AI21 labs

Jamba

Training a Foundational LLM

Announcements

Introducing Jamba: AI21's Groundbreaking SSM-Transformer Model

Debuting the first production-grade Mamba-based model delivering best-in-class quality and performance.

Announcements

Built for the Enterprise: Introducing AI21's Jamba-Instruct Model

An instruction-tuned version of our hybrid SSM-Transformer Jamba model, Jamba-Instruct is built for reliable commercial use, with best-in-class quality and performance.

Coming to Databricks Marketplace and External Model Serving

Agenda

AI21 labs

- Jamba
  - Transformer vs. Mamba
  - Advantages of hybrid architecture
  - LLM training

- What is the future: compound AI systems

Pioneers at the forefront of AI development

Prof. Amnon Shashua
Chairman

mobileye
ONE ZERO
בנקאות פרטית דיגיטלית
ORCAM

Prof. Yoav Shoham
Co-CEO & Co-Founder

Stanford University
Google

Ori Goshen
Co-CEO & Co-Founder

CrowdX

Our Journey

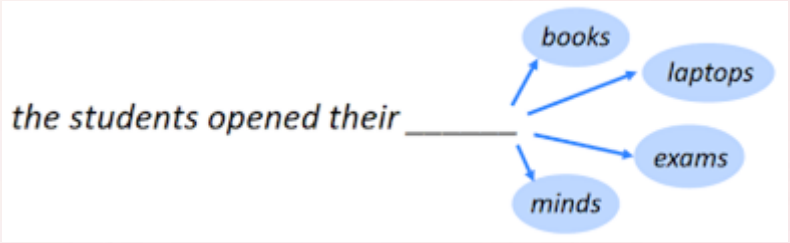
- Founded in 2017: NLP/ML research lab
- Wordtune (10M+ users) → specialized language models
- Focus: enterprise-ready AI systems with smaller language models (~1B to ~60B parameters),
- Major investors: Google, NVIDIA, Intel Capital

What are Large Language Models?

Tokenizatio n

+

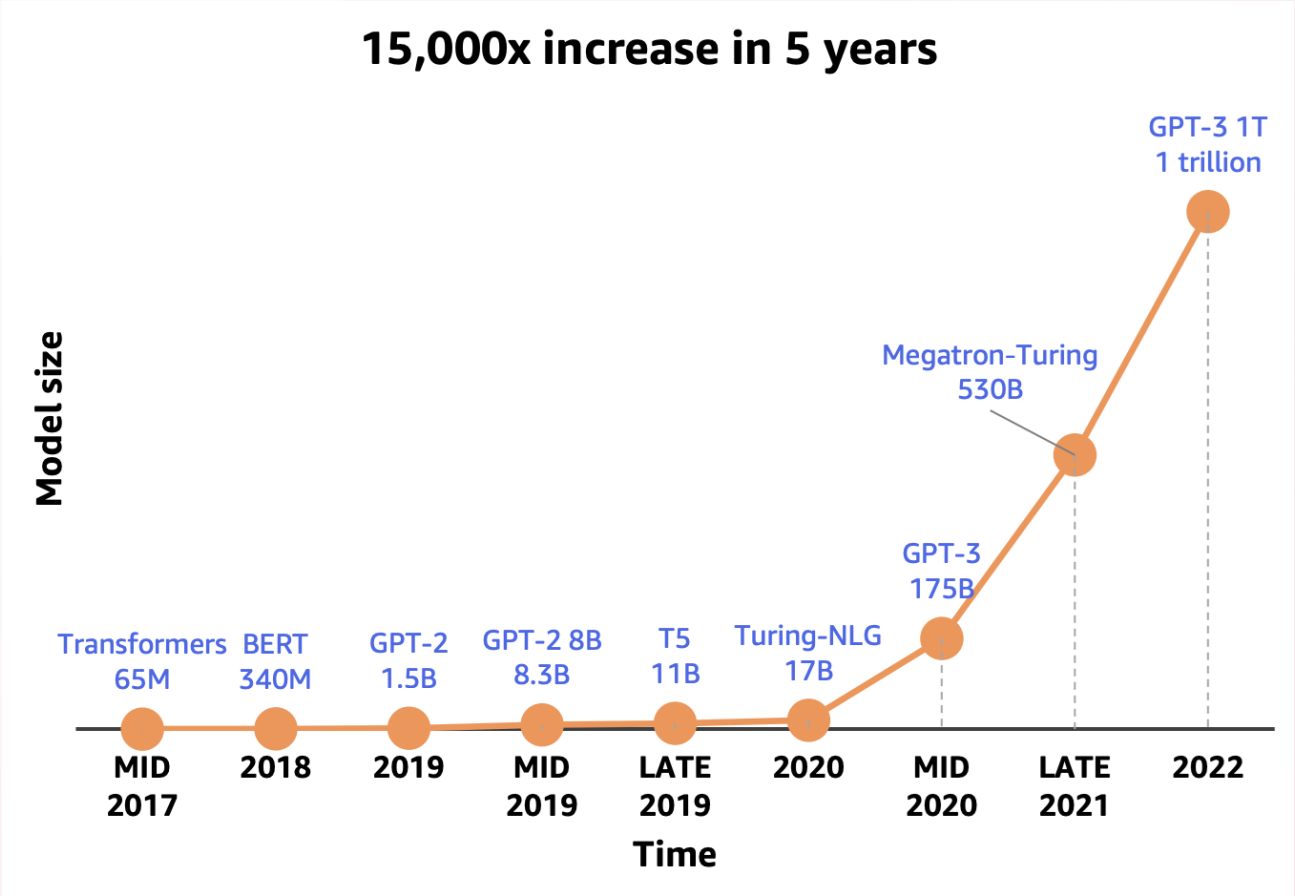
Next Token Prediction



=

LLM Chat Application

What changed?



Timeline

2007 — RNN

2017 — Attention is All You Need

2018 — BERT

2020 — GPT3

2022 — ChatGPT (3.5)

2023 — GPT4

2023 — Mamba

2017 — AI21 Founded

2019 — SenseBERT

2020 — Jurassic

2022 — MRKL

2023 — Jurassic-2

2024 — Jamba

wordtune

Jamba

	Transformer	Mamba	Jamba
Highest Quality Output	✓		✓
High Throughput		✓	✓
Low Memory Footprint		✓	✓

Jamba architecture under the hood

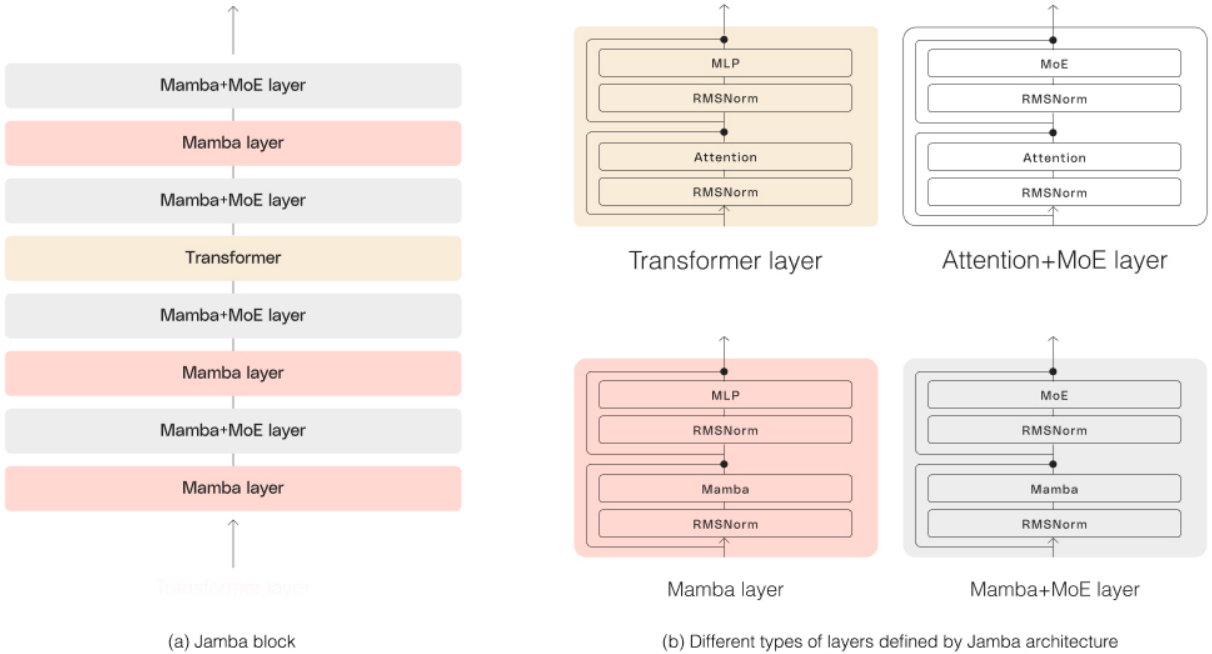
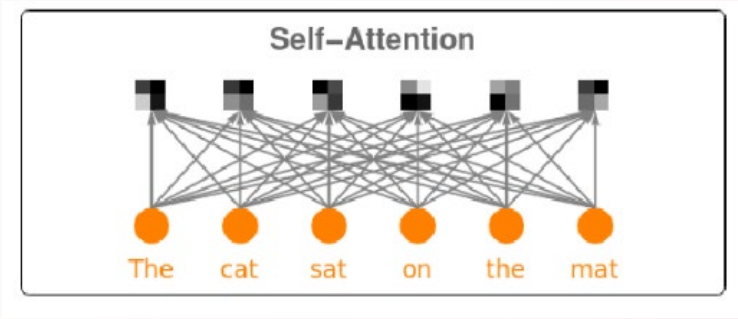


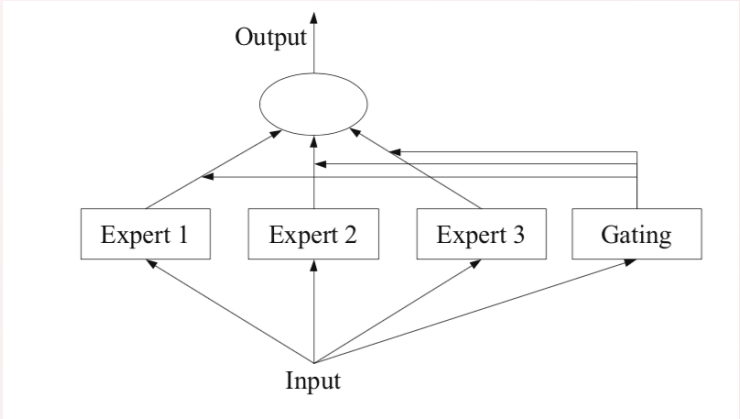
Diagram showing (a) a single Jamba block, (b) Different types of layers. Jamba implementation includes 4 Jamba blocks, each containing 8 layers, a 1/7 ratio of attention/Mamba layers, and MoE applied every 2 layers

Significant Advancements in LLMs

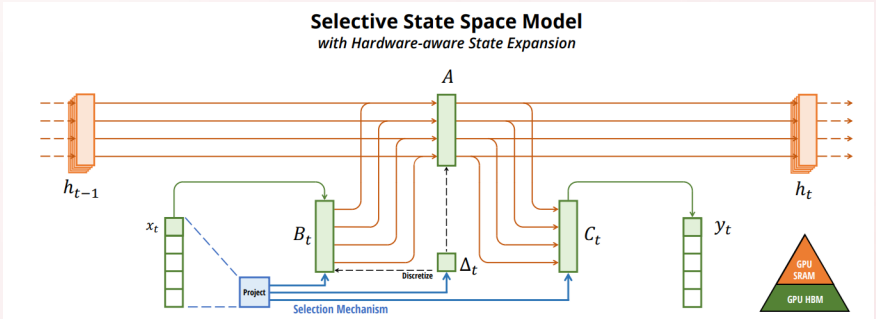
Attention (2016)



Mixture of Experts (2021)



Mamba (2023)



Jamba architecture - How to compare models?

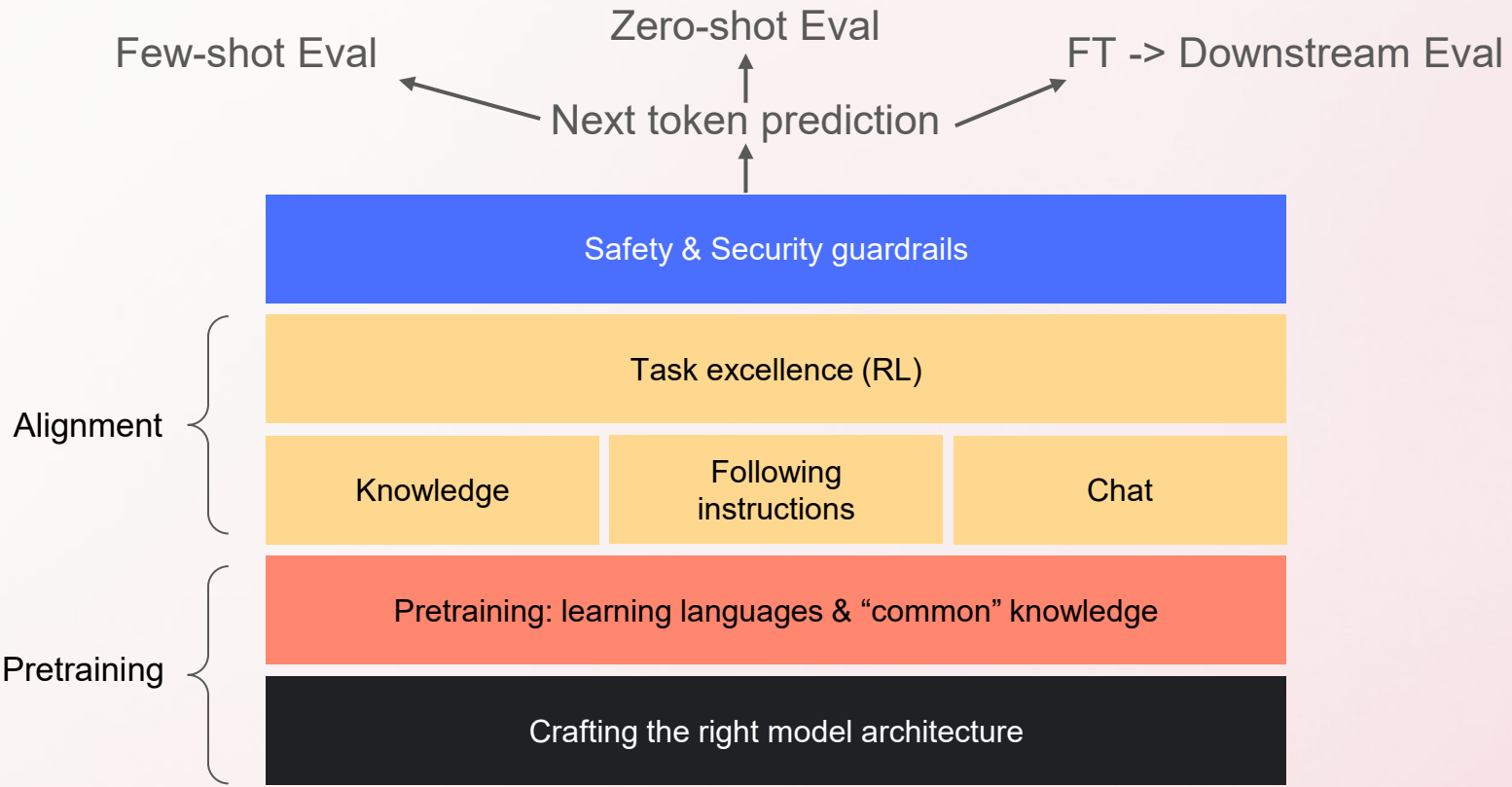
- “Vanilla Transformer” - **Overall parameters** (“model’s capacity” / “available parameters”)
- Mixture of Experts - Overall parameters + **Active parameters**
- Jamba - Overall parameters + Active parameters **+ Cache size**

	Available params	Active params	KV cache (256K context, 16bit)
LLAMA-2	6.7B	6.7B	128GB
Mistral	7.2B	7.2B	32GB
Mixtral	46.7B	12.9B	32GB
Jamba	52B	12B	4GB

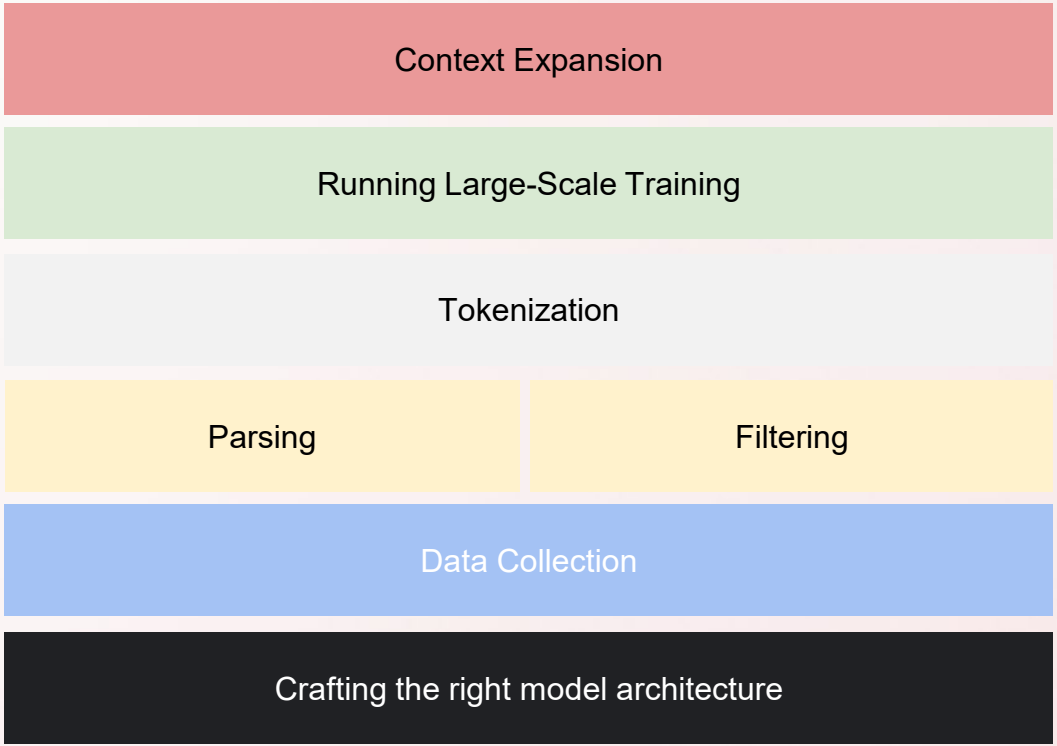
Gains:

- Very long context window
- High throughput
- Maintains high quality

LLM 101:What does it take to build an LLM?



Base Model - Pretraining



What is Alignment?

I need a list of 3 innovative ways to use peanut butter that go beyond sandwiches and salad dressing. Number the items on the list using lowercase roman numerals.

Complex instruction following

A:

i. Peanut Butter Energy Balls
ii. Peanut Butter Hummus
iii. Peanut Butter Smoothie

> Format Constraint: Pass
> List Length Constraint: Pass

B:

1. Peanut Butter Energy Balls
2. Peanut Butter Hummus

> Format Constraint: Fail
> List Length Constraint: Fail

Quality (human preference)

A:

i. Peanut Butter Energy Balls: Mix peanut butter with oats, honey, and dried fruit for a healthy snack.
ii. Peanut Sauce: Combine peanut butter, soy sauce, garlic, ginger, and lime juice for an Asian-inspired dipping sauce or marinade.
iii. Peanut Butter Smoothie: Blend peanut butter, banana, milk, and ice for a protein-packed drink.

B:

i. Peanut butter in chili
ii. Peanut butter in smoothies
ii. Peanut butter in baked goods

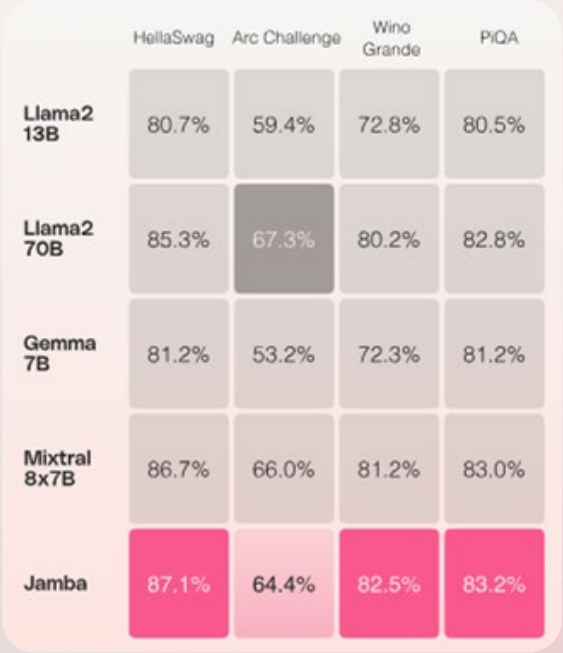
A > B, because A is more informative, comprehensive, helpful, engaging…

Jamba - What does it REALLY take to train an LLM (+ new arch)

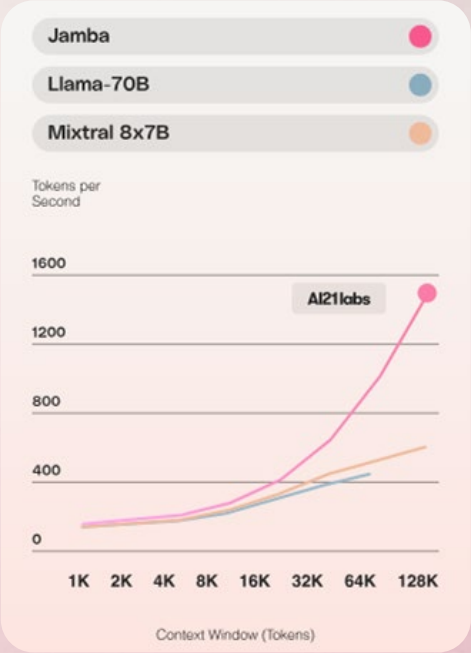
- Perform ablations, starting from small scale and increasing to the target size (100-1000s of small runs)
  - Data ablations
    - Data mixing, parsing, filtering, math, reasoning, multi lingual, coding, …
  - Arch ablations
    - Mamba vs. Transformer
    - MoE ablations, number of experts in each layer, top experts to pick, load balancing
- Decide on evals to run (zero shot, few shot, PPL, fine tuning)
- Compute
  - Training infrastructure for 1000s of GPUs
  - Different types of parallelism
  - Monitoring, CP saving, …

And it works really well

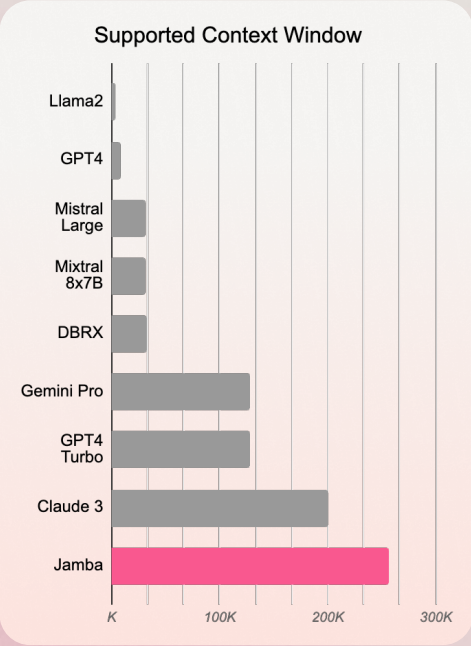
Reasoning Quality



Throughput



Long context



Jamba highlights

Mamba + Transformer MoE

Novel 7B architecture w/ 12B active parameters (52B total)

Best-in-class

Outperforms all models in its size-class

3X Throughput

Over other similar-sized models

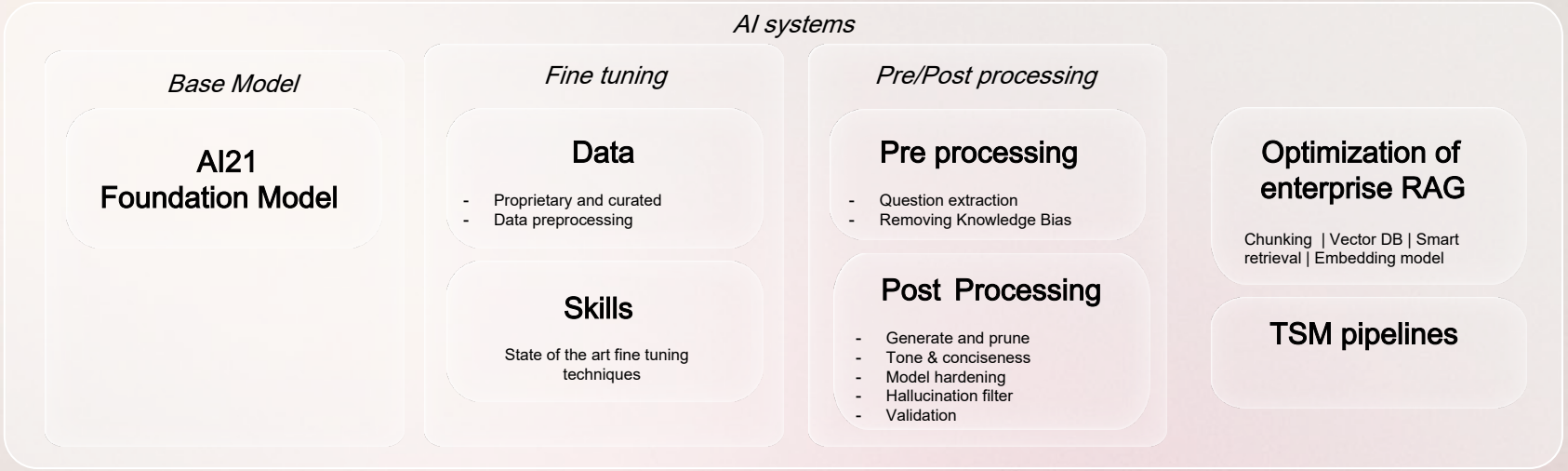
140k context on 1 GPU

256k max context size

From large language models to compound AI systems

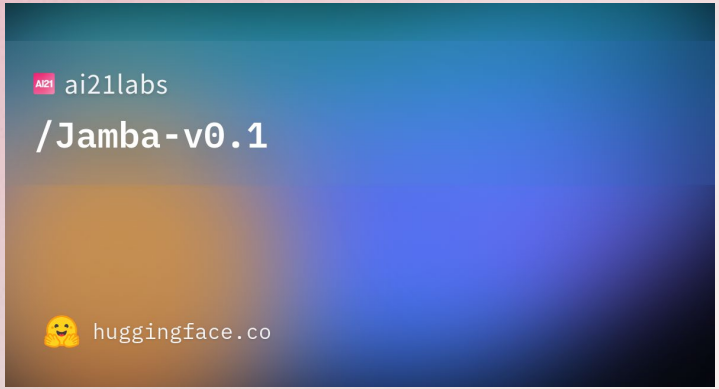
- Moving from discrete model API calls to a robust AI System with controllability, guardrails, long term memory, interpretability, etc.
- The system will provide seamless access to:
  - Multiple specialized (small) models to handle different tasks in the pipeline
  - Access to additional tools: external API calling, code execution, web search, etc.
  - Vector DBs
  - Interoperability with other technology for data input and output
  - …

What do we mean and how do we build AI systems - Task Specific Models



- Optimized for enterprise use cases (support, increasing knowledge-workers efficiency)
- **No prompting** or fine tuning needed for customer
- **Model hardening** to avoid “jailbreaking” and harmful or undesirable behaviour
- Easy to maintain - no model adaptations between versions
- Low memory footprint results in low latency & low cost
- “Out of the box” capabilities ensure faster time to deployment

Jamba-v0.1 on 🤗



https://huggingface.co/ai21labs/Jamba-v0.1

AI21Studio

🎉 **Just released! Jamba-Instruct is now available in public preview.** Try it now

https://studio.ai21.com/home/chat/single-chat

AI21labs

Thank you!