
Dharma-AI
March 202

www.dharma-ai.com

GPT-5 Claude Gemini

deepseek LLaMA by Meta xAI

# LLM

Large Language Models

Hoje, LLMs tem entre
400Bi+ e 2 Tri+ parâmetros

# SLM

Small Language Models

SLMs vão de
8Bi até 100Bi

MISTRAL AI_ cohere

Qwen Gemma

Ai2

THINKING MACHINES

DHARMA-AI

**Somos o 1º Lab de AI na América Latina**

# SSLM

Specialized Small Language Models

SSLMs são tão **eficientes** quanto SLM,
mas treinados com dados da **sua indústria e da sua organização**
Funcionam muito bem para alta volumetria de requisições, tarefas específicas/especializadas e são ideias para segurança dos seus dados e informações

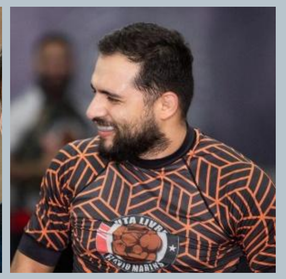
**Fontes:**; https://www.weforum.org/agenda/2024/04/how-to-manage-ais-energy-demand-today-tomorrow-and-in-the-future/; https://www.newyorker.com/news/daily-comment/the-obscene-energy-demands-of-ai; https://www.nytimes.com/2023/10/10/climate/ai-could-soon-need-as-much-electricity-as-an-entire-country.html;

# Por que a DHARMA-AI?

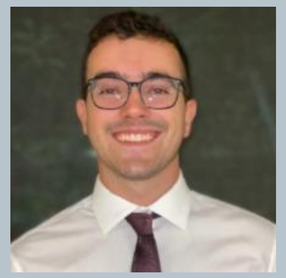
**Gabriel Renault**
-> CEO & Founder

**Elisa Mussumeci**
-> CTO & Founder

**Francisco Alves**
-> COO & Co-Founder



**FelipeGochi**
-> Head of Engineer & Co-Founder



**Gabriel Pimenta**
-> Head of R&D & Co-Founder

**Gustavo Lucchetti**
-> Head of Data Science & Co-Founder

Startup acelerada por:

NVIDIA Inception Program

Google for Startups

aws startups
**Generative AI Accelerator 2025 Cohort**

Only 35 startups selected globally

Only 3 from Brazil

USD 1MM in GPUs Credits

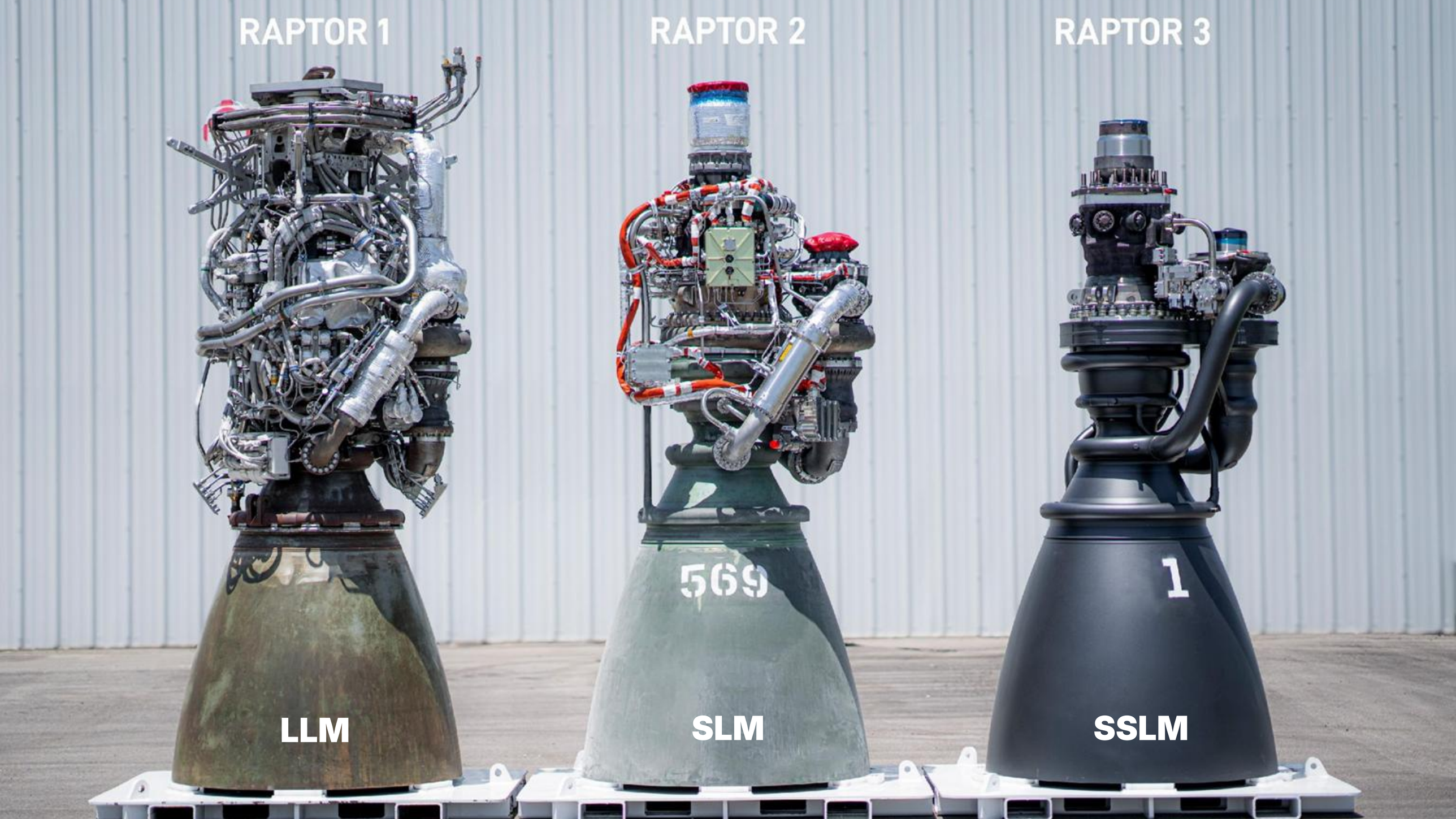

As três startups brasileiras que serão aceleradas pela Amazon

Valor Econômico

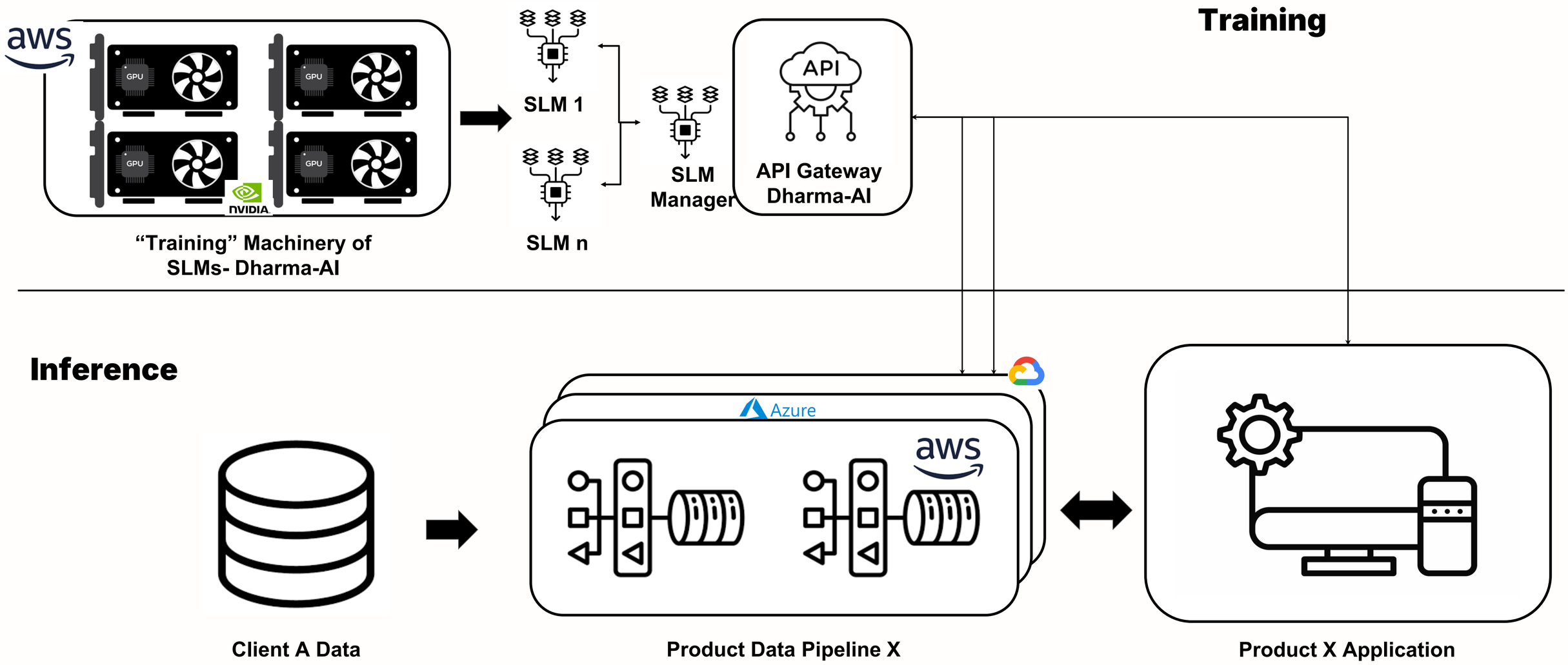
Pipeline

## HIGHLIGHTS OF THE CO-FOUNDERS

- Equipe com experiência **empreendedora em startup de IA com Exit** (fundada em 2015 e vendida em jan/2019). É a nossa 2ª startup juntos!
- **Mais de 15 anos em IA**, desenvolvendo e entregando projetos de inteligência artificial (desde 2009)
- **~200 projetos de IA** bem-sucedidos, com **mais de R$ 2 bilhões em ganhos** para nossos clientes
- C**ooperação em pesquisa** com as melhores universidades do Brasil: **UFPB, FGV Emap e IMPATech**
- Vencedores do chamamento público do STF (Supremo Tribunal Federal), **superando Palantir** e H2O entre 82 empresas


RAPTOR 1
RAPTOR 2
RAPTOR 3
569
1
LLM
SLM
SSLM

# Desenvolvemos tecnologia própria para treinar os melhores SSLMs

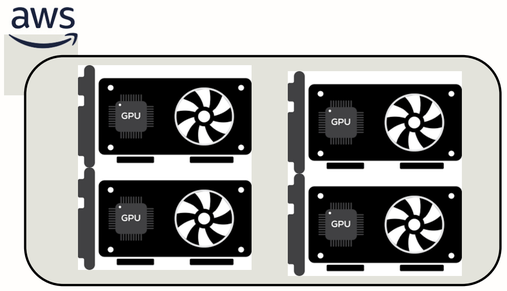


# DHARMA-AI
# Smart Training LAB

## Training data sets layer

## Models layer

## Trainninglayer




Ambiente de Treinamento de SLM - Dharma-AI

Sinthetic data generation

CPT training datasets

Fine Tunning training datasets

Gemma

Qwen

deepseek

Amazon Nova

Claude

LLaMA by Meta



Partial Fine Tuning

Full Fine Tuning

CPT training

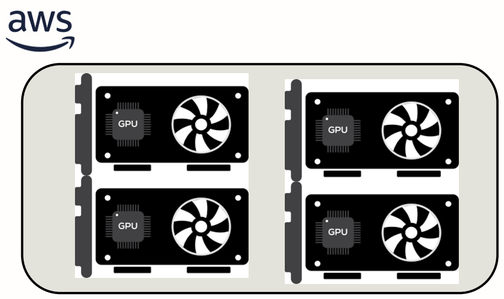
RL training

PEFT Methods

ModelParametrization

# DHARMA-AI
# Smart Training LAB

| Training data sets layer | Models layer | Traininglayer |
|---|---|---|




SLM Training Environment - Dharma-AI

Sinthetic data generation

CPT training datasets

Fine Tunning training datasets

Gemma

Qwen

deepseek

Amazon Nova

Claude

LLaMA by Meta

Partial Fine Tuning

PEFT Methods

Full Fine Tuning

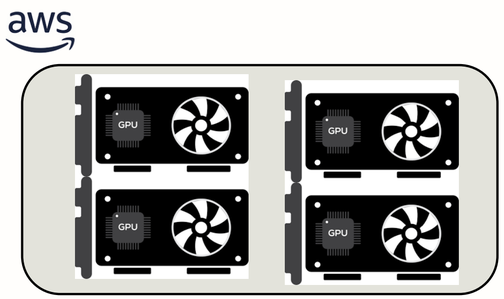
ModelParametrization

CPT training

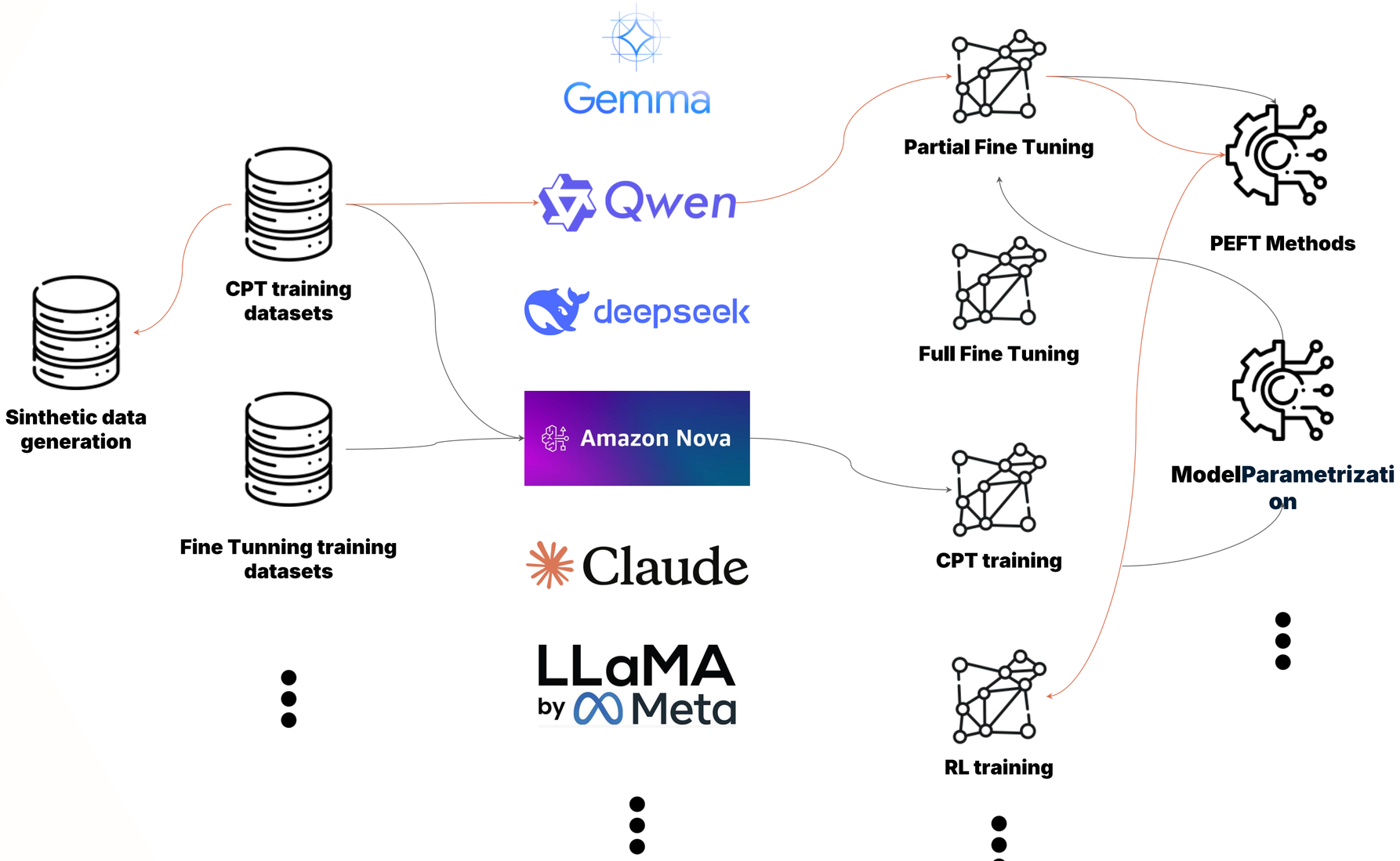
RL training

# DHARMA-AI
# Smart Training LAB

| Training data sets layer | Models layer | Traininglayer |
|---|---|---|




SLM Training Environment - Dharma-AI

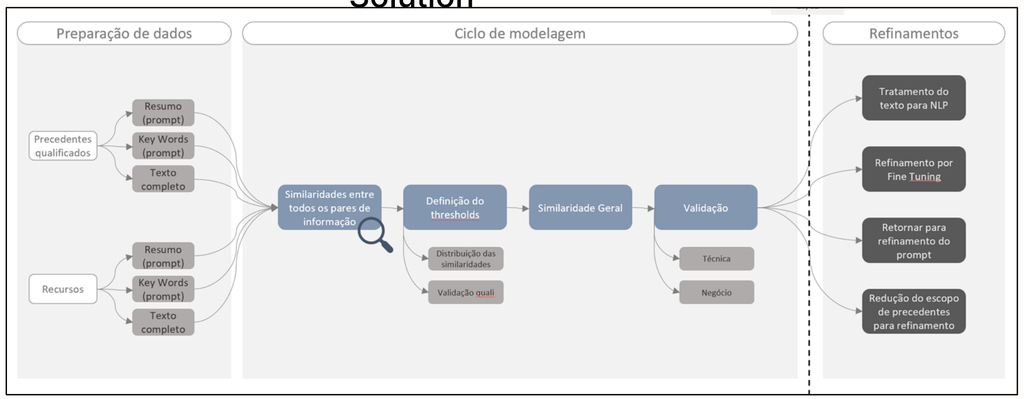


# SSLMs e seu impacto na prática

Caso real em um grande tribunal de justiça brasileiro usando IA para identificação de precedentes qualificados

Claim Input

LLM + Prompts + RAG Solution




Qualified Precedent Recommendation

| | #TESES |
|---|---|
| SÚMULA VINCULANTE | 59 |
| REPERCUSSÃO GERAL | 656 |
| IAC DE BRASÍLIA | 15 |
| REPETITIVOS | 864 |

**RG 15**

**RG 8**

**~1,600 Precedentes qualificados/mês**

✕ Média de 2.500 recursos de segunda instância por dia

✕ **23 prompts** para ter 88% acurácia

✕ **2+ trilhões de parâmetros do** GPT 4o

= ~ R$ 1MM/mês

Com DHARMA´s SSLMs
**= < 100k/ mês**

# SSLM e seu impacto na prática

Precentes Qualificados no judiciário brasileiro

| | Accuracy | Cost per 100 Calls | Response time per 100 Calls |
|---|---|---|---|
| DHARMA-AI | 86% | $ 0,003 | 6s |
| GPT - 4 | 68% | $ 0,608 | 33.5min |
| Qwen2.5 | 8% | $ 0,003 | 5s |

**~200x mais eficiente em custo**

**~300x mais eficiente em tmepo**

# We Make AI Your Model

## Outros Exemplos de SSLMs aplicados na prática

Correção de Redação do ENEM

| | Best Answer judged by GPT itself | Cost per Essay Online | Cost per Essay Offline |
|---|---|---|---|
| DHARMA-AI | 74% | R$ 0,30 | R$ 0,05 |
| GPT - 4 | 26% | R$ 0,90 | R$ 0,50 |

Resolução de questões sobre a constituição brasileira

| | Accuracy | Cost per Test Correction |
|---|---|---|
| DHARMA-AI | 93% | $ 0,15 |
| GPT - 4 | 86% | $ 2,83 |
| Microsoft Phi-3 | 40% | $ 0,14 |

DHARMA-AI

# OCRs ruins matam boas soluções de IA

✕ **Perda de qualidade e acurácia na fonte dos dados**

Erros no OCR se propagam em toda solução da IA

✕ **Textos manuscritos são facilmente ignorados**

A maioria dos OCRs trabalham com tecnologia antiga de reconhecimento de caracter

✕ **Não entende context**

Não entende bem dados tabelados, busca por um template fixo e etc

**30–50%**

Dos dados empresariais contém partes manuscritas, scan de qualidade ruim, imagens de baixa resolução e etc

*Source: Gartner / IDC Document Intelligence Reports*

*O custo de ter um OCR ruim*
*Não é o OCR em si*
*É toda a solução que vem depois dele*

*Traditional OCR*


SMART OCR

# A única solução de processamento inteligente de documentos que pode ser especializada para a sua empresa

Todo player te Entrega um produto fixo. Nós partimos de um lugar melhor e melhoramos

DHARMA-AI

## Out of the box — already the best.

| Rasterização em grandes volumes | OCR Lite vs Full | Image & PDF support | Smart NLP extraction |
|---|---|---|---|
| 200K+ pages | cost control | all formats | structured output |

## Fazemos com que seja seu

*No other commercial OCR solution offers this.*

**Treinado no seu tipo de documento**

**Construímos um case em cima da sua realidade**

**Podemos deployer no seu ambiente**

*Competitors offer fixed products. None offer custom training on client data.*

## SMART OCR - BENCHMARK RESULTS

Dharma Smart OCR outperforms every commercial model and every open-source vanilla alternative across the board.

# #1

Bench Score
among Commercial Solucions
and Vanilla SLMs

DHARMA-AI

Cloud Vision API

**35%**

Melhor que Google Vision

GPT-4o

**45%**

Melhor que  Document AI and GPT-4o

Amazon Textract

**50%**

Melhor que Amazon Textract

MISTRAL AI OCR

**62%**

Melhor que Mistral OCR 3

Qwen dots.ocr GLM-OCR deepseek

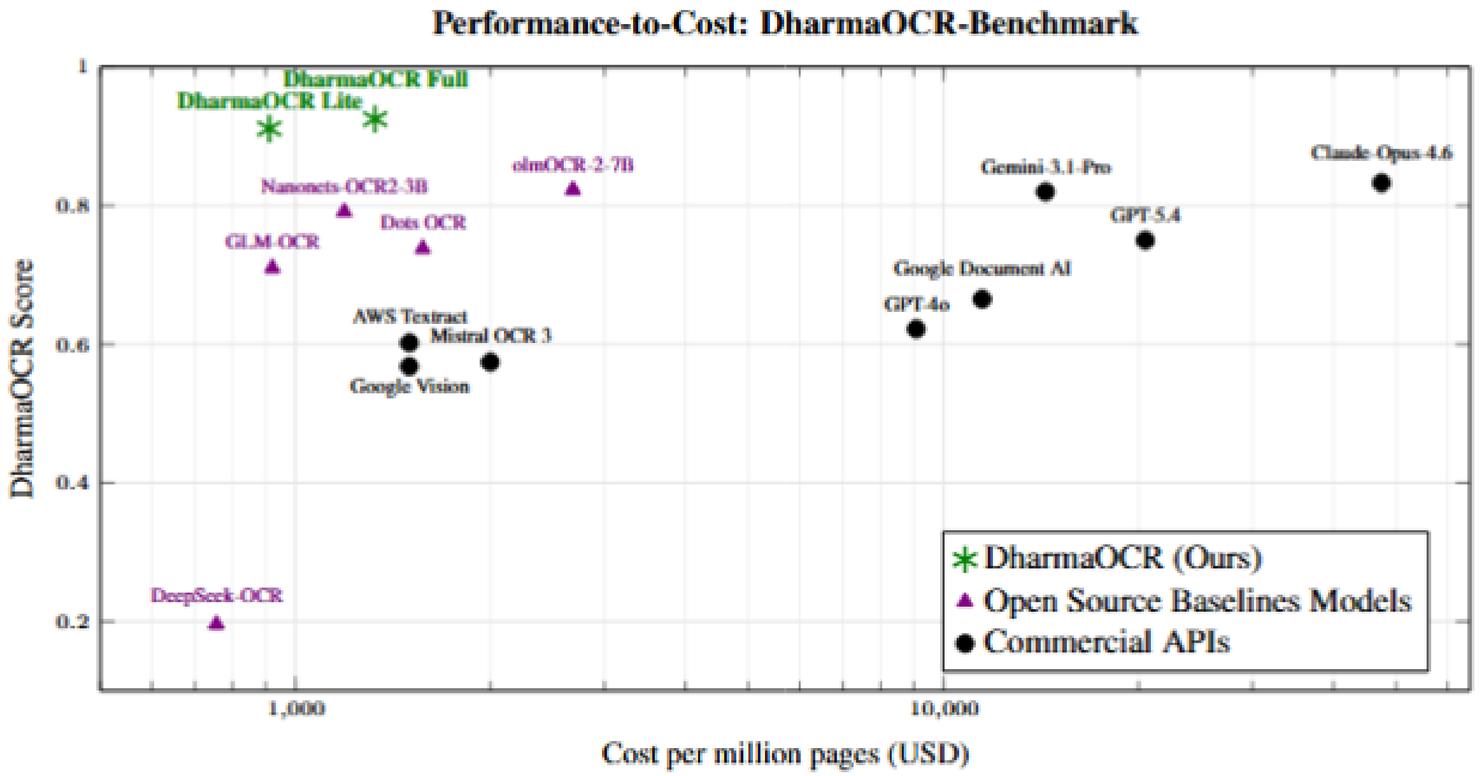
**Up to 376 %**

Melhor que os open-sources

*Source: DharmaOCR Internal Benchmark · olmOCR-2-7B Full Fine-tuning (SFT) · Tested against 25+ models across Tuned, Open-Source Vanilla and Commercial categories*

## SMART OCR - BENCHMARK RESULTS

Dharma bate todos os benchmarks de modelos comerciais e open-sources vanillas




Dharma's intelligent document processing não apenas bate em **qualidade,** mas em **custo também**

*Source: DharmaOCR Internal Benchmark · olmOCR-2-7B Full Fine-tuning (SFT) · Tested against 25+ models across Tuned, Open-Source Vanilla and Commercial categories*

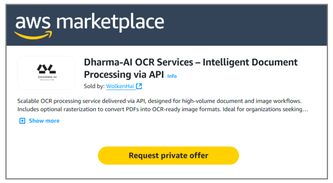
# We Make AI Your Model

- Applying to **substitute AWS Textract**

- Being the **1st AI model from LATAM at AWS Bedrock**

Others and examples real performance of our SSLMs — performance and costs:

OCR:




| OCR (OpticalCharacterRecognition) | Quality (Text Extraction Accuracy) | Price per 1000 pages |
|---|---|---|
| DHARMA-AI<br>OCR Dharma-AI: | 👍 👍 👍 | U$ 0,60 a U$ 1,50 |
| Amazon Textract<br>Smart OCR AWS Textract: LLM | 👍 👍 👍 | U$ 25,00 a U$ 50,00 |
| Document AI Google Cloud<br>SmartOCR Google: LLM | 👍 👍 👍 | U$ 6,00 a U$ 30,00 |

| | | |
|---|---|---|
| Amazon Textract<br>OCR AWS: Textract | 👍 | U$ 0,60 a U$ 1,50 |
| Google Vision API<br>OCR Google | 👍 👍 | U$ 0,60 a U$ 1,50 |

DHARMA-AI

# Challenges we solve

**Big B2B**

GenAI Solutions for High Volume Agents

**Reduzimos custo de inferência que inviabiliza uso de IA em larga escala**

**Reduzimos tempo de inferência** for Agents that need really fast

**We guarantee your Language Model safety**: No Hallucination / Auditable training data / Runs in private cloud, your cloud or even On Premise

**90% emissions in CO2 with AI**

We don't need **things to be big.**

We need things **to be SMART.**

DHARMA-AI

# Why DHARMA-AI?

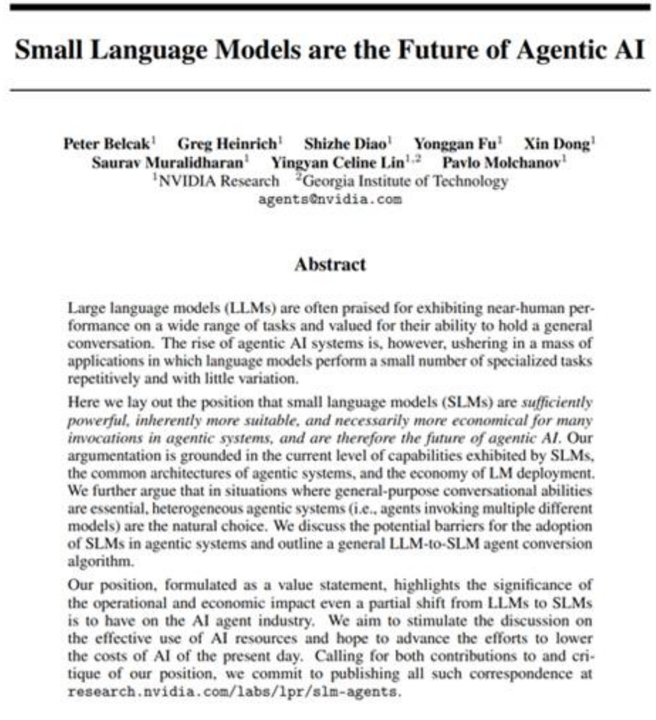
DHARMA-AI

**Article published on MIT in Mar/2025**


https://mitsloanreview.com.br/slms-nossa-proxima-fronteira/

NVIDIA

**Article published on Arxiv in Jun/2025**



Small Language Models are the Future of Agentic AI

Peter Belcak[1] Greg Heinrich[1] Shizhe Diao[1] Yonggan Fu[1] Xin Dong[1]
Saurav Muralidharan[1] Yingyan Celine Lin[1,2] Pavlo Molchanov[1]
[1]NVIDIA Research [2]Georgia Institute of Technology
agents@nvidia.com

**Abstract**

Large language models (LLMs) are often praised for exhibiting near-human performance on a wide range of tasks and valued for their ability to hold a general conversation. The rise of agentic AI systems is, however, ushering in a mass of applications in which language models perform a small number of specialized tasks repetitively and with little variation.

Here we lay out the position that small language models (SLMs) are *sufficiently powerful, inherently more suitable, and necessarily more economical for many invocations in agentic systems, and are therefore the future of agentic AI*. Our argumentation is grounded in the current level of capabilities exhibited by SLMs, the common architectures of agentic systems, and the economy of LM deployment. We further argue that in situations where general-purpose conversational abilities are essential, heterogeneous agentic systems (i.e., agents invoking multiple different models) are the natural choice. We discuss the potential barriers for the adoption of SLMs in agentic systems and outline a general LLM-to-SLM agent conversion algorithm.

Our position, formulated as a value statement, highlights the significance of the operational and economic impact even a partial shift from LLMs to SLMs is to have on the AI agent industry. We aim to stimulate the discussion on the effective use of AI resources and hope to advance the efforts to lower the costs of AI of the present day. Calling for both contributions to and critique of our position, we commit to publishing all such correspondence at `research.nvidia.com/labs/lpr/slm-agents`.

:2506.02153v1 [cs.AI] 2 Jun 2025

https://arxiv.org/pdf/2506.02153v1


gabriel.renault@dharma-ai.com
+55 21 98187 2663