

The background of the slide is a dark, atmospheric scene. A person is standing on a highly reflective surface, possibly a wet floor or a polished metal surface. The person is silhouetted against the bright light trails. The light trails are numerous, curved, and glow in shades of blue and cyan, creating a sense of motion and digital energy. The overall mood is futuristic and technological.

Latam GPT: Training the First Foundational Model from Latinoamérica

Open Source **Regional Llms**



LatamGPT

Latam-GPT

La Revolución Mexicana (1910-1920) fue un proceso de profundo cambio en la estructura social y política de México, que buscó reemplazar al régimen porfirista, caracterizado por la concentración de poder y riqueza en pocas manos, y la explotación de la población indígena y campesina.

- **Contexto histórico:** El Porfiriato, que duró de 1876 a 1911, vio un crecimiento económico, pero a costa de un régimen autoritario y desigual.
- **Causas:** La desigualdad social, la falta de justicia y el descontento con el gobierno de Porfirio Díaz fueron el caldo de cultivo para el estallido revolucionario.
- **Liderazgo:** Francisco I. Madero, Emiliano Zapata, Pancho Villa y Venustiano Carranza fueron figuras clave en la lucha.
- **Reformas:** La Revolución Mexicana llevó a la promulgación de la Constitución de 1917, que estableció derechos laborales, la reforma agraria y la separación de la Iglesia y el Estado.
- **Consecuencias:** Aunque el conflicto armado terminó, el legado de la Revolución se refleja en la política y la identidad nacional mexicana, con un énfasis en la justicia social y la democracia.
- **Puntos clave:**
 1. **Cambio de régimen:** Fin del Porfiriato y comienzo de gobiernos más democráticos.
 2. **Reformas sociales:** La Constitución de 1917 y leyes laborales.
 3. **Liderazgo revolucionario:** Madero, Villa, Zapata, y Carranza.
 4. **Constitución de 1917:** Un hito en el constitucionalismo social.
 5. **Legado:** La Revolución Mexicana es un símbolo de lucha por la justicia y la democracia en México.

ChatGPT

ChatGPT 5.4 Thinking

¿Cuál fue la importancia de la Revolución Mexicana en la transformación política y social de México?

Thought for a couple of seconds

La Revolución Mexicana fue fundamental porque cambió tanto la política como la estructura social de México.

En lo político, puso fin al Porfiriato, un régimen largo y autoritario, y abrió el camino para construir un nuevo Estado. Su resultado más importante fue la Constitución de 1917, que estableció nuevas bases legales para el país e incorporó derechos sociales que antes no existían de forma clara.

En lo social, buscó responder a las grandes desigualdades que vivían campesinos, obreros y sectores populares. Impulsó el reparto de tierras, fortaleció los derechos laborales y promovió una mayor intervención del Estado en educación y justicia social.

En conjunto, la Revolución Mexicana fue importante porque no solo derrocó una dictadura, sino que transformó a México en un país con nuevas instituciones, más derechos para la población y una idea distinta de nación y ciudadanía.

AI Stack



Application layer

Foundational model training

Architecture research

GPU design and development

AI Stack



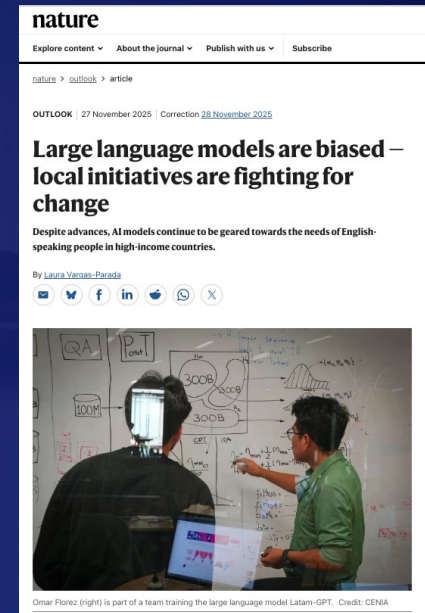
- Application layer
- Foundational model training**
- Architecture research
- GPU design and development

LatamGPT 

LatamGPT

- **First foundational model trained in Latin America**
- **Budget:**
 - Size: 70B params
 - Data: 300B tokens
 - FLOPs: 126M FLOPs (128/256 H200 GPUs)

- **Dataset** based on regional sources
Universities, national libraries, congresses, many of which have not been crawled yet
- **Open source code, training dataset, and benchmarks**
- **Stages:**
 - Pre-Training (PT)
 - Supervised Fine Tuning (SFT)
 - Alignment (DPO)





Team



Source: "LatamGPT Team" (2026)

LatamGPT

Strategic Collaborating Entities

50+ Partner Institutions in Latin America



Source: "latamgpt.org" (2026)

Outline

SCALING LAWS

- Kaplan vs Chinchilla laws
- Predict model performance in downstream tasks

CONTINUED PRE-TRAINING

- Reuse, do not re-train
- Use Perplexity for Curriculum Learning
- Forget VS. Adaptation

COMPUTE

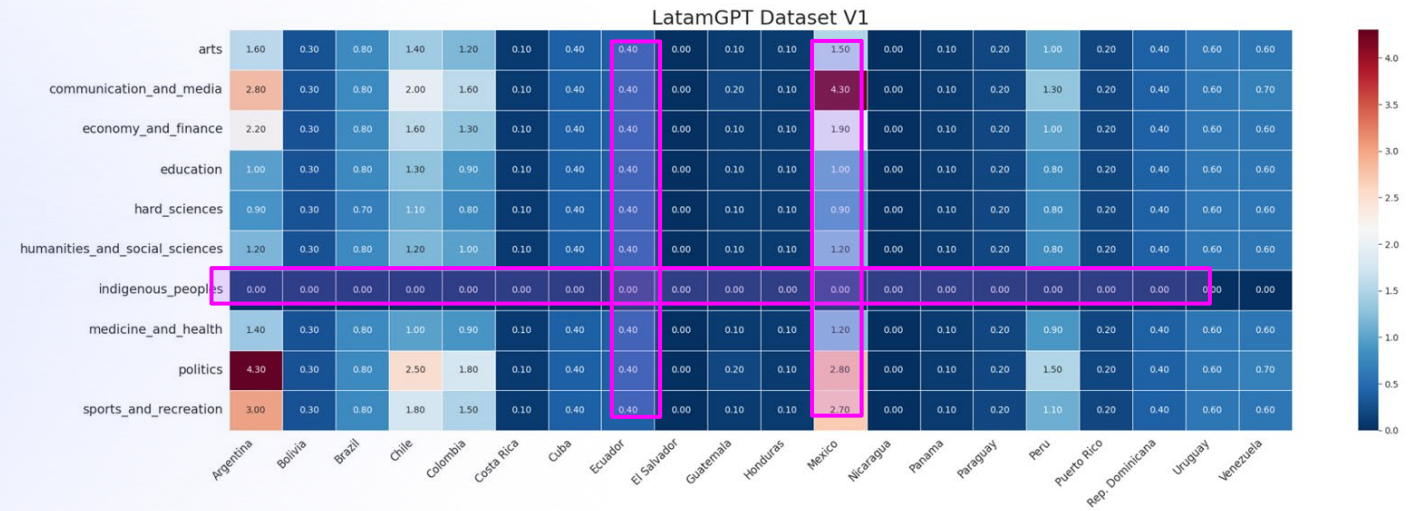
- Data, Tensor, Pipeline parallelism
- Multi Head, Grouped Query, Flash Attention
- Throughput optimization

DATA

- Generate Data Mixture
- Synthetic dataset generation

Data

LatamGPT pre-training dataset V1



Common Crawl (ES, PT, EN)

Zipf distribution of web text

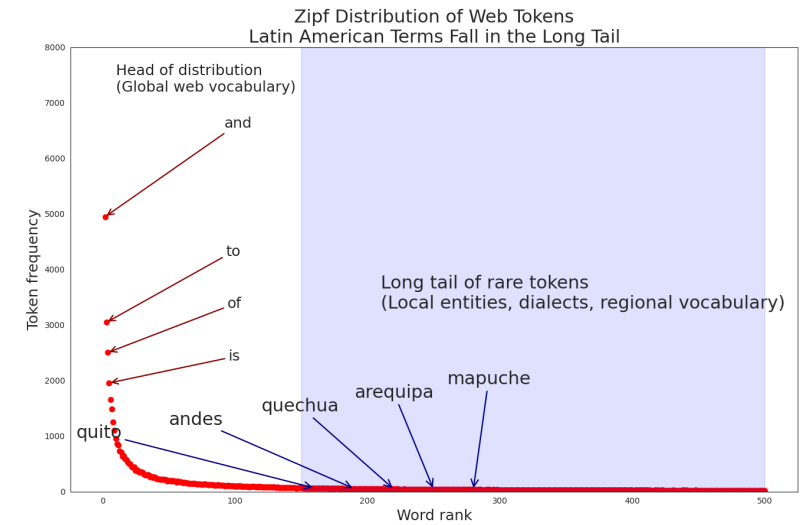
Natural language follows Zipf's law. A small number of tokens dominate the corpus while most tokens appear in the long tail

Regional vocabulary lives in the tail

Names, entities, and dialect terms from Latin America appear with very low frequency in global web crawls like Common Crawl

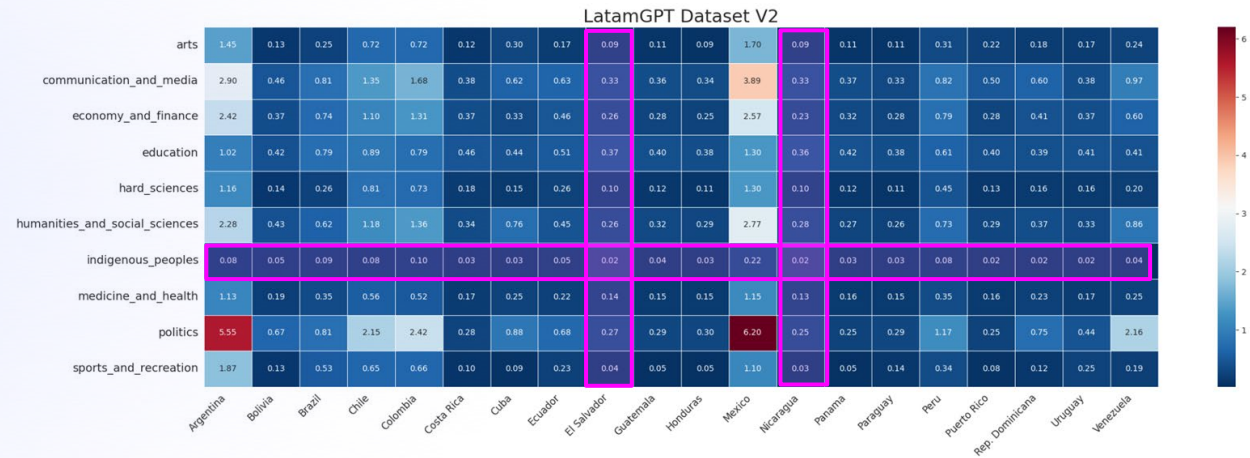
Impact on model training

Rare tokens provide little training signal, so regional knowledge is underrepresented unless the dataset is intentionally rebalanced (e.g., LatamGPT).



LatamGPT pre-training dataset V1

Include focused data collection and synthetic data generation



Scaling Laws

Neural Scaling Laws

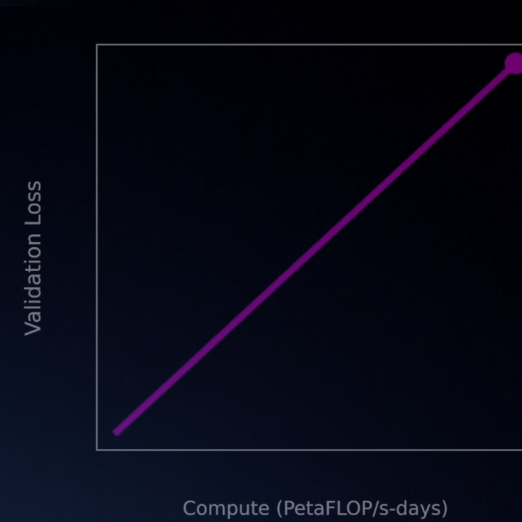
- **Validation Loss vs Compute**
 - Performance vs Compute plot
 - Bitter lesson
 - **Empirical Risk Minimization:** increase in compute, data, parameters **decreases validation loss**
 - **Learning** doesn't scale infinitely, but it **does predictably**



Source: AI can't cross this line and we don't know why, Welch Labs (2025).

Neural Scaling Laws

- **Validation Loss vs Compute**
 - Performance vs Compute plot
 - Bitter lesson
 - **Empirical Risk Minimization:** increase in compute, data, parameters **decreases validation loss**
 - **Learning** doesn't scale infinitely, but it **does predictably**

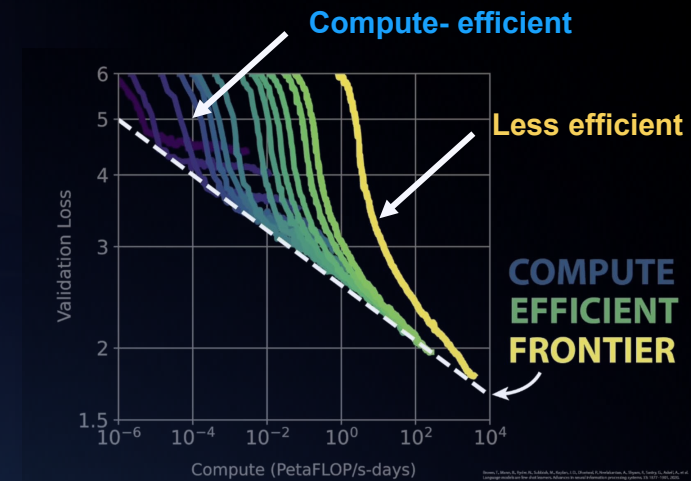


Source: AI can't cross this line and we don't know why, Welch Labs (2025).

Neural Scaling Laws

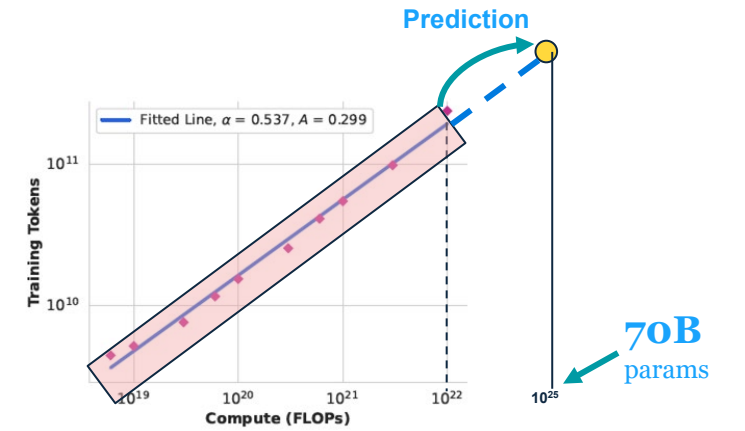
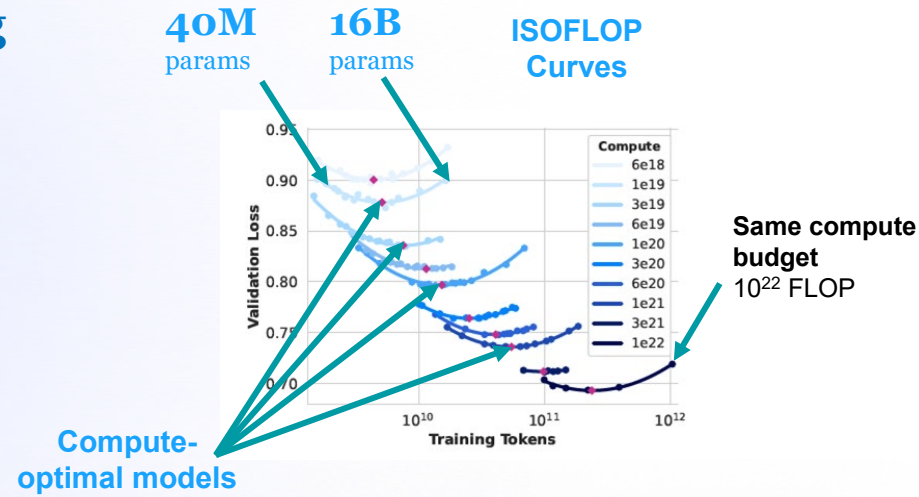
Compute-Efficient Frontier

- White line represents the best achievable **validation loss** (for a given amount of compute)



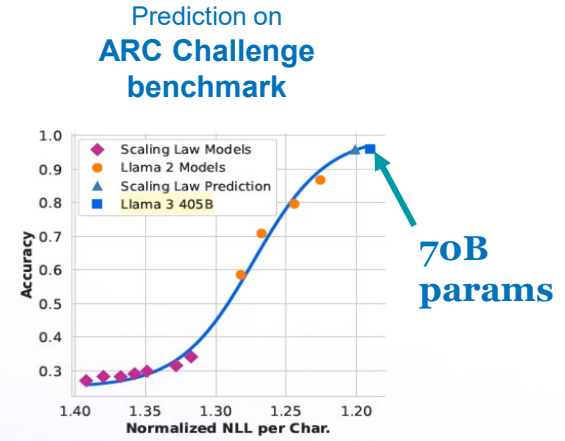
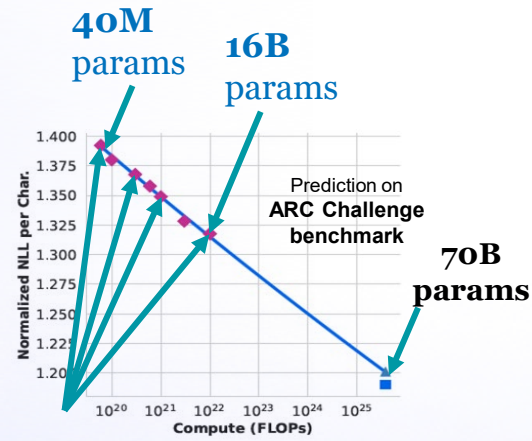
Source: AI can't cross this line and we don't know why, Welch Labs (2025).

Predict size of training dataset (# Tokens)



Predict model's accuracy on Downstream Task

Compute-optimal models



Pre-training

Pre-Training (from Scratch)

- **Compute Efficiency:**

- Chinchilla-based Compute (C) budget:
 Predicted FLOPs ~ $6 \times (\text{model size}) \times (\text{dataset size})$
 840 Million PFLOPs ~ $6 \times (70\text{B parameters}) \times (2 \text{ Trillion tokens})$

- **Time estimate:**

$$39 \text{ days} = \frac{840 \times 10^6 \times \text{PFLOPs}}{32 \text{ instances} \times 8 \text{ H200 GPUs} \times 1.979 \text{ PFLOPs@fp32} \times 0.5 \text{ MFU} \times 86400 \text{ secs}}$$

$$840\text{M PFLOPs} / 128 \text{ H200 GPUs} = 39 \text{ days}$$




Continued Pre-Training (CPT)

- **Compute Efficiency:**

- Chinchilla-based Compute (C) budget:
 Predicted FLOPs ~ $6 \times (\text{model size}) \times (\text{dataset size})$
 126 Million PFLOPs ~ $6 \times (70\text{B parameters}) \times (300 \text{ Billion tokens})$

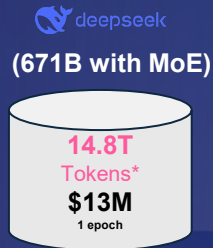


- **Time estimate:**

$$5.78 \text{ days} = \frac{126 \times 10^6 \times \text{PFLOPs}}{32 \text{ instances} \times 8 \text{ H200 GPUs} \times 1.979 \text{ PFLOPs@fp32} \times 0.5 \text{ MFU} \times 86400 \text{ secs}}$$

$$\frac{126\text{M PFLOPs}}{128 \text{ H200 GPUs}} = 6 \text{ days}$$


Comparison



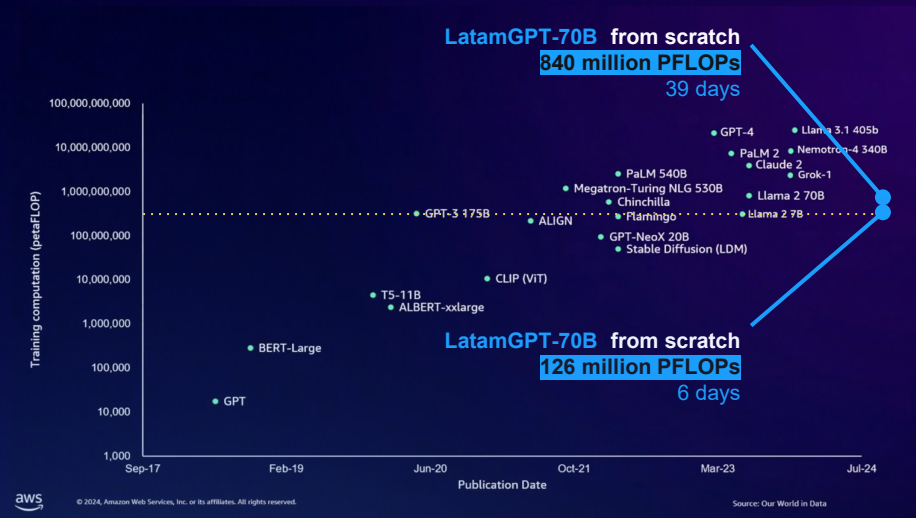
- 3.3 ZetaFLOPs
- 14 months
- 256 GPUs NVIDIA H200 GPUs



- 840 Million PFLOPs
- Build general-purpose LM with random initialization
- Data strategy: Broad corpus
- LR strategy: **Standard decays with warmup**
- Data distribution similar to the pre-training dataset prevents catastrophic forgetting
- Prioritize **non-web content** and **high-quality web data**



- 126 Million PFLOPs
- Continued Pre-Training
- Enhance accuracy on specific new data domains
- **General Blend -> QA Blend**
- High initial LR with cosine decay, no warmup
- **9%-16% Avg. Accuracy on tasks like MMLU, HellaSwag**





Continued Pre-training

Continued Pre-Training (CPT)

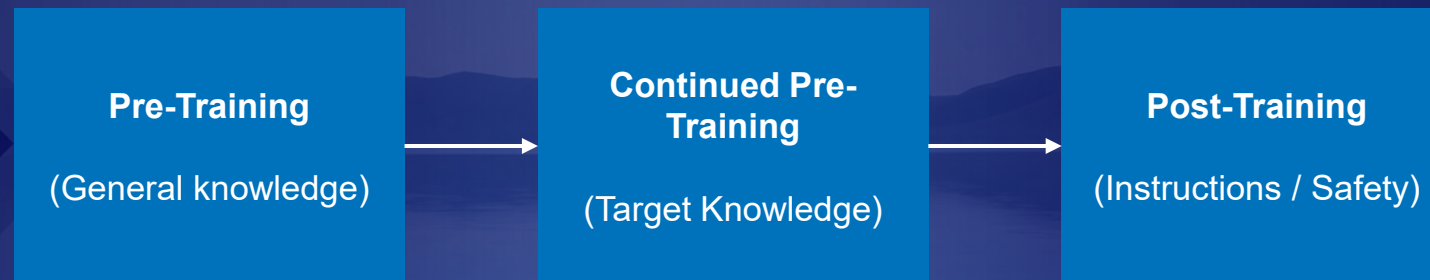
- Rather than training from scratch, start from a pretrained checkpoint
 - Example: LLaMA-70B (15.6T tokens)
 - CPT on new domain data

Continue optimization on a new token distribution to improve specific capabilities or update knowledge.

CPT can be seen as a transition between training regimes with different token distributions and learning rates.

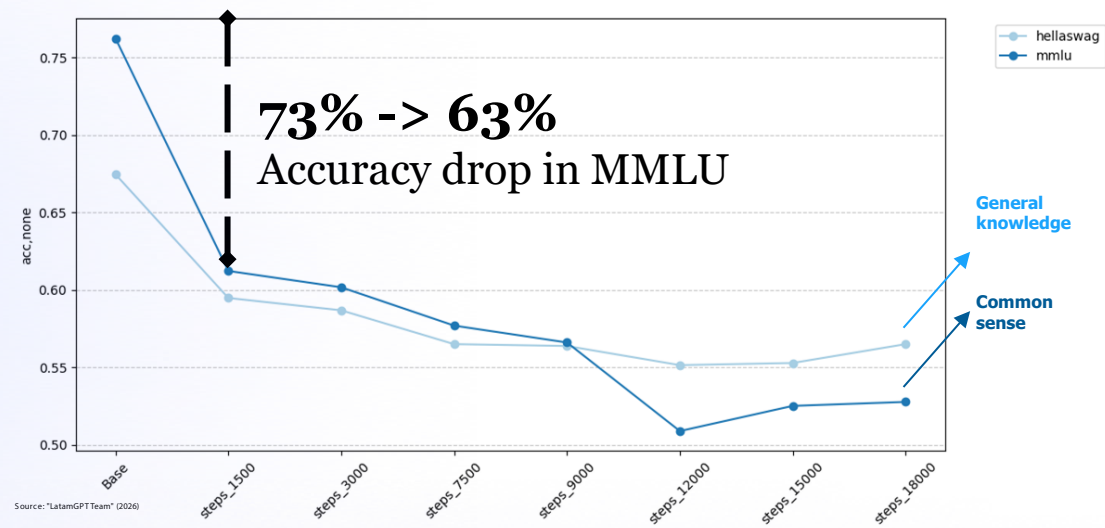


Continued Pre-Training (CPT)



What could go wrong?
**Catastrophic
Forgetting**

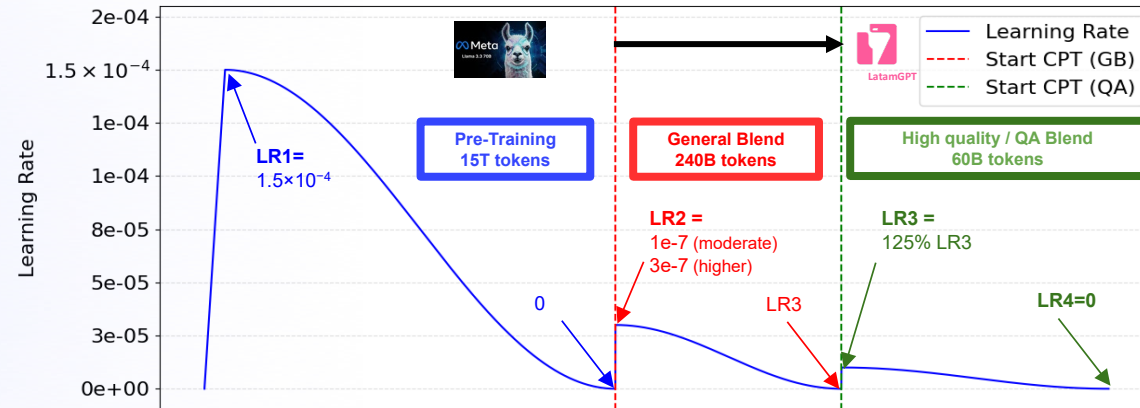
Accuracy on MMLU and Hellaswag



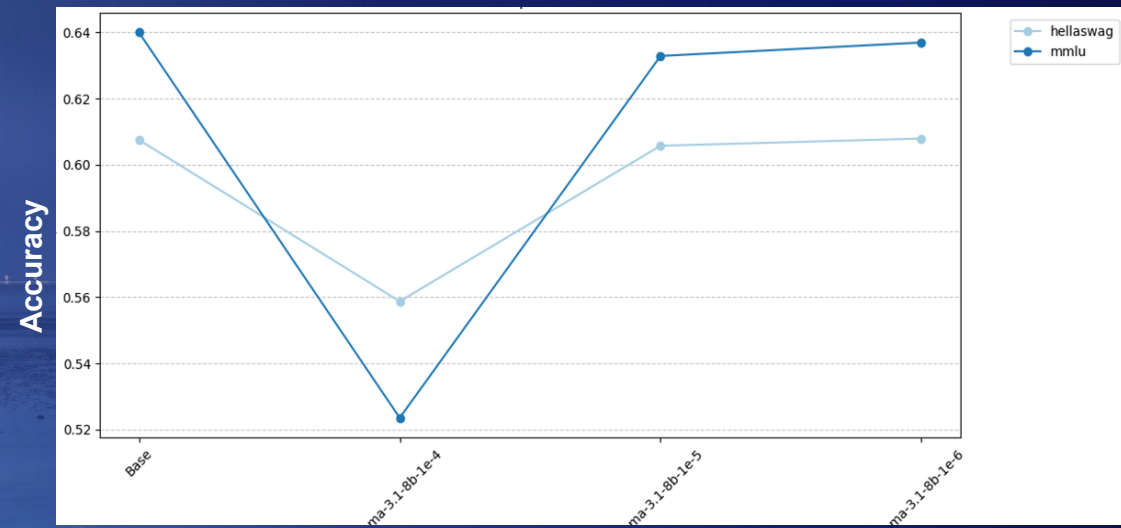
Forget Vs Adaptation

LR scheduling by steps

Cosine LR schedule $\eta(t) = \eta_{init} \times \frac{1 + \cos(\pi \cdot \frac{t}{T})}{2}$



Accuracy on MMLU and Hellaswag

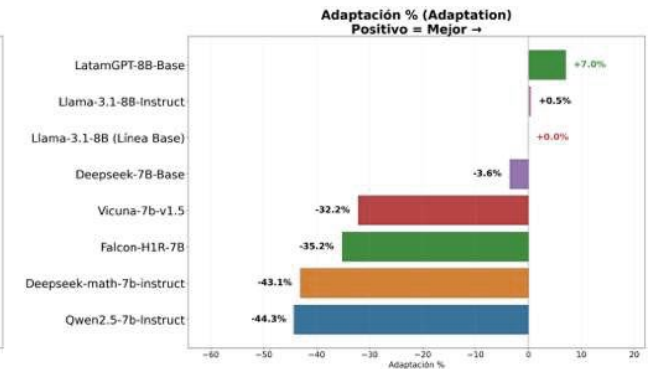
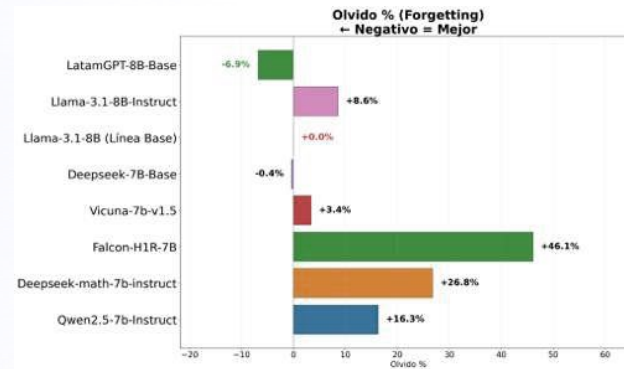


Source: "LatamGPT Team" (2026)

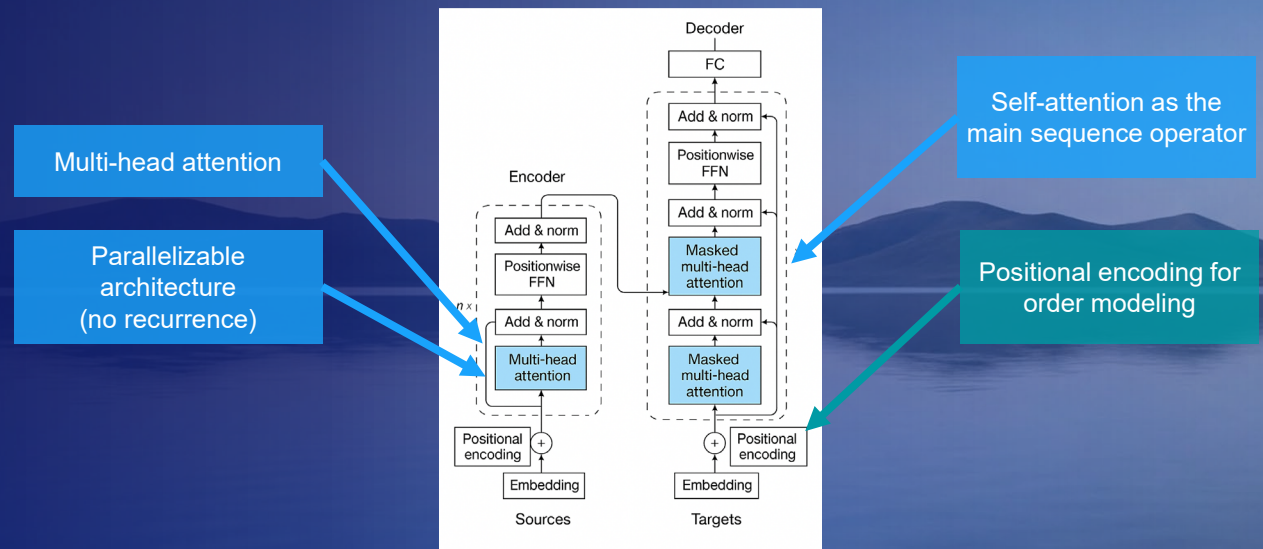
Forget vs. Adaptation

$$\text{Forgetting \%} = \frac{\text{PPL}_{\text{MMLU}}^{\text{checkpoint}} - \text{PPL}_{\text{MMLU}}^{\text{baseline}}}{\text{PPL}_{\text{MMLU}}^{\text{baseline}}} \times 100$$

$$\text{Adaptation \%} = \frac{\text{PPL}_{\text{Choco}}^{\text{baseline}} - \text{PPL}_{\text{Choco}}^{\text{checkpoint}}}{\text{PPL}_{\text{Choco}}^{\text{baseline}}} \times 100$$

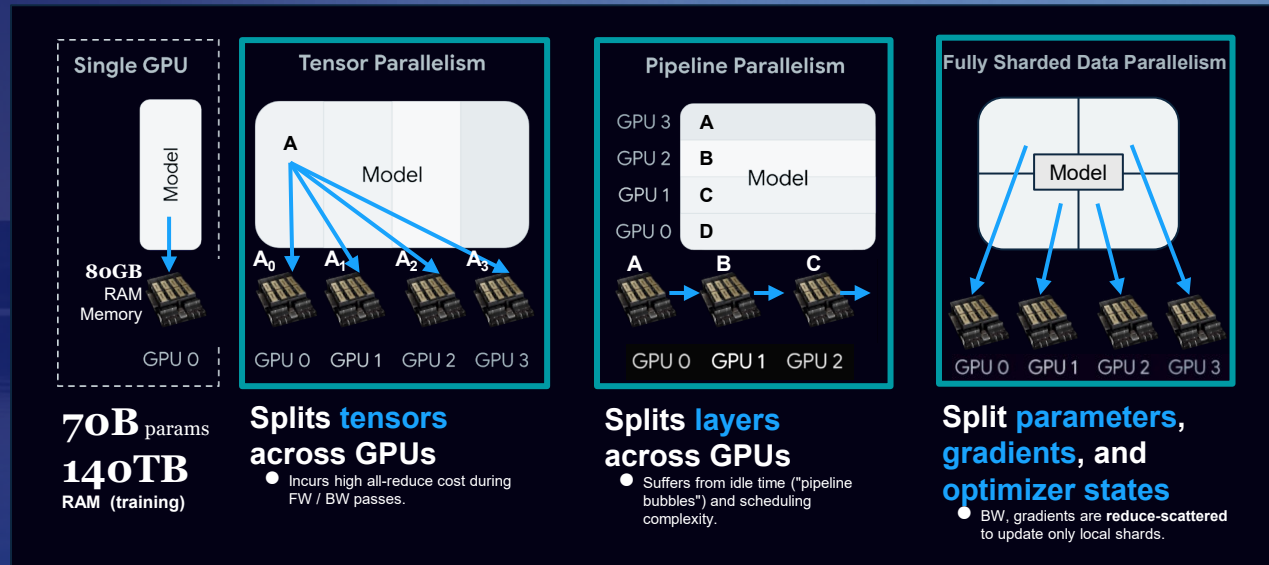


Compute



Parallelization

Parallel Computing (LatamGPT 70B)



70B params
140TB RAM (training)

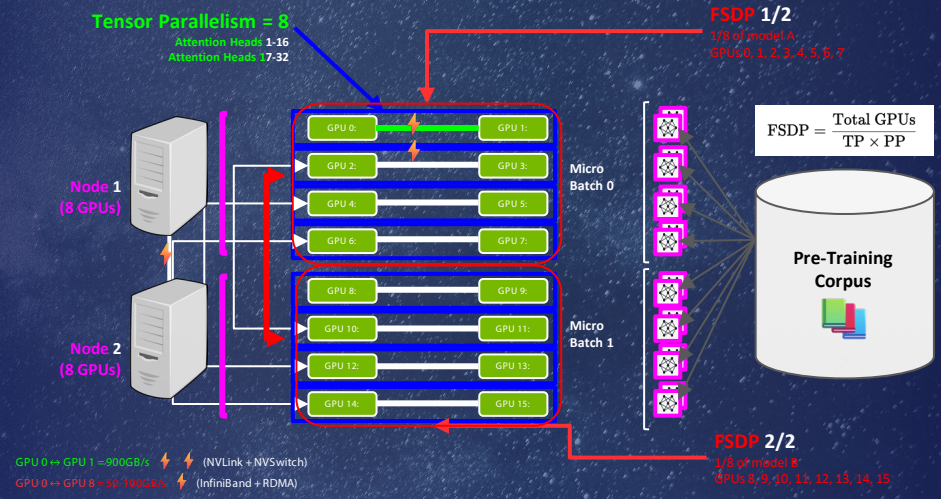
Splits tensors across GPUs
 • Incurs high all-reduce cost during FW / BW passes.

Splits layers across GPUs
 • Suffers from idle time ("pipeline bubbles") and scheduling complexity.

Split parameters, gradients, and optimizer states
 • BW, gradients are **reduce-scattered** to update only local shards.

Parallelism





VIDEO

LatamGPT

LatamGPT To peruano, junto con organizaciones internacionales, ha implementado medidas para proteger el sitio y su entorno, asegurando su legado para las generaciones futuras.

7. Investigación y Descubrimientos

La investigación arqueológica en Machu Picchu continúa, y se han descubierto nuevos hallazgos que aportan a la comprensión de la civilización inca. Por ejemplo, se han identificado tumbas, ofrendas y restos humanos, lo que ha enriquecido el conocimiento sobre la vida y la muerte en la época inca.

8. Cultura Viva

A pesar de su antigüedad, Machu Picchu sigue siendo un lugar vivo. Las comunidades andinas, como los pueblos de Aguas Calientes y Ollantaytambo, mantienen tradiciones, festivales y prácticas que se remontan a la época inca, lo que hace que el legado de Machu Picchu sea una parte integral de la identidad cultural andina.

Información que la gente quizás no conozca:

- Machu Picchu no es un "descubrimiento"; los pueblos andinos ya conocían y utilizaban el sitio.
- Hiram Bingham no "descubrió" el sitio, sino que lo divulgó a nivel internacional en 1911.
- Machu Picchu es Patrimonio de la Humanidad desde 1983 y una de las Siete Maravillas del Mundo Moderno.
- El sitio es un centro de investigación y se continúan descubriendo nuevos aspectos de la civilización inca.
- Machu Picchu es un lugar vivo con tradiciones, festivales y comunidades andinas que mantienen su legado.

Send a Message

localhost:3000

ChatGPT

ChatGPT

...mas representativos de la civilización andina. Su relevancia es histórica, cultural, arquitectónica, espiritual y contemporánea.

1. Importancia histórica

- Fue construida en el siglo XV durante el gobierno del emperador inca Pachacútec.
- Representa el máximo desarrollo político, administrativo y arquitectónico del Imperio Inca.
- Es una evidencia tangible del alto nivel de organización, ingeniería y planificación urbana alcanzado por los incas.
- Fue dada a conocer internacionalmente en 1911 por el explorador Hiram Bingham.

2. Importancia arquitectónica y tecnológica

- Construida con piedra tallada sin uso de mortero (técnica de encaje perfecto).
- Adaptada magistralmente a la geografía montañosa andina.
- Incluye terrazas agrícolas, templos, plazas y complejos residenciales.
- Demuestra avanzados conocimientos en ingeniería hidráulica, antisísmica y agrícola.

3. Importancia cultural y espiritual

- Se cree que fue un centro ceremonial, político y posiblemente una residencia real.
- Refleja la cosmovisión andina: la armonía entre el ser humano, la naturaleza y lo sagrado.
- Está vinculada al culto al Sol, representado en estructuras como el Templo del

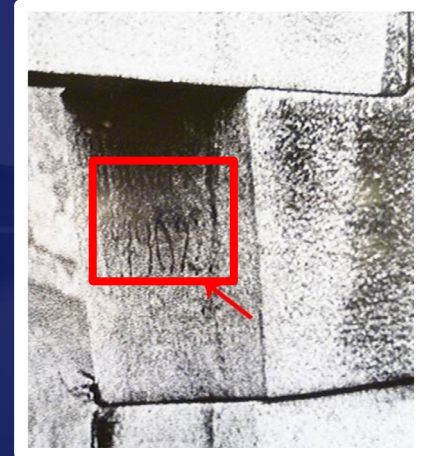
Ask anything

ChatGPT can make mistakes. Check important info.

1
9
1
1



1
9
0
2





Benchmarks

<https://huggingface.co/datasets/latam-gpt/CHOCLO>

Choclo (Wikidata)

- Build knowledge graph from Wikidata based on popular Entities related to Latin America
- 200K questions and answers**

The screenshot shows the Hugging Face dataset card for 'latam-gpt/CHOCLO'. It includes a search bar, navigation tabs for 'Models' and 'Datasets', and a list of dataset details: 'Tasks: Question Answering', 'Modalities: Text', 'Formats: csv', 'Sub-tasks: open-domain-qa', 'Size: 100K - 1M', 'Tags: culture, latin-america, benchmark, knowledge', 'Libraries: Datasets', and 'License: mit'. At the bottom, there are links for 'Dataset card', 'Data Studio', 'Files', 'Community', and 'Settings'.

copuchat.latamgpt.org

Trueque (Human)

- Collect questions and answers based on a live chat from Latin American users during 6 Months (Copuchat)
- Clean and filter each Question and Answer
- 2,000 Q&A collected**

The screenshot shows the Copuchat chat interface. It features a header with the Copuchat logo and 'PAG #33014 - Listo'. Below the header is a 'Chat' section with a 'Nueva Conversación' button and a search icon. The main area is a dark chat window with a text input field containing the placeholder 'Escribe tu mensaje y presiona Enter...' and a red 'Enviar' button at the bottom.

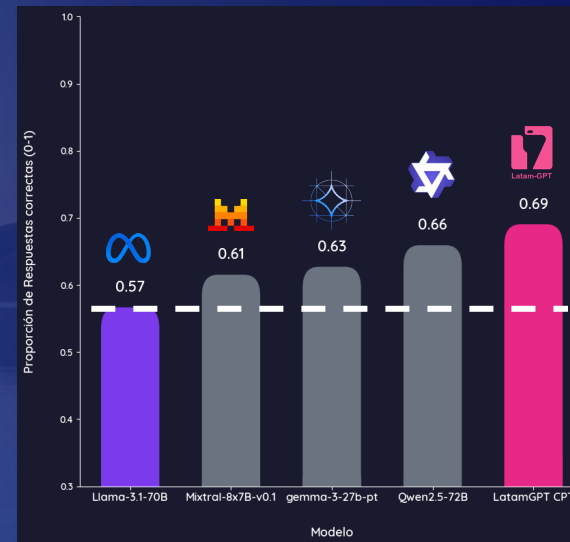
Source: Alexandra Garcia & LatamGPT team (CENIA, 2026)

Benchmark Truque



Source: Alexandra Garcia & LatamGPT team (CENIA, 2026)

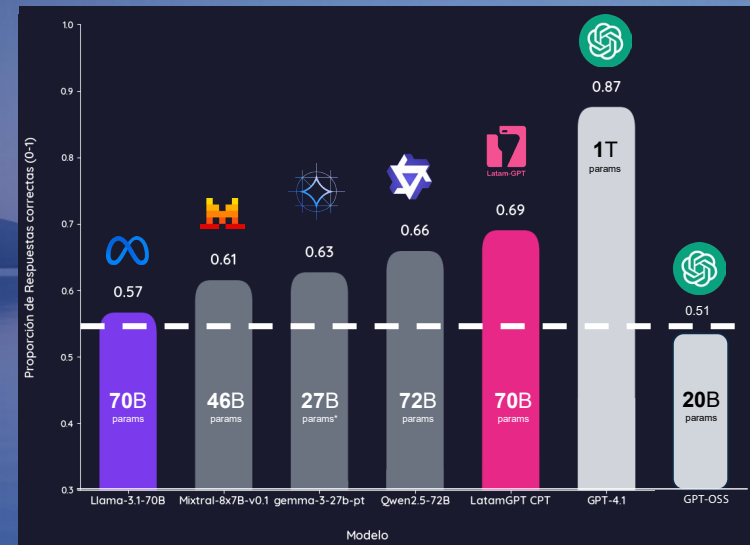
Truque Benchmark



Source: Alexandra Garcia & LatamGPT team (CENIA, 2026)

Latam-GPT Benchmark

* Largest model available in its family (e.g., Gemma 3)



Source: Alexandra Garcia & LatamGPT team (CENIA, 2026)

Future Work



Multimodal LatamGPT

Prompt on ChatGPT:

“Generate four realistic images of women from Peru, Chile, Argentina, and Mexico.”

(Generated on February, 2026)



Latin America in 2030

- By 2030, AI will write most code, contracts, and do most knowledge work
- Whoever owns the models captures the productivity surplus, the rest of the world rents it

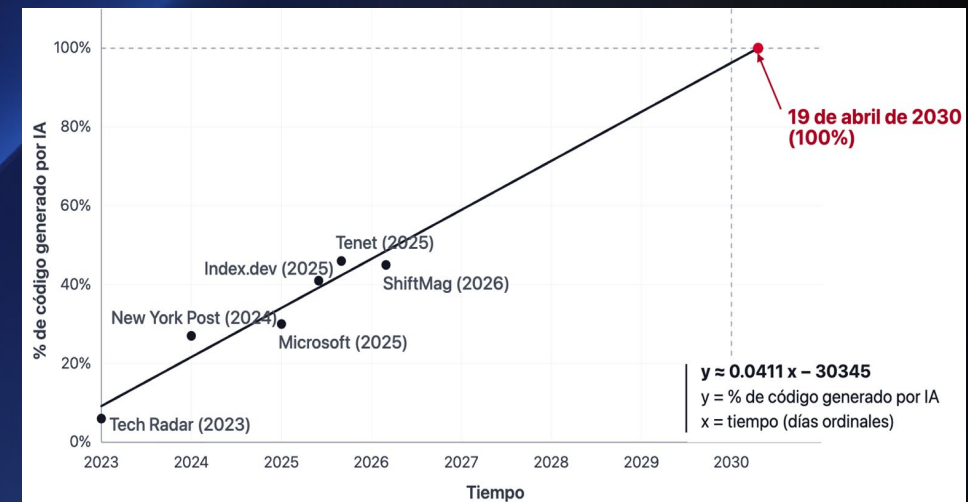
Who decides what gets fine-tuned, what gets censored, what gets priced out, what gets deprecated?



Latin America in 2030

- By 2030, AI will write most code, contracts, and do most knowledge work
- Whoever owns the models captures the productivity surplus, the rest of the world rents it

Who decides what gets fine-tuned, what gets censored, what gets priced out, what gets deprecated?



Thank You





Omar Florez

Team Lead for Pre-Training of LatamGPT

in



