The catastrophic $\mathrm{lr} = 10^{-4}$ candidate lies well outside

Table 10: Trained model inventory across the 1 B, 8 B, and 70 B scales. Gains are signed percentage changes relative to the corresponding same-scale Llama base model. The base scores are MMLU 38.4% and Trueque 1.3% at 1B; MMLU 64.2% and Trueque 15.8% at 8 B; and MMLU 76.4% and Trueque 39.0% at 70 B. Bold rows indicate the selected candidate at each scale.

Configuration	LR	Schedule	Role	MMLU Δ	Trueque Δ
1 B candidates (Llama 3.2 1 B base)					
1 B stem 0.15 lr=3e-4 partial	3×10^{-4}	partial	Catastrophic	-13.1%	$+3.5\%$
1 B stem 0.15 lr=3e-5 partial	$\mathbf{3 \times 10^{-5}}$	**partial**	**Selected at limit**	$\mathbf{-3.6\%}$	$\mathbf{+4.1\%}$
1 B stem 0.15 lr=3e-6 partial	3×10^{-6}	partial	At floor	$+0.7\%$	$+0.9\%$
1 B stem 0.30 lr=3e-6 partial	3×10^{-6}	partial	Replay variant	$+1.0\%$	$+1.6\%$
1 B stem 0.15 lr=1e-7 partial	1×10^{-7}	partial	Under-adapt	-0.2%	$+0.2\%$
8 B candidates (Llama 3.1 8 B base)					
8 B stem 0.15 lr=1e-4 partial	1×10^{-4}	partial	Catastrophic	-11.3%	$+7.0\%$
8 B stem 0.15 lr=1e-5 partial	$\mathbf{1 \times 10^{-5}}$	**partial**	**Selected at limit**	$\mathbf{-1.5\%}$	$\mathbf{+9.6\%}$
8 B stem 0.15 lr=1e-5 zero	1×10^{-5}	zero	Schedule comparison	-1.8%	$+8.8\%$
8 B stem 0.30 lr=1e-5 partial	1×10^{-5}	partial	Replay variant	-1.2%	$+8.0\%$
8 B stem 0.15 lr=1e-6 partial	1×10^{-6}	partial	PPL-skip base	-0.3%	$+5.2\%$
8 B stem 0.15 lr=1e-6 zero	1×10^{-6}	zero	Schedule comparison	-0.4%	$+7.8\%$
8 B stem 0.30 lr=1e-6 partial	1×10^{-6}	partial	Replay variant	-0.05%	$+6.4\%$
8 B PPL-skip 25%	1×10^{-6}	partial	PPL-skip, 75% data	-0.5%	$+5.2\%$
8 B PPL-skip 50%	1×10^{-6}	partial	PPL-skip, 50% data	-0.4%	$+4.6\%$
8 B PPL-skip 75%	1×10^{-6}	partial	PPL-skip, 25% data	-0.3%	$+3.6\%$
8 B stem 0.15 lr=1e-7 partial	1×10^{-7}	partial	Under-adapt	$+0.1\%$	$+0.5\%$
70 B candidates (Llama 3.1 70 B base)					
CPT Phase 0 ($\mathrm{lr} = 10^{-6}$)	1×10^{-6}	zero	Phase 0 only	0.0%	$+2.0\%$
CPT Phase 1 ($\mathrm{lr} = 10^{-6}$)	$\mathbf{1 \times 10^{-6}}$	**zero**	**Selected**	$\mathbf{+0.2\%}$	$\mathbf{+5.0\%}$
CPT Phase 1 ($\mathrm{lr} = 10^{-5}$)	1×10^{-5}	partial	Choclo regression	-4.5%	$+2.0\%$
CPT Phase 0+1 ($\mathrm{lr} = 10^{-4}$)	1×10^{-4}	partial	Catastrophic	-23.3%	$+2.0\%$

this region, incurring 11.3% MMLU forgetting while producing less Trueque gain than the in-region $\mathrm{lr} = 10^{-5}$ candidate.

Across the right-hand panels, Choclo tends to move in the same direction as MMLU and HellaSwag. Most in-region candidates produce small gains or losses on all three benchmarks, while aggressive parameter updates reduce them together. This association suggests that Choclo shares some sensitivity with the general-domain benchmarks to the magnitude of the parameter update.

However, this relationship does not imply that Choclo has reached the same degree of performance saturation as MMLU or HellaSwag. Choclo accuracy remains near 25%, leaving substantially more absolute room for improvement. The small variation observed across the current recipes therefore suggests that these configurations do not yet fully exploit the available Choclo headroom. One possible explanation is that Choclo requires forms of regional linguistic and cultural adaptation that are not captured solely by increasing the learning rate or training compute. Further gains may depend more strongly on the composition, coverage, and supervision quality of the regional data.

The Trueque panels show a different structure. Trueque improves across several candidates while MMLU and HellaSwag remain close to their base scores. This indicates that Trueque is more sensitive to the intended effect of the regional training mixture and retains measurable adaptation

headroom under the tested recipes.

Taken together, the panels distinguish two forms of benchmark behavior. MMLU and HellaSwag primarily measure whether general capabilities are preserved. Trueque responds clearly to the intended regional adaptation. Choclo occupies an intermediate position: it has substantial absolute headroom, but the current recipes produce only modest gains, indicating that exploiting this headroom may require more targeted data and training interventions.

Figure 10 places the same experiments in terms of effective training compute, showing which benchmarks improve with additional continued pre-training and which remain near the corresponding base-model performance.

Additional compute breaks the base-model plateau most clearly on Trueque, the primary adaptation benchmark. MMLU and HellaSwag remain tightly clustered around their same-scale base scores for candidates inside the adaptation region, indicating that these recipes preserve general capabilities. Only catastrophic configurations move sharply downward.

Choclo again requires a more nuanced interpretation. In Figure 9, Choclo tends to vary with MMLU and HellaSwag across candidate recipes, suggesting shared sensitivity to the magnitude of parameter updates. Nevertheless, its absolute accuracy remains near 25%, substantially below the accuracy observed on MMLU and HellaSwag. Choclo therefore retains greater potential headroom even though the current continued pre-training recipes produce only modest improvements. Its limited gains should not be interpreted as evidence that the benchmark itself is saturated.

Among the displayed models in the 70 B regime, the selected LatamGPT recipe achieves stronger results than the shown DeepSeek, Qwen, and Gemma baselines on the Latin American benchmarks, with the clearest advantage on Trueque and additional gains on Choclo. At the same time, it remains broadly comparable on MMLU and HellaSwag.

This pattern is consistent with the objective of regional continued pre-training. The additional training shifts the model toward Latin American linguistic, cultural, institutional, and factual distributions, producing larger gains on benchmarks that directly measure the target domain. MMLU and HellaSwag instead measure broad capabilities on which the base model already performs strongly. Similar performance on these benchmarks therefore indicates successful capability preservation rather than an absence of adaptation.

3.4.3 Perplexity-Based Filtering

We next examine how the regional adaptation signal is distributed across examples with different perplexity levels under the base model. In particular, we test whether the examples that contribute to Trueque improvements are concentrated in the highest-perplexity tail or spread across a broader portion of the continued pre-training corpus.

Using WayraPPL [Florez and LatamGPT Team, 2025], a 55M-parameter perplexity estimator distilled from Llama 3.1 8 B logits, we score 83,890 Phase 1 samples. The resulting quartile boundaries are $Q_{25} = 10.30$, $Q_{50} = 12.87$, and $Q_{75} = 16.72$, as shown in Figure 11.

We conduct three perplexity-skip ablations at the 8 B scale using $\mathrm{lr} = 10^{-6}$. The experiments remove the lowest-perplexity 25%, 50%, and 75% of examples, respectively. Figure 12 reports their effect on Trueque performance.

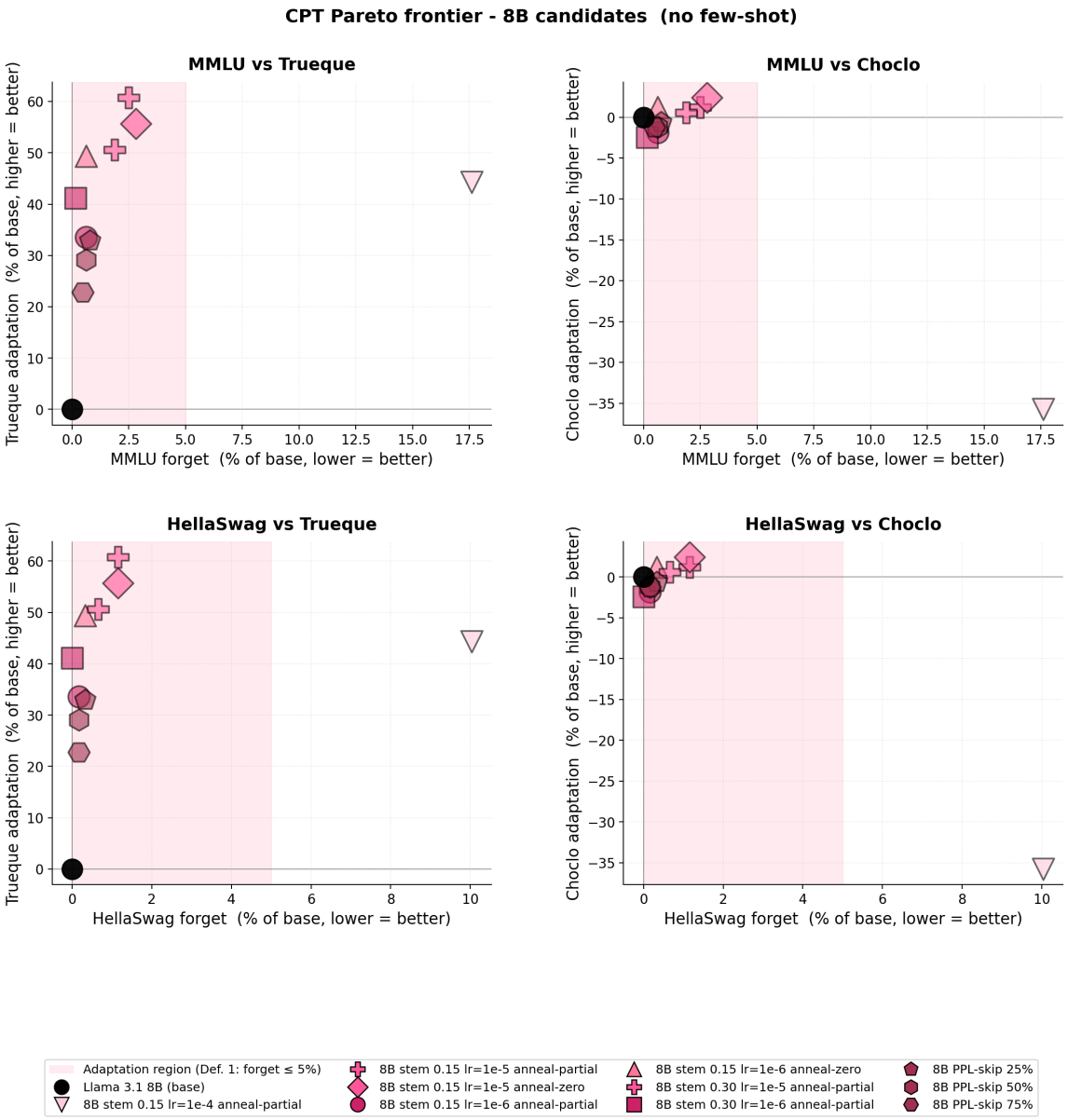


Figure 9: CPT Pareto frontier at the 8 B scale under no-few-shot evaluation. Each panel plots normalized forgetting on the x-axis against normalized adaptation on the y-axis, expressed as a percentage of the Llama 3.1 8 B base score. The top row compares MMLU forgetting with Trueque adaptation on the left and Choclo adaptation on the right. The bottom row compares HellaSwag forgetting with Trueque adaptation on the left and Choclo adaptation on the right.

Removing the lowest-perplexity quartile produces little change on Trueque, suggesting that examples already modeled confidently by the base model provide limited marginal adaptation under this specific training configuration. This result does not imply that these examples are low quality or unnecessary for the complete training mixture. Rather, it indicates that their incremental contribution to the measured Trueque gain is smaller under the tested 8 B recipe.

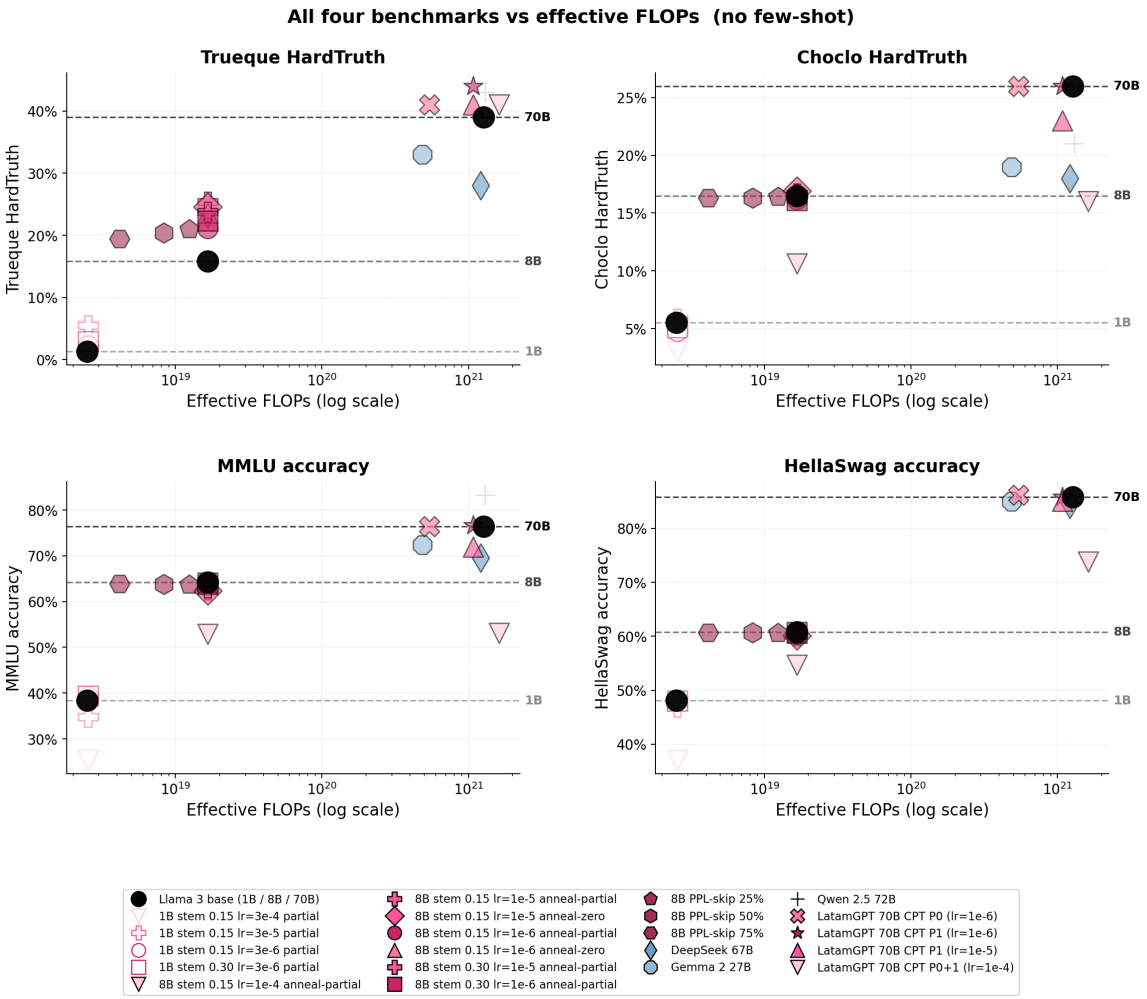


Figure 10: Benchmark performance against effective training FLOPs on a logarithmic x-axis. Black circles mark the 1 B, 8 B, and 70 B Llama base models, while dashed horizontal lines show the corresponding base-model accuracy. The selected LatamGPT configuration produces its clearest gains on Trueque and additional improvements on Choclo while preserving performance close to the base model on MMLU and HellaSwag. Although Choclo varies with the general-domain benchmarks across training recipes, its absolute accuracy remains near 25%, indicating substantially greater remaining headroom.

More aggressive filtering produces a different result. Trueque performance begins to decline once filtering exceeds 25% and decreases monotonically as additional examples are removed. The regional factual signal measured by Trueque is therefore not concentrated exclusively in the highest-perplexity quartile. Instead, it is distributed across the upper three quartiles, including examples that are only moderately surprising to the base model but still contribute complementary Latin American knowledge.

This finding has direct implications for constructing the adaptation mixture. Selecting only the highest-perplexity examples would reduce training volume, but it would also discard useful regional evidence. Perplexity is therefore informative for identifying examples with lower marginal adaptation

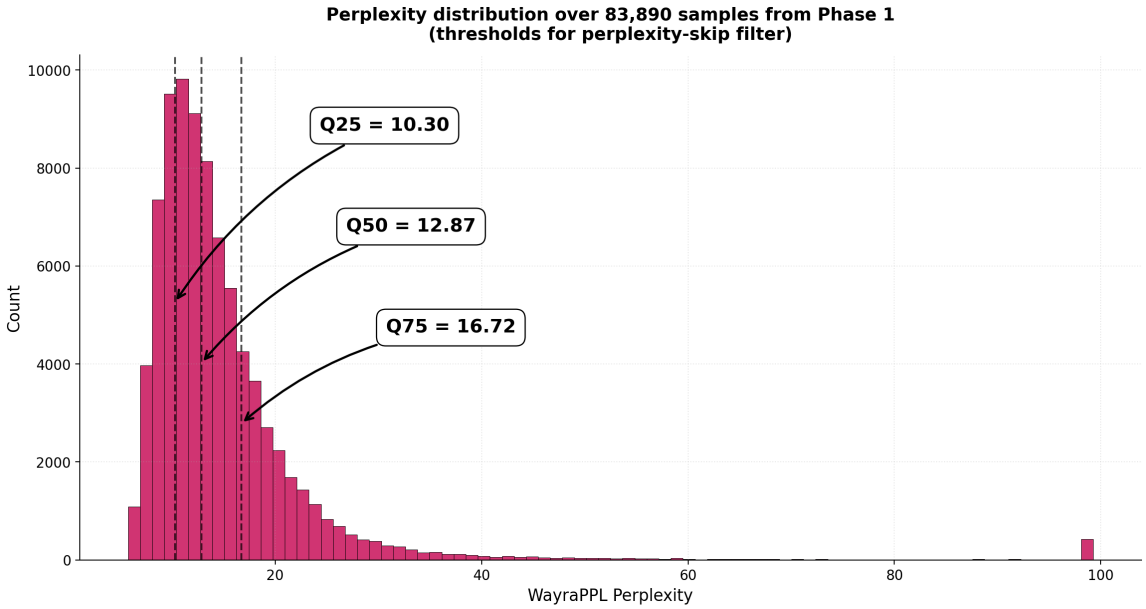


Figure 11: Distribution of perplexity scores over 83,890 Phase 1 samples. The quartile boundaries are $Q_{25} = 10.30$, $Q_{50} = 12.87$, and $Q_{75} = 16.72$. Lower-perplexity examples are predicted more confidently by the base model and may therefore provide smaller marginal gains for regional adaptation under a fixed training budget.

value, but it should not be used as a hard proxy for data quality or as the sole selection criterion.

Phase 1 follows this broader principle. It concentrates informative regional material through stricter quality filtering, targeted upsampling, institutional and alliance-contributed data, synthetic regional content, and memory replay, while retaining examples across a broad range of perplexity values. The ablation supports this selective but heterogeneous construction: moderate filtering preserves most of the measured adaptation benefit, whereas aggressive filtering removes regional signal required to improve Trueque.

These findings refine prior results suggesting that approximately 30% of pre-training tokens may be sufficient for general English evaluations [Ibrahim et al., 2024, Marion et al., 2024]. For regional continued pre-training, the efficient frontier depends on the target capability. Reducing examples with lower marginal utility can improve data efficiency, but restricting training to only the highest-perplexity tail weakens regional factual adaptation.

4 Post-training

An initial round of post-training yielded poor downstream performance. After supervised fine-tuning (SFT) an early continual pre-training (CPT) checkpoint, the resulting model performed poorly across our evaluations and was particularly weak on multiple-choice and alternative-selection tasks, often failing to commit reliably to a single option. This initial SFT stage used a minimal data mixture comprising Tülu 3 and approximately 100K Wikipedia-derived cultural QA pairs, with no safety data.

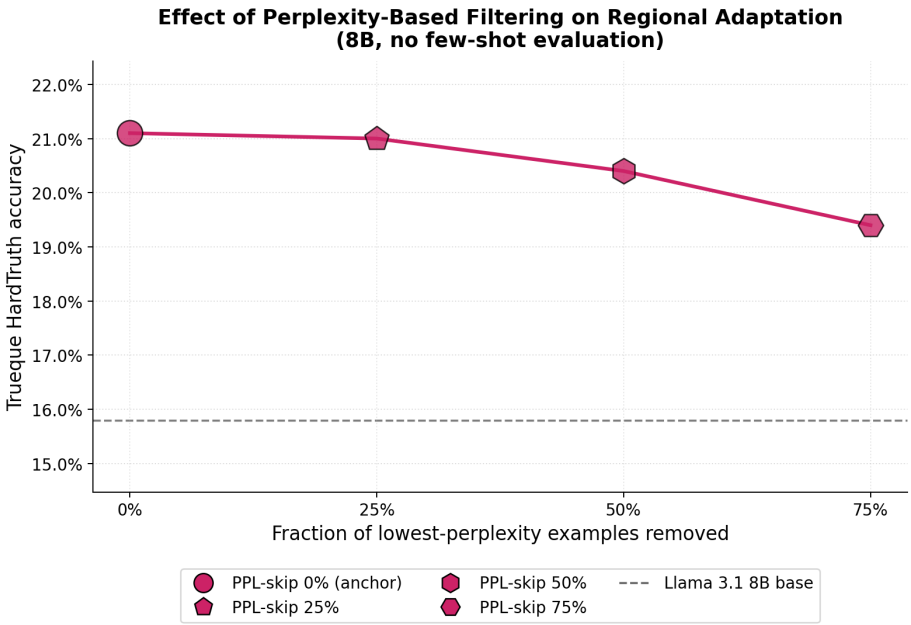


Figure 12: Effect of perplexity-based filtering on Trueque performance at the 8 B scale. Removing the lowest-perplexity quartile produces little change under the tested recipe, whereas more aggressive filtering progressively reduces regional adaptation. The result indicates that the Trueque adaptation signal is distributed across the upper three perplexity quartiles rather than concentrated exclusively in the highest-perplexity tail.

The experiments in this section were therefore designed to determine whether these shortcomings could be addressed through changes to the SFT data and training recipe, or whether they primarily reflected limitations inherited from the underlying CPT checkpoint.

To investigate these questions efficiently, we conducted the main post-training experiments at 8 B scale, where rapid iteration over data mixtures and hyperparameters was feasible. We pursued three lines of investigation. First, because the most acute failure involved multiple-choice behavior, we assembled a unified MCQ fine-tuning dataset (`sft_mcqa_unified`) and compared two integration strategies: including it directly in the SFT mixture or applying it in a dedicated post-SFT stage. Second, because the initial SFT mixture contained no safety data, we evaluated a separate safety fine-tuning stage and measured its effect on safety behavior. Third, as a control, we applied the same SFT recipe to the original Llama 3.1 70 B model without continual pre-training. The alternative-selection failure did not appear in this setting, suggesting that it originated primarily in the CPT stage rather than in the SFT recipe itself. Nevertheless, the final SFT mixture retains the MCQ dataset as a precaution against residual selection bias.

All subsequent experiments in Sections 4.1–4.3 are ablation studies conducted on the 8 B model; the resulting insights inform the final 70 B training recipe (Section 4.4).

4.1 Supervised Fine-tuning: 8 B

Our central question is whether `sft_mcqa_unified` is more effective when incorporated directly into the main SFT mixture or applied in a separate fine-tuning stage after SFT. We conduct SFT

on the LatamGPT CPT checkpoint using a multilingual data mixture with sequence packing and prompt masking. At 8 B scale, we use a learning rate of $\mathrm{lr} = 5 \times 10^{-6}$, following the Tülu 3 SFT configuration [Lambert et al., 2024]; the corresponding 70 B run uses $\mathrm{lr} = 2 \times 10^{-6}$, as reported in Table 12. All runs use the `deepspeed_tulu3` ZeRO-3 configuration.

We compare two base SFT mixtures: **Base-1**, which excludes MCQ Unified, and **Base-2**, which includes `sft_mcqa_unified`. The two configurations are otherwise identical. Each resulting base checkpoint is then subjected to a second fine-tuning stage using safety data, yielding **FT-1** from Base-1 and **FT-2** from Base-2. For this second stage, we sweep three learning rates: 5×10^{-6}, 5×10^{-7}, and 5×10^{-8}.

Table 11 summarizes the dataset composition of each configuration. The best-performing setup—**Base-2 + FT-2**, which incorporates MCQ Unified during the main SFT stage and then applies dedicated safety fine-tuning—was subsequently replicated at 70 B scale, as described in Section 4.4.

Table 11: SFT data mixture configurations for 8 B ablations. Fine-tune stages (FT-1, FT-2) sweep $\mathrm{lr} \in \{5 \times 10^{-6}, 5 \times 10^{-7}, 5 \times 10^{-8}\}$. MCQ Unified = `sft_mcqa_unified` (~370K examples from ten benchmarks; see Table 7).

Config	Dataset	Size
Base-1	Wikipedia QA Iberoamérica (1M)	1,000K
	Tülu 3 SFT	939K
	Copuchat (February 2025)	23K
	Identity	15K
Base-2	Wikipedia QA Iberoamérica (1M)	1,000K
	Tülu 3 SFT	939K
	Copuchat (February 2025)	23K
	Identity	15K
	MCQ Unified	370K
FT-1 (from Base-1)	Safety SFT Sample	50K
	MCQ Unified	370K
FT-2 (from Base-2)	Safety SFT Sample	50K

4.2 Multiple Choice Questions (MCQ) Fine-tuning: 8 B

MCQ fine-tuning applies `sft_mcqa_unified` (described in Section 2) as a dedicated fine-tuning stage on the SFT checkpoint. The objective is to reduce systematic choice bias and improve calibrated accuracy on the Choclo and LatamGPT benchmarks. The negative rationalization technique applied to `synthetic_latam_mcq` produces contrastive supervision signal that is particularly effective at correcting position bias in multiple-choice settings.

4.3 Safety Fine-tuning: 8 B

To validate whether the Spanish-translated safety dataset provides a useful training signal, we perform a controlled safety fine-tuning experiment at the 8 B scale. Starting from the SFT checkpoints described in Section 4.1, we apply a second fine-tuning stage using the translated safety data and sweep learning rates $\{5 \times 10^{-6}, 5 \times 10^{-7}, 5 \times 10^{-8}\}$. We compare the two SFT starting points

(Base-1 without MCQ Unified, Base-2 with MCQ Unified), producing checkpoints denoted FT-1 (from Base-1) and FT-2 (from Base-2). Training uses the same sequence-packing and DeepSpeed ZeRO-3 configuration as the main SFT stage.

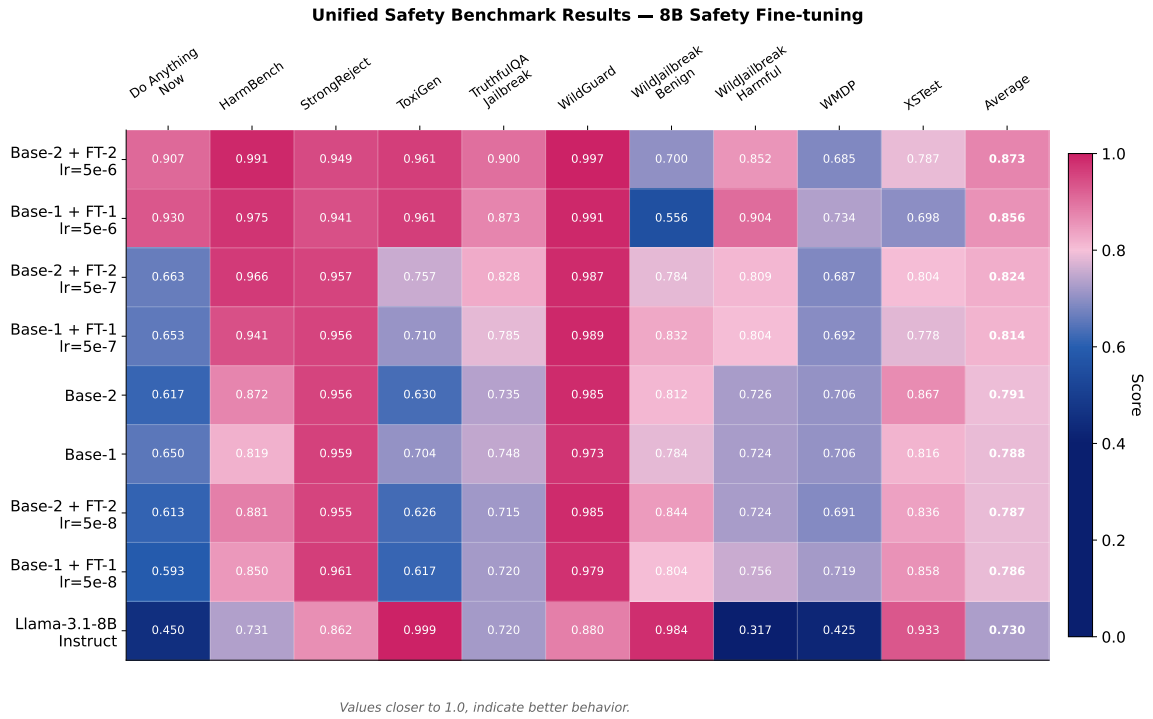


Figure 13: Safety fine-tuning benchmark results for the 8 B model. The unified evaluation is reported on a 0–1 scale (higher is better) and is computed as the unweighted mean across the included benchmarks; the suite includes jailbreak, toxicity, hazardous-knowledge and over-refusal evaluations (see cited references for the complete list).

Figure 13 shows that the translated safety data provides a clear and consistent safety signal. For all tested learning rates, safety fine-tuned checkpoints outperform `Llama-3.1-8B-Instruct` on the unified safety average. The best configuration, **Base-2 + FT-2** with $\mathrm{lr} = 5 \times 10^{-6}$, reaches an average score of 0.873 compared with 0.730 for `Llama-3.1-8B-Instruct`, an absolute gain of 0.143 points. The second-best configuration, **Base-1 + FT-1** (same learning rate), reaches 0.856.

These results indicate that applying safety supervision as a dedicated post-SFT fine-tuning stage consistently improves the unified safety average, which we interpret as evidence that the translated safety mixture provides a cleaner optimization signal after the model has already learned general instruction-following behavior.

4.4 Supervised Fine-tuning: 70 B

Based on the 8 B ablations, the 70 B SFT uses the **Base-2** mixture followed by safety fine-tuning (**FT-2**): the same configuration that achieved the best overall results on the 8 B model. The Base-2 mixture includes `sft_mcqa_unified` directly in the SFT stage rather than as a separate fine-tuning step (see Section 2 for the full data description).

The 70B post-training recipe follows the Tülu 3 70 B SFT configuration [Lambert et al., 2024] unchanged, summarized in Table 12. Reusing an established recipe lets us attribute downstream evaluation differences to data and model choices rather than to training-recipe effects.

Table 12: Supervised fine-tuning hyperparameters for LatamGPT 70 B, following the Tülu 3 70B recipe [Lambert et al., 2024].

Hyperparameter	**Value**
Learning rate	2×10^{-6}
LR schedule	Linear (warmup then linear decay)
Warmup ratio	0.03
Number of epochs	2
Effective batch size	128
Max sequence length	4,096 tokens
Optimizer	AdamW
Weight decay	0.0
Sequence packing	Enabled (with prompt masking)
Distributed training	DeepSpeed ZeRO-3 (`deepspeed_tulu3`)

5 Evaluation

We evaluate LatamGPT models on four benchmarks: MMLU, HellaSwag, Trueque, and Choclo. Each benchmark is evaluated under two prompt formats and two few-shot conditions, yielding four configurations per benchmark.

Prompt formats. Both formats use plain-text prompts without chat template tokens. The **unprompted** format presents the question directly, allowing the model to complete freely:

```
Pregunta: ¿A qué familia pertenece Tillandsia kammii?
Respuesta:
```

The **prompted** format prepends a structured set of rules that constrain the response language, length, and style:

```
Responde de forma precisa y concisa la siguiente pregunta:
¿A qué familia pertenece Tillandsia kammii?
REGLAS PARA TU RESPUESTA:
1. Máximo 100 palabras -- sé directo y conciso
2. Enfócate SOLO en responder lo que se pregunta
3. Formato de párrafo corrido -- NO uses listas ni viñetas
4. Prioriza la información más relevante y actual
5. Si la pregunta requiere hechos verificables, basate en
   información precisa
6. Evita detalles tangenciales o comparaciones no pedidas
```

```
7. Mantén un tono neutral, informativo y profesional
8. Si hay múltiples aspectos relevantes, sintetiza lo esencial
9. No expliques tu razonamiento
10. Responde en el mismo idioma de la pregunta (español o portugués)
Respuesta:
```

All models are evaluated with the same prompts. The prompted format is similar to zero-shot chain-of-thought prompting in that it guides the model's output without modifying its weights.

Few-shot condition. For all four benchmarks, the few-shot condition prepends in-context question-answer pairs to the prompt before the target question, using the same format as the main question. This is the standard few-shot evaluation convention and applies identically to both unprompted and prompted formats.

5.1 Pre-training Benchmarks

5.1.1 MMLU

MMLU [Hendrycks et al., 2020] measures accuracy across 57 academic subjects. Table 13 and Figure 14 report results under all four configurations.

A key finding is that prompted and unprompted scores are nearly identical across all models, differing by at most 0.4 points. For instance, CPT Phase 1 ($\text{lr} = 10^{-6}$) scores 76.6% in both formats at 0-shot, and Llama 3.1 70B scores 76.4% unprompted versus 76.2% prompted. This is expected for a multiple-choice benchmark evaluated by log-likelihood scoring, in which the rules in prompted format (answer in Spanish, be concise, avoid lists) are orthogonal to selecting the highest-probability choice. This near-zero sensitivity to prompt format on MMLU serves two purposes. First, it validates that the prompted format does not artificially inflate or deflate scores on knowledge benchmarks, making prompted results on Trueque and Choclo genuine signal about regional knowledge rather than format artifacts [Sclar et al., 2024, Mizrahi et al., 2024]. Second, it confirms that CPT Phase 1 ($\text{lr} = 10^{-6}$) retains pre-trained knowledge in a format-robust manner.

The 23-point drop for CPT Phase 0+1 ($\text{lr} = 10^{-4}$) is symmetric across prompted and unprompted (53.1% vs. 53.2%), which rules out prompt sensitivity as an explanation and confirms that the loss resides in the model's parametric knowledge rather than in its ability to interpret evaluation formats [Kirkpatrick et al., 2017]. This symmetry is a technically important property, it means the forgetting-adaptation tradeoff can be measured equivalently in both evaluation modes, and that the choice of format does not confound comparisons across learning rate configurations.

5.1.2 HellaSwag

HellaSwag [Zellers et al., 2019] measures commonsense NLI via sentence completion. Table 14 and Figure 15 report results across all four configurations.

As with MMLU, prompted and unprompted scores are essentially identical across all base and CPT models, with maximum differences of 0.1 to 0.3 points. HellaSwag requires ranking sentence continuations by plausibility, a task whose outcome is determined by the model's commonsense representations and is insensitive to output-style constraints.

Table 13: MMLU accuracy (%). **Bold** indicates best LatamGPT result per column.

Model	Unprompted		Prompted	
	0-shot	5-shot	0-shot	5-shot
Open baselines				
Llama 3.1 70B	76.4	78.9	76.2	78.9
DeepSeek 67B	69.5	71.6	69.5	71.6
Gemma 2 27B	72.4	75.2	72.4	75.2
Qwen 2.5 72B	83.3	86.1	83.4	86.0
LatamGPT models				
CPT Phase 0+1 (lr=10^{-4})	53.1	54.8	53.2	55.3
CPT Phase 0 (lr=10^{-6})	76.4	78.9	76.5	78.8
CPT Phase 1 (lr=10^{-6})	**76.6**	**78.7**	**76.6**	**78.7**
CPT Phase 1 (lr=10^{-5})	71.9	75.4	71.6	75.4

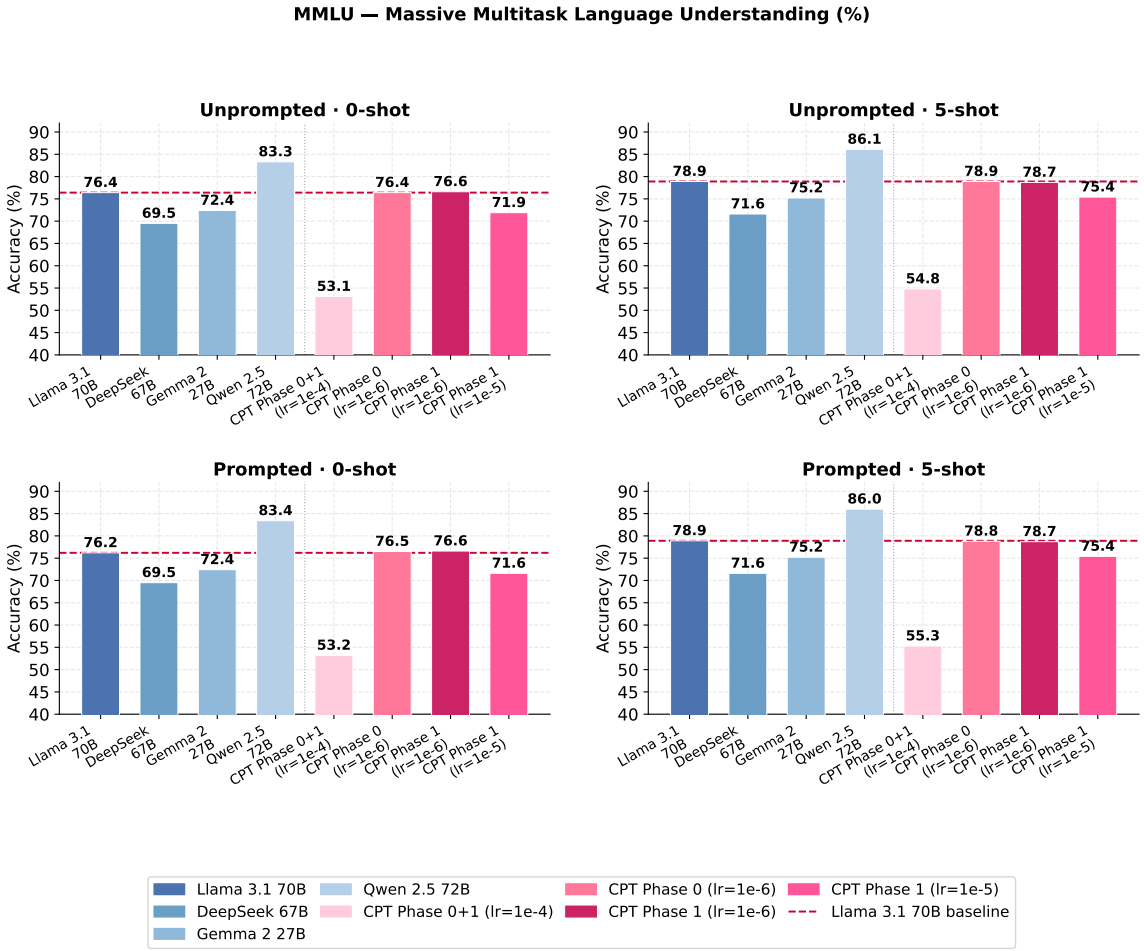


Figure 14: MMLU accuracy (%) across four evaluation configurations. Top row: unprompted format under 0-shot (left) and 5-shot (right). Bottom row: prompted format under 0-shot (left) and 5-shot (right). Prompted and unprompted scores are nearly identical across all models. Dashed line indicates Llama 3.1 70B baseline.

An important contrasting data point is that a supervised fine-tuned model on Llama 3.1 70B (not shown in figures) shows a 4.6-point degradation on HellaSwag under the prompted format (86.1% unprompted vs. 81.5% prompted at 0-shot). This pattern is consistent with findings that instruction fine-tuning can reduce the robustness of base model representations to prompt variations [Luo et al., 2023]. CPT Phase 1 ($\text{lr} = 10^{-6}$) shows no such degradation: 85.8% unprompted versus 85.8% prompted at 0-shot. The preservation of format-robust representations after CPT is a desirable property for the subsequent SFT stage: a base model whose commonsense representations remain intact and format-independent provides a stronger initialization for instruction fine-tuning than one where those representations have already been disrupted by sensitivity to prompt format.

CPT Phase 0 ($\text{lr} = 10^{-6}$) surpasses the Llama 3.1 70B baseline in both formats at 0-shot (86.2% vs. 85.8%) and all open baselines. This is a non-trivial result: HellaSwag measures commonsense reasoning about everyday situations, and the Phase 0 corpus was not curated for commonsense coverage. A plausible interpretation is that exposure to 296.5 B tokens of Latin American web text and news documents reinforces the linguistic structures and world models that HellaSwag probes, as these documents contain dense everyday descriptions of events, causality, and social interactions in Spanish and Portuguese.

Table 14: HellaSwag normalized accuracy (%). **Bold** indicates best LatamGPT result per column.

Model	**Unprompted**		**Prompted**	
	0-shot	5-shot	0-shot	5-shot
Open baselines				
Llama 3.1 70B	85.8	87.6	85.7	87.6
DeepSeek 67B	84.5	86.4	84.4	86.4
Gemma 2 27B	85.0	86.5	85.0	86.5
Qwen 2.5 72B	86.0	87.2	86.0	87.2
LatamGPT models				
CPT Phase 0+1 (lr=10^{-4})	73.8	76.0	73.6	76.1
CPT Phase 0 (lr=10^{-6})	**86.2**	**88.0**	**86.2**	**88.1**
CPT Phase 1 (lr=10^{-6})	85.8	87.6	85.8	87.5
CPT Phase 1 (lr=10^{-5})	85.1	86.7	84.9	86.3

5.2 Regional Benchmarks

5.2.1 Trueque

Trueque is a Spanish-language factual truthfulness benchmark scored via HardTruth, a binary metric that credits only fully correct answers. Table 15 and Figure 16 report HardTruth scores across all four configurations.

In the unprompted format, CPT Phase 1 ($\text{lr} = 10^{-6}$) achieves 0.44 at no few-shot and 0.51 with few-shot, already surpassing the Llama 3.1 70B baseline (0.39 / 0.46) without any format guidance. This is direct evidence that regional knowledge is genuinely encoded in the model weights after CPT, rather than being elicited by the prompt. In the prompted format, CPT Phase 1 ($\text{lr} = 10^{-6}$) achieves 0.53 at no few-shot, +0.15 points above the prompted Llama 3.1 70B baseline (0.38), and 0.58 with few-shot. The gap between unprompted and prompted scores is larger for LatamGPT CPT models than for baselines, suggesting the prompted rules help the model express its regional knowledge

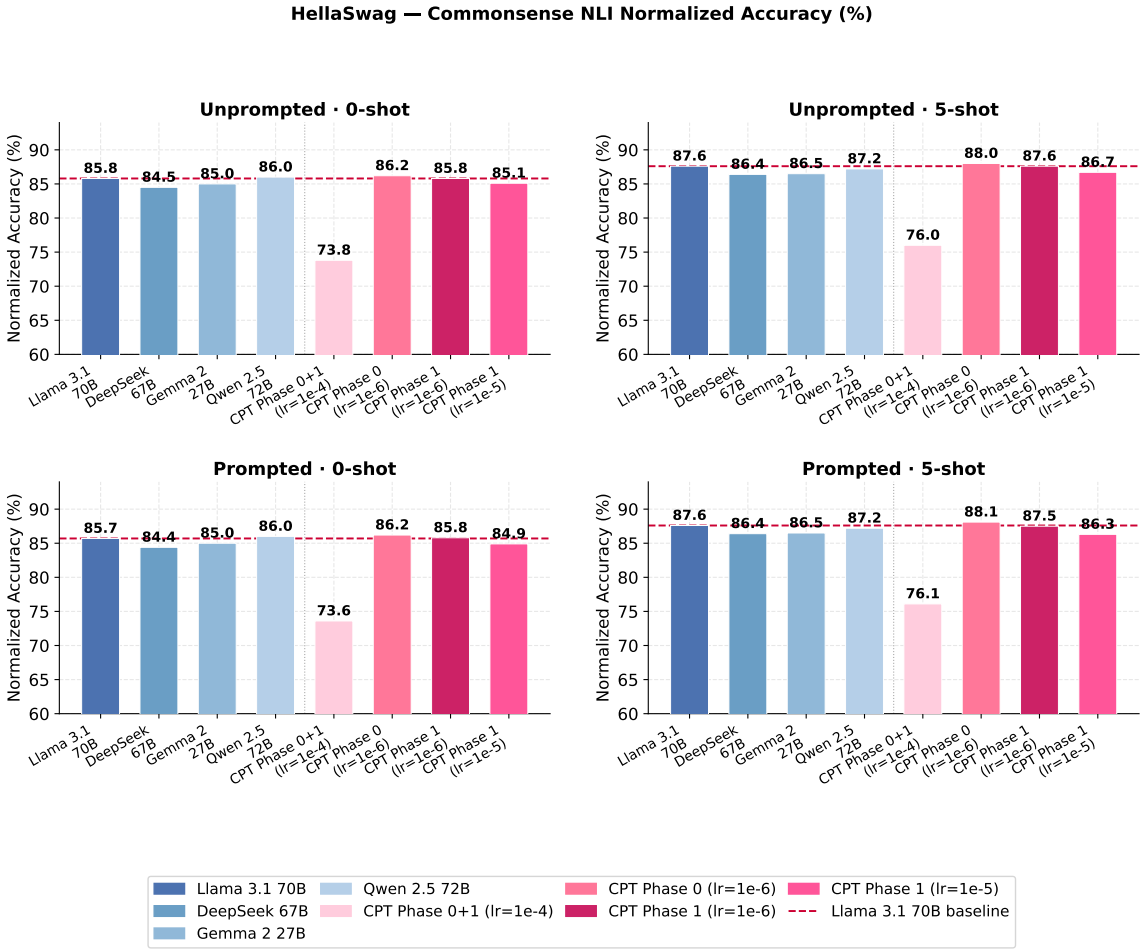


Figure 15: HellaSwag normalized accuracy (%) across four evaluation configurations. Top row: unprompted format under 0-shot (left) and 5-shot (right). Bottom row: prompted format under 0-shot (left) and 5-shot (right). Prompted and unprompted scores are nearly identical for all CPT models, confirming format-robust commonsense reasoning. Dashed line indicates Llama 3.1 70B baseline.

more precisely, producing answers that the judge can score more reliably.

Notably, the few-shot gain for the Llama 3.1 70B baseline in the prompted format (+0.13 points) is larger than for CPT Phase 1 (+0.05 points). This is consistent with the baseline needing the in-context examples to calibrate its response format for Latin American factual questions in Spanish, while the CPT model has already internalized that distribution and gains less from additional calibration.

Finally, CPT Phase 0 ($\text{lr} = 10^{-6}$) reaches the highest score overall at prompted with few-shot (0.60), outperforming CPT Phase 1 (0.58) despite Phase 1 being trained on higher-quality institutional data. A plausible explanation is that Phase 0's broader exposure to the full 296.5 B-token corpus provides an advantage on Trueque, which tests a wide distribution of Latin American facts across 20 countries, even after Phase 1 narrows the focus to high-quality institutional sources.

Table 15: Trueque HardTruth (0–1) across all four configurations. **Bold** indicates best LatamGPT result per column.

Model	Unprompted		Prompted	
	No few-shot	With few-shot	No few-shot	With few-shot
Open baselines				
Llama 3.1 70B	0.39	0.46	0.38	0.51
DeepSeek 67B	0.28	0.30	0.31	0.42
Gemma 2 27B	0.33	0.32	0.33	0.43
Qwen 2.5 72B	0.43	0.55	0.45	0.57
LatamGPT models				
CPT Phase 0+1 (lr=10^{-4})	0.41	0.51	0.46	0.57
CPT Phase 0 (lr=10^{-6})	0.41	0.50	0.48	**0.60**
CPT Phase 1 (lr=10^{-6})	0.44	0.51	**0.53**	0.58
CPT Phase 1 (lr=10^{-5})	0.41	0.49	0.50	0.58

5.2.2 Choclo

Choclo is a LatAm-specific factual QA benchmark covering regional knowledge across 20 countries, evaluated on a 10% sample (10,485 questions). Table 16 and Figure 17 report HardTruth scores across all four configurations.

CPT Phase 1 ($\text{lr} = 10^{-6}$) achieves the highest HardTruth in the prompted with-few-shot configuration (0.35), +0.07 points above the prompted Llama 3.1 70B baseline (0.28), and 0.33 at prompted no few-shot (+0.09 above baseline). As with Trueque, the gain from prompting is larger for CPT models than for baselines: CPT Phase 1 improves from 0.26 unprompted to 0.33 prompted at no few-shot (+0.07), while Llama 3.1 70B improves from 0.26 to 0.24 (no gain). This pattern is consistent with the interpretation that CPT encodes regional knowledge in the model weights but the generation policy defaults to imprecise answers without format guidance; the prompted rules act as a precision constraint that surfaces knowledge already present in the representations.

CPT Phase 0+1 ($\text{lr} = 10^{-4}$) drops to 0.16 unprompted at no few-shot, below all open baselines including DeepSeek 67B (0.18) and Gemma 2 27B (0.19). This indicates that aggressive learning rates not only cause English benchmark forgetting but also disrupt the internal representations used for regional factual retrieval, consistent with catastrophic interference theory [Kirkpatrick et al., 2017]: aggressive parameter updates overwrite weight configurations that support both English and regional knowledge simultaneously. The model partially recovers to 0.24 under the prompted with-few-shot condition, suggesting that some regional knowledge survives but can no longer be accessed through free generation.

Qwen 2.5 72B, the strongest open baseline on MMLU (83.3%), scores below Llama 3.1 70B on Choclo across all configurations (0.21–0.27 vs. 0.24–0.28). This illustrates that general benchmark performance does not predict regional factual knowledge: Qwen was heavily optimized for Chinese and English benchmarks, and its stronger reasoning capacity does not compensate for limited exposure to Latin American-specific content during pre-training.

CPT Phase 0 ($\text{lr} = 10^{-6}$) surpasses all open baselines in both prompted configurations (0.28 / 0.33), confirming that LatAm-specific CPT consistently improves regional factual knowledge even before

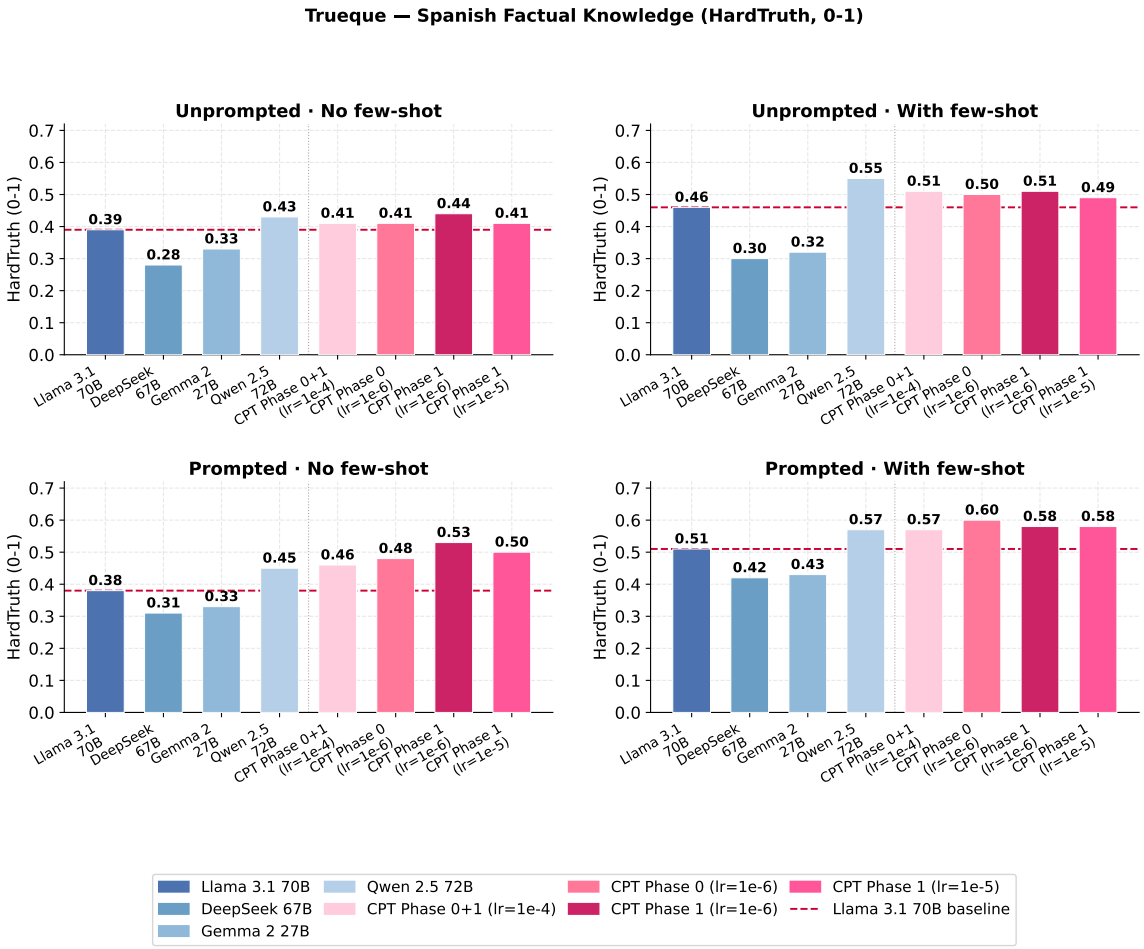


Figure 16: Trueque HardTruth (0–1) across four evaluation configurations. Top row: unprompted format without few-shot (left) and with few-shot (right). Bottom row: prompted format without few-shot (left) and with few-shot (right). Few-shot refers to in-context question-answer pairs prepended to the prompt. Dashed line indicates Llama 3.1 70B baseline for each configuration.

the Phase 1 high-quality institutional data is introduced.

5.3 Post-training Benchmarks

LatamGPT 70B SFT 1.0 is evaluated on three benchmark suites: Trueque 0.2, a regional Latin American factual benchmark; general English benchmarks (MMLU, HellaSwag, TruthfulQA); and safety benchmarks. General and safety benchmarks are evaluated under the default OLMES configuration [Groeneveld et al., 2024]. The comparison set for these includes Qwen3-80B-A3B [Yang et al., 2025], Qwen2.5-72B [Yang et al., 2024], Llama-3.1-70B-Instruct and Llama-3.3-Nemotron-Super-49B [Dubey et al., 2024], Llama-3.1-Nemotron-70B [NVIDIA et al., 2024], Nemotron-3-Super-120B [NVIDIA, 2025a], and Tülu 3-70B [Lambert et al., 2024].

Table 16: Choclo HardTruth (0–1) across all four configurations on the 10% sample (n = 10,485). **Bold** indicates best LatamGPT result per column.

Model	Unprompted		Prompted	
	No few-shot	With few-shot	No few-shot	With few-shot
Open baselines				
Llama 3.1 70B	0.26	0.28	0.24	0.28
DeepSeek 67B	0.18	0.21	0.18	0.22
Gemma 2 27B	0.19	0.21	0.19	0.22
Qwen 2.5 72B	0.21	0.25	0.22	0.27
LatamGPT models				
CPT Phase 0+1 (lr=10^{-4})	0.16	0.19	0.20	0.24
CPT Phase 0 (lr=10^{-6})	0.26	0.28	0.28	0.33
CPT Phase 1 (lr=10^{-6})	0.26	0.29	0.33	**0.35**
CPT Phase 1 (lr=10^{-5})	0.23	0.27	0.27	0.31

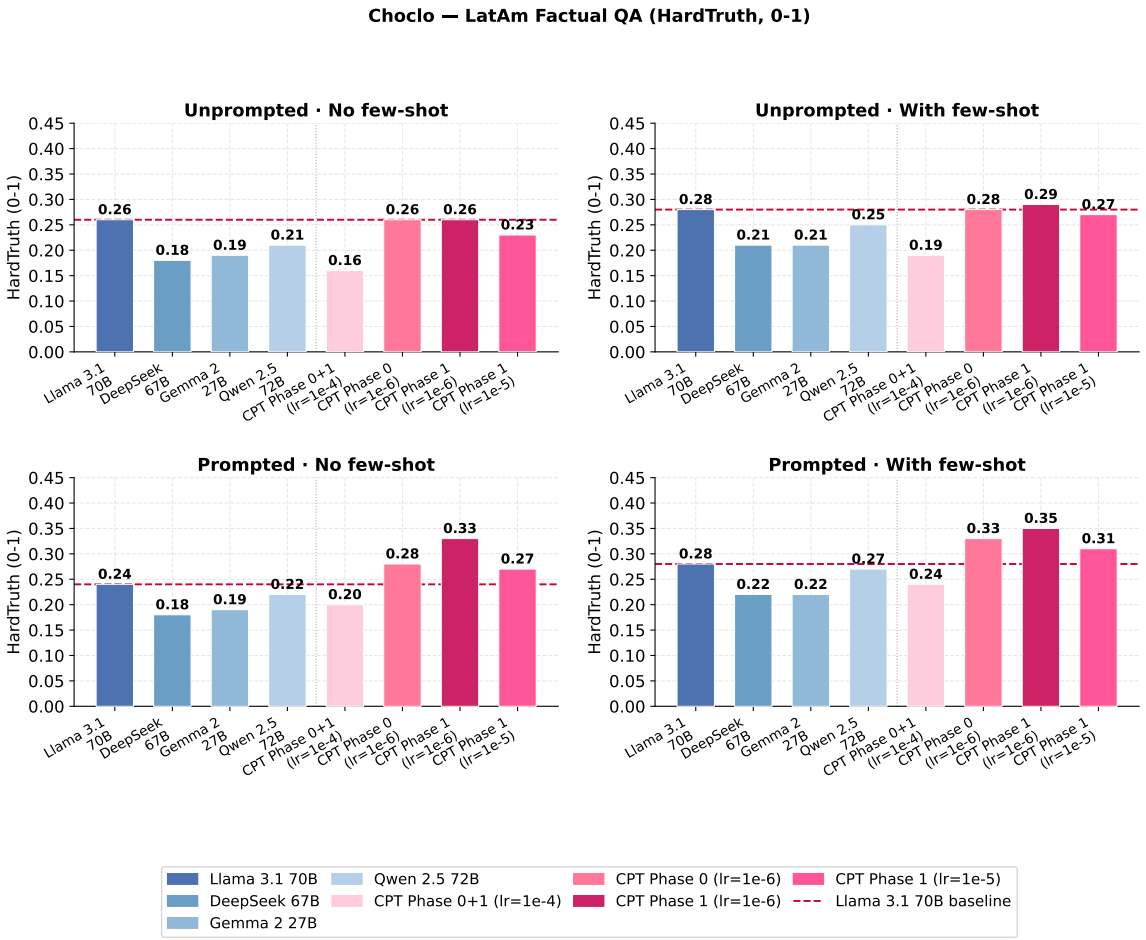


Figure 17: Choclo HardTruth (0–1) across four evaluation configurations on the 10% sample (n = 10,485). Top row: unprompted format without few-shot (left) and with few-shot (right). Bottom row: prompted format without few-shot (left) and with few-shot (right). Dashed line indicates Llama 3.1 70B baseline for each configuration.

5.3.1 Trueque 0.2

Trueque 0.2 is an internal extension of the Trueque regional factual benchmark [LatamGPT Team, 2025] comprising $n = 2{,}296$ Latin American questions; the public beta [LatamGPT Team, 2025] contains 500 questions. All models are evaluated with 5-shot prompting; each response is scored by Qwen2.5-72B [Yang et al., 2024] using 3 judge in-context examples on two dimensions: Hard Truth (binary per-question accuracy, %) and Level of detail (1–10 scale assigned by the judge). The comparison set includes Gemma-4 [Gemma Team et al., 2025], Llama 3.1/3.3 [Dubey et al., 2024], Tülu 3 [Lambert et al., 2024], Qwen2.5 and Qwen3 [Yang et al., 2024, 2025], and NVIDIA Nemotron models [NVIDIA et al., 2024, NVIDIA, 2025b,a] (Table 17).

With 65.8% Hard Truth, the model outperforms Llama 3.1 70B Instruct (63.2%) and Qwen2.5-72B (62.9%), both models with broader multilingual pre-training but no Latin American specialization. This confirms that Latam-specific CPT followed by SFT yields a consistent gain on regional factual knowledge.

Table 17: Trueque 0.2 results for open-source models ($n = 2{,}296$ Latin American factual questions). 5-shot evaluation; Qwen2.5-72B [Yang et al., 2024] judge with 3-shot in-context examples. Level of detail: 1–10 (detail quality). Hard Truth: binary accuracy (%). Models ordered by Hard Truth. **Bold** = LatamGPT 70B SFT 1.0.

Model	Level of detail	Hard Truth (%)
Qwen3.5-122B-A10B	5.97	70.8
Gemma-4-31B	6.08	70.6
Gemma-4-26B-A4B	6.25	69.3
Llama-3.1-Tülu-3-70B	5.71	67.9
Llama-3.1-Tülu-3-70B-SFT	5.47	67.2
LatamGPT 70B SFT 1.0	**5.73**	**65.8**
Nemotron-Super-120B	5.35	64.8
Qwen3-Next-80B-A3B	5.77	64.8
Llama-3.1-Nemotron-70B	5.83	63.3
Llama-3.1-70B-Instruct	4.78	63.2
Qwen2.5-72B-Instruct	5.59	62.9
Gemma-3-27B	5.32	59.4
Nemotron-Nano-30B	5.16	57.6
Llama-3.3-Nemotron-Super-49B	5.31	55.3
Gemma-4-E4B	4.60	49.4
Gemma-4-E2B	3.99	38.4

Hard Truth varies considerably across the 20 geographic strata of Trueque 0.2, which comprises 19 individual countries plus a cross-national "Latam" stratum of questions about concepts, people, or culture involving at least two Latin American countries. Table 18 disaggregates performance for a representative selection of models; column N reports the stratum size and rows are sorted by LatamGPT 70B SFT 1.0 score.

Table 18: Trueque 0.2 Hard Truth (%) by country for selected open-source models. N = questions per stratum. The "Latam" category covers questions about concepts, people, or culture spanning at least two Latin American countries. Sorted by LatamGPT 70B SFT 1.0 score (descending). Same evaluation setup as Table 17. **Bold** column/row = LatamGPT 70B SFT 1.0 / Latam category.

Country	N	LatamGPT	Gemma-4-31B	Tülu-3-70B	Llama-3.1-70B
Latam	**17**	**100.0**	**100.0**	**100.0**	**100.0**
Argentina	145	**84.1**	89.0	82.1	77.9
Puerto Rico	55	**78.2**	72.7	83.6	81.8
Nicaragua	77	**77.9**	66.2	80.5	79.2
Honduras	63	**77.8**	65.1	74.6	73.0
Brasil	98	**77.6**	82.7	81.6	83.7
Perú	118	**77.1**	81.4	76.3	71.2
Guatemala	71	**76.1**	74.6	74.6	71.8
Cuba	119	**73.1**	73.1	78.2	77.3
Bolivia	63	**73.0**	77.8	73.0	68.3
Colombia	205	**70.7**	79.5	69.3	68.3
Paraguay	100	**67.0**	67.0	61.0	56.0
Venezuela	51	**66.7**	70.6	70.6	66.7
Costa Rica	93	**63.4**	51.6	61.3	55.9
Rep. Dominicana	175	**63.4**	72.6	66.9	65.1
Uruguay	122	**62.3**	62.3	68.0	55.7
Ecuador	288	**61.1**	74.0	63.2	57.3
México	188	**60.1**	73.9	67.6	60.6
Panamá	75	**56.0**	62.7	57.3	53.3
Chile	96	**26.0**	36.5	35.4	26.0
El Salvador	75	**22.7**	32.0	30.7	16.0

5.3.2 General Benchmarks

On general English benchmarks (Table 19), the model scores 78.4% MMLU, 90.8% HellaSwag, and 55.3% TruthfulQA, placing it in the middle of the 70B-class instruction-tuned comparison set, just below Tülu 3-70B (79.1% MMLU) and within reach of the Llama 3.1 70B Instruct base (82.4%, 91.9%, 66.8% respectively).

A 4-point MMLU gap and a more pronounced 11-point TruthfulQA gap relative to Llama 3.1 70B Instruct are consistent with the known cost of continued pre-training on a non-English corpus followed by domain-specific SFT: shifting the model's parameter budget toward a target language and domain inevitably compresses capacity for other distributions, a phenomenon broadly documented in continued and domain-adapted LLMs [Kirkpatrick et al., 2017]. HellaSwag degrades by less than one point, suggesting that low-level commonsense reasoning is robust to the adaptation.

This trade-off is precisely what Trueque 0.2 quantifies in the target domain. LatamGPT achieves 65.8% Hard Truth, outperforming Llama 3.1 70B Instruct (63.2%) and Qwen2.5-72B (62.9%), two models with broader multilingual pre-training but no Latin American specialization. The regional gain offsets a portion of the English-benchmark cost, reflecting the adaptation-forgetting balance that domain-focused CPT is designed to achieve.

Table 19: General benchmark scores (%) evaluated with the default OLMES configuration [Groeneveld et al., 2024]. MMLU [Hendrycks et al., 2020], HellaSwag [Zellers et al., 2019], TruthfulQA [Lin et al., 2022]. **Bold** = LatamGPT 70B SFT 1.0.

Model	MMLU Avg	MMLU STEM	MMLU Hum.	MMLU Soc.	MMLU Other	HellaSwag	TruthfulQA
Qwen3-80B-A3B-Instruct	85.8	83.9	86.9	89.3	84.2	93.6	60.2
Qwen2.5-72B-Instruct	83.9	80.0	85.6	88.8	83.0	93.9	69.6
Llama-3.1-70B-Instruct	82.4	75.0	86.3	88.4	82.7	91.9	66.8
Llama-3.1-Nemotron-70B-Instruct	81.0	72.0	85.8	87.8	81.8	92.5	62.7
Llama-3.1-Tülu-3-70B-SFT	79.2	69.0	83.7	87.4	80.2	92.1	55.7
Llama-3.1-Tülu-3-70B	79.1	69.4	83.3	87.2	80.3	91.6	67.6
LatamGPT 70B SFT 1.0	**78.4**	**68.9**	**82.9**	**86.7**	**78.8**	**90.8**	**55.3**
Llama-3.3-Nemotron-Super-49B	77.4	67.3	81.5	85.4	79.1	90.2	59.7
Nemotron-3-Super-120B	74.3	68.4	76.8	82.4	72.1	93.7	56.4

5.3.3 Safety Benchmarks

On safety benchmarks (Table 20), the model ranks third with a Safety Avg of 84.4%, substantially above Llama 3.1 70B Instruct (74.7%). Building on the Tülu 3 safety recipe, which leads at 91.0%, provides a strong foundation: DAN (70.3%), HarmBench (89.1%), StrongReject (92.2%), TrustLLM (84.2%), and WildGuard (95.7%) all indicate solid resistance to adversarial and jailbreak inputs.

The notable exception is WildJailbreak (benign), on which the model achieves 63.2%, well below the Tülu 3-70B-SFT reference (80.0%) and Llama 3.1 70B Instruct (90.8%). This pattern, strong on harmful inputs but more conservative behavior on benign ones, is consistent with a known side effect of safety alignment: fine-tuning on curated safety data can increase over-refusal on borderline benign prompts [Röttger et al., 2023]. Future work could mitigate this behavior through more carefully calibrated refusal data or preference optimization explicitly targeting false-positive refusals.

Table 20: Safety benchmark scores (%, higher is safer) evaluated with the default OLMES configuration [Groeneveld et al., 2024]. Safety Avg is the unweighted mean across all benchmarks. DAN [Shen et al., 2023], HarmBench [Mazeika et al., 2024], StrongReject [Souly et al., 2024], ToxiGen [Hartvigsen et al., 2022], TrustLLM [Huang et al., 2024], WildGuard [Han et al., 2024], WJB = WildJailbreak [Jiang et al., 2024], XSTest [Röttger et al., 2023]. **Bold** = LatamGPT 70B SFT 1.0.

Model	*Safety Avg*	*DAN*	*HarmBench*	*WJB (harmful)*	*TrustLLM*	*StrongReject*	*WJB (benign)*	*WildGuard*	*XSTest*	*ToxiGen*
Llama-3.1-Tülu-3-70B-SFT	91.0	91.0	90.0	85.6	88.0	95.3	80.0	98.7	90.2	99.9
Llama-3.1-Tülu-3-70B	85.3	72.0	85.3	61.0	76.0	85.4	98.4	96.4	93.3	100.0
LatamGPT 70B SFT 1.0	**84.4**	**70.3**	**89.1**	**82.2**	**84.2**	**92.2**	**63.2**	**95.7**	**82.4**	**99.9**
Llama-3.3-Nemotron-Super-49B	76.1	48.7	83.1	26.9	65.8	82.3	97.6	89.5	90.9	99.9
Llama-3.1-70B-Instruct	74.7	73.3	71.2	20.4	51.8	87.1	90.8	78.5	92.0	99.1
DeepSeek-LLM-67B-Chat	71.5	42.0	76.6	13.3	66.8	81.8	98.4	78.5	86.4	100.0
Llama-3.1-Nemotron-70B-Instruct	69.6	41.0	70.6	12.6	43.2	85.9	99.6	81.3	92.4	99.9
Hermes-3-Llama-3.1-70B	62.4	54.7	45.3	8.2	33.2	74.8	99.2	64.8	90.0	91.4

6 Discussion

6.1 What Worked and What Did Not

Several empirical findings proved consistent across ablations and are likely to generalize to CPT of other large models on low-resource regional data.

Learning rate is the dominant variable in CPT. Across all Trueque and Choclo ablations, the Phase 1 peak learning rate is the strongest predictor of both adaptation quality and forgetting severity. At $\mathrm{lr} = 10^{-5}$, MMLU drops by up to 10 points relative to the Llama 3.1 baseline regardless of decay schedule length, confirming that excessive learning rate is the primary cause of catastrophic forgetting. CPT Phase 0+1 ($\mathrm{lr} = 10^{-4}$) shows the most extreme case, with MMLU falling 23 points below baseline. The $\mathrm{lr} = 10^{-6}$ setting occupies the optimal point on the forgetting-adaptation frontier, recovering nearly all baseline MMLU and HellaSwag performance while yielding +0.15 points on Trueque HardTruth over the Llama 3.1 base (0.53 vs. 0.38, prompted no few-shot).

Sequence packing with prompt masking is essential. Without sequence packing, training at long context on 128 H200 GPUs produces significant padding overhead, reducing effective GPU utilization. With packing enabled, MFU improves measurably. More importantly, packing without prompt masking causes the model to receive gradient signal from padding tokens and from the packed document boundaries, which degrades instruction-following quality post-SFT. This is consistent with findings reported in Tülu 3 [Lambert et al., 2024].

FP8 is stable for CPT. FP8 training with `fp8_amax_history_len` $=$ 1,024 was numerically stable throughout Phase 0 and Phase 1. No NaN losses were observed after applying the `fp8_amax_compute_algo` $=$ `max` setting. This contrasts with initial attempts using `fp8_amax_history_len` $=$ 16, which produced intermittent gradient norm spikes during domain shift at the Phase 0 to Phase 1 transition.

Phase 0 generalization exceeds Phase 1 on English benchmarks. CPT Phase 0 ($\mathrm{lr} = 10^{-6}$) matches the Llama 3.1 70B baseline on MMLU (76.4%) and surpasses it on HellaSwag (86.2% vs. 85.8%). This likely reflects the Phase 0 data mixture containing a substantial replay fraction of English general knowledge, preventing forgetting before LatAm-specific data is introduced in Phase 1. Additionally, exposure to 296.5 B tokens of Latin American web text, news, and institutional documents appears to reinforce the linguistic structures and world models that HellaSwag probes, even though the corpus was not curated for commonsense coverage.

CPT preserves format-robust representations. Across MMLU and HellaSwag, prompted and unprompted scores for CPT Phase 1 ($\mathrm{lr} = 10^{-6}$) differ by at most 0.1 points, matching the behavior of the Llama 3.1 70B base model. By contrast, the `sft-amd` supervised fine-tuned model shows a 4.6-point HellaSwag degradation under the prompted format (86.1% unprompted vs. 81.5% prompted), consistent with findings that instruction fine-tuning can reduce representational robustness to prompt variations [Sclar et al., 2024, Luo et al., 2023]. The absence of this effect in CPT Phase 1 confirms that conservative-lr CPT acquires regional knowledge without introducing the prompt sensitivity associated with weight-modifying fine-tuning.

Regional knowledge is encoded in weights, not elicited by prompts. CPT Phase 1 ($\mathrm{lr} = 10^{-6}$) already surpasses the Llama 3.1 70B baseline on Trueque in the unprompted format (0.44 vs. 0.39, no

few-shot), without any structured guidance constraining the output. This is parametric evidence: the regional knowledge acquired during CPT is stored in the model weights and is accessible through free generation, not dependent on prompt engineering to surface. The additional gain under the prompted format (+0.09 points) reflects improved answer precision rather than knowledge acquisition.

General benchmark strength does not predict regional knowledge. Qwen 2.5 72B outperforms Llama 3.1 70B on MMLU by 6.9 points (83.3% vs. 76.4%), reflecting stronger general reasoning capacity. Yet on Choclo, Qwen 2.5 72B scores below Llama 3.1 70B across all four configurations (0.21–0.27 vs. 0.24–0.28). This dissociation illustrates that general benchmark performance does not predict regional factual coverage: Qwen's stronger reasoning capacity does not compensate for limited exposure to Latin American-specific content during pre-training. For regional AI development, training data coverage is a necessary condition that cannot be substituted by model scale or general reasoning ability.

High learning rates disrupt regional and English knowledge simultaneously. CPT Phase 0+1 ($\text{lr} = 10^{-4}$) drops to 0.16 HardTruth on Choclo unprompted, below all open baselines including DeepSeek 67B (0.18) and Gemma 2 27B (0.19), despite being trained entirely on Latin American data. This indicates that aggressive parameter updates do not merely overwrite English representations. They also disrupt the internal configurations that support regional factual retrieval, consistent with catastrophic interference theory [Kirkpatrick et al., 2017]. The practical implication is that optimizing solely for regional adaptation at high learning rates is counterproductive: it degrades both what the model already knew and what it was trying to learn.

The MMLU change sign is a single-run saturation diagnostic. The direction of the MMLU change at the selected learning rate provides a low-cost diagnostic of base-model saturation. A positive MMLU delta (the model improves on MMLU after CPT) indicates under-adaptation: the learning rate is below the floor and the model has not reached the feasible region. A near-zero delta indicates the model is within the adaptable set $\mathbb{F}_s$. A negative delta indicates forgetting. This diagnostic requires only one CPT run at the candidate learning rate and can be used to triage the multi-scale search before committing to further ablations.

Adaptation headroom is benchmark-specific, not global. Choclo HardTruth saturates at the 8B and 70B base levels. Across all in-region 8B candidates, Choclo adaptation remains between -0.4% and $+1.0\%$, the visual signature of saturation in the Pareto frontier (Figure 9). This is consistent with Choclo's Wikipedia- and Wikidata-derived content already being represented in Llama 3 pre-training. In contrast, Trueque continues to improve monotonically with CPT across all scales. Saturation is therefore benchmark-specific rather than global: a CPT recipe that saturates Choclo may still have substantial headroom on Trueque, and the selection objective should target the benchmark with the largest remaining adaptation potential.

6.2 Replication Guidance for Other Regional LLMs

LatamGPT v1.1 CPT provides a blueprint replicable for other regional language contexts with access to a 70B-class foundation model and sufficient compute. The following recommendations are grounded in our empirical findings.

Scale-aware recipe. The central transferable result is not a fixed set of hyperparameters but a validation procedure. Search at small scale, estimate the large-scale learning-rate limit using the scaling rule below, then verify with a candidate at the target scale before committing the full compute

budget. Table 21 summarizes the scale-aware recipe used in LatamGPT.

Table 21: Scale-aware CPT recipe. Learning rate follows approximately $1/\sqrt{N}$ from 1B to 8B and $1/N$ from 8B to 70B. Annealing preference reverses between the ceiling and floor learning rates at 8B; validate separately at each target scale.

Decision factor	**1B**	**8B**	**70B**
Replay (STEM) ratio	0.15	0.15	0.15
Peak learning rate	3×10^{-5}	1×10^{-5}	1×10^{-6}
Annealing schedule	partial	partial	to zero

Minimum viable compute. CPT Phase 0 at per-node batch size 18 and $\mathrm{lr}_{\max} = 10^{-6}$ requires 16 nodes of 8-way GPU clusters equivalent to `p5en.48xlarge` (H100/H200 class). Scaling to fewer nodes is feasible by proportionally increasing training duration, but at per-node batch sizes below 8 we observed gradient norm instability. For teams without access to 100+ GPUs, we recommend initializing from a 7B to 13B CPT checkpoint first and scaling to 70B with confirmed data mixtures.

Learning rate selection. Use the base model's pre-training peak learning rate as an upper bound and conduct a brief grid search over $\{10^{-6}, 10^{-5}, 10^{-4}\}$ on a 2 to 5B token slice before committing to a full CPT run. For 70B models, $\mathrm{lr} = 10^{-6}$ has proven effective for domain CPT in our experiments and in prior work [Ibrahim et al., 2024]. Note that the learning-rate limit tightens more rapidly from 8B to 70B than the $1/\sqrt{N}$ rule would predict from the 1B-to-8B transition alone. Always validate the predicted limit with at least one 70B candidate before committing the full compute budget.

Data mixture. We recommend a memory replay fraction of at least 20 to 30% general English or multilingual data throughout Phase 1 to mitigate catastrophic forgetting. Pure regional data in Phase 1 leads to rapid forgetting within the first 10% of Phase 1 training steps.

Infrastructure alternatives. For teams without AWS HyperPod access, the parallelism strategy (TP=8, FSDP=16) is directly transferable to any cluster of nodes with 8 NVLink-connected GPUs per node. The NeMo recipe is portable to other Kubernetes-based clusters by replacing the `p5en.48xlarge` instance type with an equivalent node configuration. FSx for Lustre can be replaced with GPFS, Lustre, or NFS with comparable token throughput if the file system is co-located with compute.

Tokenization and packing. Arrow-format tokenization with first-fit-decreasing (FFD) sequence packing achieves near-100% packing efficiency at 8k context. Prompt masking must be enabled for SFT; omitting it causes letter collapse in MCQ fine-tuning due to gradient signal leaking across document boundaries.

6.3 Limitations

Unbalanced country and topic distribution. The CPT corpus is not balanced across the 20 covered countries. Mexico and Argentina each contribute 13.3% of tokens; Chile 7.4%; Brazil 6.2%; while several smaller Central American and Caribbean nations contribute under 2%. Choclo performance is likely to correlate with token density per country: models may perform significantly better on Chilean and Argentine factual questions than on Honduran or Paraguayan ones.

Synthetic data quality ceiling. Synthetic datasets use Qwen 2.5 72B as the generator. Model-generated data introduces a quality ceiling: the generator's own factual errors and hallucinations

propagate into training datasets.

Indigenous and minority languages. The corpus covers only Spanish and Portuguese. Indigenous languages of Latin America, including Quechua, Guaraní, Nahuatl, Aymara, and over 400 others, are absent from both the CPT corpus and the evaluation benchmarks.

Legal heterogeneity. As documented in the corpus catalog, Latin America lacks text and data mining exceptions to copyright comparable to those available in the EU, USA, and Japan [Schirru et al., 2024]. This restricts the fraction of high-quality institutional data that can be included in future corpus versions and places academic CPT initiatives at a systematic disadvantage relative to commercial actors.

7 Conclusion

We presented LatamGPT 70B, a fully open language model continually pre-trained and post-trained for Latin America. Starting from Llama 3.1 70B, we performed continual pre-training on a 296.5 B-token corpus assembled from data associated primarily with 20 Latin American countries and through collaborations with 79 academic, governmental, and civil society institutions. These institutional partnerships contributed 58.8 B tokens of regionally grounded data that are difficult to obtain through public web crawls alone.

A relevant empirical finding is the tradeoff between regional adaptation and preservation of the capabilities inherited from the base model. At a peak learning rate of $\mathrm{lr} = 10^{-6}$, CPT Phase 1 improves over the Llama 3.1 base model by 39% on Trueque HardTruth (0.53 vs. 0.38, prompted without few-shot examples) and by 25% on Choclo HardTruth (0.35 vs. 0.28, prompted with few-shot examples), while keeping MMLU and HellaSwag performance within 0.5 points of the pre-trained baseline. More aggressive updates, however, lead to substantial forgetting: CPT Phase 0+1 at $\mathrm{lr} = 10^{-4}$ loses 23 MMLU points. These results show that continued pre-training can substantially improve regional knowledge without sacrificing general capabilities, but only within a carefully calibrated optimization regime.

A transferable outcome of this work is therefore not a single learning-rate value or training recipe, but a methodology for selecting one. Our experiments show that the safe learning-rate range narrows as model scale increases. The learning rate selected at 8 B scale, $\mathrm{lr} = 10^{-5}$, does not transfer directly to 70 B and results in a 3% reduction in Choclo performance. In our setting with 15% replay data, scaling from 8 B to 70 B required an approximately tenfold reduction in learning rate.

A practical strategy is therefore to explore broadly at smaller scale, use the observed scaling trend to narrow the candidate range, and validate a small number of configurations at the target scale before committing the full compute budget. More broadly, larger and more capable base models appear increasingly sensitive to aggressive parameter updates, particularly on benchmarks where their performance is already near an effective ceiling.

We also find that CPT Phase 1 at $\mathrm{lr} = 10^{-6}$ preserves the base model's commonsense capabilities in a format-robust manner, producing identical MMLU and HellaSwag scores under prompted and unprompted evaluation. This contrasts with the measurable prompt sensitivity observed in supervised fine-tuned models and suggests that the selected CPT configuration preserves general representations rather than merely retaining performance under a particular evaluation format.

Beyond the model itself, this work releases a set of open resources for regional LLM development: the 296.5 B-token LatamGPT-Corpus with source traceability and a three-tier licensing framework; the data-collection, governance, filtering, and preprocessing pipeline; reproducible recipes for continual pre-training, supervised fine-tuning, safety alignment, and evaluation; and results on Trueque and Choclo across multiple prompting and few-shot configurations. We additionally document the institutional governance mechanisms used to construct the corpus, including memoranda of understanding, framework agreements, and jurisdiction-specific data-access conditions. Together, these resources provide a reusable foundation for regional AI initiatives operating under heterogeneous technical, institutional, and legal constraints.

Future work will extend LatamGPT along three complementary directions. First, we aim to expand support for Indigenous languages to better represent the linguistic diversity of Latin America and the Caribbean. Second, we plan to develop multimodal versions of LatamGPT that incorporate images, audio, and other forms of regionally relevant data. Third, beyond improving the model's regional knowledge, we aim to strengthen its reasoning capabilities through dedicated training and post-training methods. This direction is essential to narrowing the growing capability gap between open regional models and state-of-the-art proprietary systems, whose recent advances increasingly rely on enhanced reasoning abilities rather than knowledge coverage alone. The model weights, corpus, and training and evaluation code will be released publicly to support further research and accelerate the development of regionally grounded and technically competitive AI systems.

Acknowledgements

We thank our institutional partners and data alliance members across 20 Latin American countries for their contributions to the LatamGPT-Corpus. We gratefully acknowledge Anton Alexander, Laura Álvarez Modernel, Paulo Aragão, and Brenda Quilarque from Amazon Web Services for computational support to train on 128 H200 GPUs via SageMaker HyperPod. We also thank the Data Observatory Foundation, represented by Rodrigo Roa, for their infrastructure support and coordination throughout this project. We specially thank the CENIA operations team for their invaluable support in managing logistics and operational efforts, and the broader LatamGPT team for their contributions across data, modeling, and evaluation.

References

Joshua Ainslie, James Lee-Thorp, Michiel de Jong, Yury Zemlyanskiy, Federico Lebrón, and Sumit Sanghai. GQA: Training generalized multi-query transformer models from multi-head checkpoints. *arXiv preprint arXiv:2305.13245*, 2023. doi: 10.48550/arXiv.2305.13245.

Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Mérouane Debbah, Étienne Goffinet, Daniel Hesslow, Julien Launay, Quentin Malartic, Daniele Mazzotta, Badreddine Noune, Baptiste Pannier, and Guilherme Penedo. The Falcon series of open language models. *arXiv preprint arXiv:2311.16867*, 2023. doi: 10.48550/arXiv.2311.16867.

Stefan Baack. A critical analysis of the largest source for generative AI training data: Common Crawl. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 2199–2208, 2024. doi: 10.1145/3630106.3659033.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023. doi: 10.48550/arXiv.2309.16609.

BigScience Workshop, Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, et al. BLOOM: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*, 2022. doi: 10.48550/arXiv.2211.05100.

Faeze Brahman, Sachin Kumar, Vidhisha Balachandran, Pradeep Dasigi, Valentina Pyatkin, et al. The Art of Saying No: Contextual noncompliance in language models. *arXiv preprint arXiv:2407.12043*, 2024. doi: 10.48550/arXiv.2407.12043.

Centro Nacional de Inteligencia Artificial. Índice latinoamericano de inteligencia artificial (ILIA) 2024. Technical report, CENIA, 2024. URL https://indicelatam.cl/wp-content/uploads/2024/10/ILIA_2024.pdf.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? Try ARC, the AI2 reasoning challenge. *arXiv preprint arXiv:1803.05457*, 2018. doi: 10.48550/arXiv.1803.05457.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Édouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, 2020. doi: 10.18653/v1/2020.acl-main.747.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT 2019*, pages 4171–4186, 2019. doi: 10.18653/v1/N19-1423.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. doi: 10.48550/arXiv.2407.21783.

Omar U. Florez and LatamGPT Team. WayraPPL: High-performance perplexity estimation of data novelty, 2025. URL https://huggingface.co/latam-gpt/Wayra-Perplexity-Estimator-55M.

Chaochen Gao, Xing Wu, Zijia Lin, Debing Zhang, and Songlin Hu. LongMagpie: A self-synthesis method for generating large-scale long-context instructions. *arXiv preprint arXiv:2505.17134*, 2025. doi: 10.48550/arXiv.2505.17134.

Tao Ge, Xin Chan, Xiaoyang Wang, Dian Yu, Haitao Mi, and Dong Yu. Scaling synthetic data creation with 1,000,000,000 personas. *arXiv preprint arXiv:2406.20094*, 2024. doi: 10.48550/arXiv.2406.20094.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, et al. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025. doi: 10.48550/arXiv.2503.19786.

Dirk Groeneveld, Iz Beltagy, Pete Walsh, Akshita Bhagia, Rodney Kinney, Oyvind Tafjord, Ananya Jha, Hamish Ivison, Ian Magnusson, Yizhong Wang, Shruti Arora, et al. OLMo: Accelerating the science of language models. *arXiv preprint arXiv:2402.00838*, 2024. doi: 10.48550/arXiv.2402.00838.

Seungju Han, Kavel Rao, Allyson Ettinger, Liwei Jiang, Bill Yuchen Lin, Nathan Lambert, Yejin Choi, and Nouha Dziri. WildGuard: Open one-stop moderation tools for safety risks, jailbreaks, and refusals of LLMs. *arXiv preprint arXiv:2406.18495*, 2024. doi: 10.48550/arXiv.2406.18495.

Eric Hartford and Cognitive Computations. Dolphin 2.9.1 Llama 3 70B. https://huggingface.co/cognitivecomputations/dolphin-2.9.1-llama-3-70b, 2024.

Thomas Hartvigsen, Saadia Gabriel, Hamid Palangi, Maarten Sap, Dipankar Ray, and Ece Kamar. ToxiGen: A large-scale machine-generated dataset for adversarial and implicit hate speech detection. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, pages 3309–3326, 2022. doi: 10.18653/v1/2022.acl-long.234.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*, 2020. doi: 10.48550/arXiv.2009.03300.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals, and Laurent Sifre. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022. doi: 10.48550/arXiv.2203.15556.

Yue Huang, Lichao Sun, Haoran Wang, Siyuan Wu, Qihui Zhang, Yuan Li, Chujie Gao, Yixin Huang, Wenhan Lyu, Yixuan Zhang, et al. TrustLLM: Trustworthiness in large language models. *arXiv preprint arXiv:2401.05561*, 2024. doi: 10.48550/arXiv.2401.05561.

Adam Ibrahim, Benjamin Thérien, Kshitij Gupta, Mats L. Richter, Quentin Anthony, Eugene Belilovsky, Irina Rish, and Timothée Lesort. Simple and scalable strategies to continually pre-train large language models. *arXiv preprint arXiv:2403.08763*, 2024. doi: 10.48550/arXiv.2403.08763.

Liwei Jiang, Kavel Rao, Seungju Han, Allyson Ettinger, Faeze Brahman, Sachin Kumar, Niloofar Mireshghallah, Ximing Lu, Maarten Sap, Yejin Choi, and Nouha Dziri. WildTeaming at scale: From in-the-wild jailbreaks to (adversarially) safer language models. *arXiv preprint arXiv:2406.18510*, 2024. doi: 10.48550/arXiv.2406.18510.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526, 2017. doi: 10.1073/pnas.1611835114.

Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James V. Miranda, Alisa Liu, Nouha Dziri, et al. Tülu 3: Pushing frontiers in open language model post-training. *arXiv preprint arXiv:2411.15124*, 2024. doi: 10.48550/arXiv.2411.15124.

LatamGPT Team. Trueque: A spanish factual knowledge benchmark for latin american language models. Hugging Face Datasets, 2025. URL https://huggingface.co/datasets/latam-gpt/trueque.

Hugo Laurençon, Lucile Saulnier, Thomas Wang, Christopher Akiki, Albert Villanova del Moral, Teven Le Scao, Leandro Von Werra, Chenghao Mou, et al. The BigScience ROOTS corpus: A 1.6TB composite multilingual dataset. In *Advances in Neural Information Processing Systems*, volume 35, pages 31809–31826, 2022. doi: 10.48550/arXiv.2303.03915.

Stephanie Lin, Jacob Hilton, and Owain Evans. TruthfulQA: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, pages 3214–3252, 2022. doi: 10.18653/v1/2022.acl-long.229.

Yuhang Luo, Fenglong Yang, and Xiaojun Wan. An empirical study of catastrophic forgetting in large language models during continual fine-tuning. *arXiv preprint arXiv:2308.08747*, 2023. doi: 10.48550/arXiv.2308.08747.

Max Marion, Ahmet Üstün, Luiza Pozzobon, Alex Wang, Marzieh Fadaee, and Sara Hooker. When less is more: Investigating data pruning for pretraining LLMs at scale. In *Proceedings of the 41st International Conference on Machine Learning*, 2024. URL https://arxiv.org/abs/2309.04564.

Mantas Mazeika, Long Phan, Xuwang Yin, Andy Zou, Zifan Wang, Norman Mu, Elham Sakthivadivel, Nathaniel Li, and Dan Hendrycks. HarmBench: A standardized evaluation framework for automated red teaming and robust refusal. *arXiv preprint arXiv:2402.04249*, 2024. doi: 10.48550/arXiv.2402.04249.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. Can a suit of armor conduct electricity? a new dataset for open book question answering. In *Proceedings of EMNLP 2018*, pages 2381–2391, 2018. doi: 10.18653/v1/D18-1260.

Moran Mizrahi, Guy Kaplan, Dan Malkin, Rotem Dror, Dafna Shahaf, and Gabriel Stanovsky. State of what art? a call for multi-prompt LLM evaluation. *Transactions of the Association for Computational Linguistics*, 12:933–949, 2024. doi: 10.1162/tacl_a_00681.

Raymond Ng, Thanh Ngan Nguyen, Yuli Huang, Ngee Chia Tai, Wai Yi Leong, Wei Qi Leong, Xianbin Yong, Jian Gang Ngui, Yosephine Susanto, Nicholas Cheng, et al. SEA-LION: Southeast Asian languages in one network. *arXiv preprint arXiv:2504.05747*, 2025. doi: 10.48550/arXiv.2504.05747.

Thuat Nguyen, Chien Van Nguyen, Viet Dac Lai, Hieu Man, Nghia Trung Ngo, Franck Dernoncourt, Ryan A. Rossi, and Thien Huu Nguyen. CulturaX: A cleaned, enormous, and multilingual dataset for large language models in 167 languages. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, 2024. URL https://doi.org/10.48550/arXiv.2309.09400.

NVIDIA. Llama-3.1-Nemotron-Ultra-253B: A technical report. *arXiv preprint arXiv:2505.00315*, 2025a. doi: 10.48550/arXiv.2505.00315.

NVIDIA. Nemotron-H: A family of accurate and efficient hybrid Mamba-Transformer models. *arXiv preprint arXiv:2504.03624*, 2025b. doi: 10.48550/arXiv.2504.03624.

NVIDIA, Bo Adler, Niket Agarwal, Ashwath Aithal, et al. Nemotron-4 340b technical report. *arXiv preprint arXiv:2406.11704*, 2024. doi: 10.48550/arXiv.2406.11704.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. MedMCQA: A large-scale multi-subject multi-choice dataset for medical domain question answering. *arXiv preprint arXiv:2203.14371*, 2022. doi: 10.48550/arXiv.2203.14371.

Guilherme Penedo, Hynek Kydlíček, Loubna Ben Allal, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro von Werra, and Thomas Wolf. The FineWeb datasets: Decanting the web for the finest text data at scale. *arXiv preprint arXiv:2406.17557*, 2024. doi: 10.48550/arXiv.2406.17557.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. In *OpenAI technical report*, 2019.

Paul Röttger, Hannah Rose Kirk, Bertie Vidgen, Giuseppe Attanasio, Kalina Bontcheva, and Dirk Hovy. XSTest: A test suite for identifying exaggerated safety behaviours in large language models. *arXiv preprint arXiv:2308.01263*, 2023. doi: 10.48550/arXiv.2308.01263.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. WinoGrande: An adversarial winograd schema challenge at scale. *Communications of the ACM*, 64(9):99–106, 2021. doi: 10.1145/3474381.

Luca Schirru, Ana Rocha de Souza, Miguel G. Valente, and André de Perdigão Lana. Text and data mining exceptions in Latin America. *IIC – International Review of Intellectual Property and Competition Law*, 55:1624–1653, 2024. doi: 10.1007/s40319-024-01511-2.

Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. Quantifying language models' sensitivity to spurious features in prompt design. In *International Conference on Learning Representations*, 2024.

Xinyue Shen, Zeyuan Chen, Michael Backes, Yun Shen, and Yang Zhang. "do anything now": Characterizing and evaluating in-the-wild jailbreak prompts on large language models. *arXiv preprint arXiv:2308.03825*, 2023. doi: 10.48550/arXiv.2308.03825.

Álvaro Soto, Rodrigo Durán, Antonia Moreno, Sebastián Adasme, Sebastián Rovira, Valeria Jordán, and Laura Poveda. Índice latinoamericano de inteligencia artificial (ILIA) 2025: Hallazgos principales, IA aplicada y talento humano. Technical report, CEPAL and CENIA, 2026. URL https://indicelatam.cl/.

Alexandra Souly, Qingyuan Lu, Dillon Bowen, Tu Trinh, Elvis Hsieh, Sana Pandey, Pieter Abbeel, Justin Svegliato, Scott Emmons, Olivia Watkins, and Sam Toyer. A StrongREJECT for empty jailbreaks. *arXiv preprint arXiv:2402.10260*, 2024. doi: 10.48550/arXiv.2402.10260.

Jianlin Su, Yu Lu, Shengfeng Pan, Ahmed Murtadha, Bo Wen, and Yunfeng Liu. RoFormer: Enhanced transformer with rotary position embedding. *arXiv preprint arXiv:2104.09864*, 2021. doi: 10.48550/arXiv.2104.09864.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. CommonsenseQA: A question answering challenge targeting commonsense knowledge. In *Proceedings of NAACL-HLT 2019*, pages 4149–4158, 2019. doi: 10.18653/v1/N19-1421.

Team Olmo, Allyson Ettinger, Amanda Bertsch, Bailey Kuehl, David Graham, David Heineman, Dirk Groeneveld, Faeze Brahman, Finbarr Timbers, Hamish Ivison, et al. Olmo 3. *arXiv preprint arXiv:2512.13961*, 2025. doi: 10.48550/arXiv.2512.13961.

Johannes Welbl, Nelson F. Liu, and Matt Gardner. Crowdsourcing multiple choice science questions. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*, pages 94–106, 2017. doi: 10.18653/v1/W17-4413.

Zhangchen Xu, Fengqing Jiang, Luyao Niu, Yuntian Deng, Radha Poovendran, Yejin Choi, and Bill Yuchen Lin. Magpie: Alignment data synthesis from scratch by prompting aligned LLMs with nothing. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2406.08464.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. mT5: A massively multilingual pre-trained text-to-text transformer. *arXiv preprint arXiv:2010.11934*, 2021. doi: 10.48550/arXiv.2010.11934.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, et al. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024. doi: 10.48550/arXiv.2412.15115.

An Yang et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. doi: 10.48550/arXiv.2505.09388.

Greg Yang, Edward J. Hu, Igor Babuschkin, Szymon Sidor, Xiaodong Liu, David Farhi, Nick Ryder, Jakub Pachocki, Weizhu Chen, and Jianfeng Gao. Tensor programs V: Tuning large neural networks via zero-shot hyperparameter transfer. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. HellaSwag: Can a machine really finish your sentence? *arXiv preprint arXiv:1905.07830*, 2019. doi: 10.48550/arXiv.1905.07830.

Wenting Zhao, Xiang Ren, Jack Hessel, Claire Cardie, Yejin Choi, and Yuntian Deng. WildChat: 1M ChatGPT interaction logs in the wild. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2405.01470.