MENLO
VENTURES

← Back to stories

The Modern AI Stack: Design Principles for the Future of Enterprise AI Architectures

By Matt Murphy, Tim Tully, Grace Ge, Derek Xiao, Katie Keller

10-minute read
January 18, 2024

The future of the modern AI stack is being decided now. More than ever, machines are capable of reasoning, creation, and creativity, and these new capabilities are driving enterprises to reconstruct their tech stacks. While the early days of this AI transformation felt like the Wild West, today, builders are converging around infrastructure, tooling, and approach. (We first documented this transition in Menlo's State of Enterprise AI report last November.)

Today, we're excited to share our thesis for how AI development will evolve, as well as the core infrastructure components that will combine to create the modern AI stack—the new runtime architecture that will drive AI applications for the coming decade.

Defining the Modern AI Stack

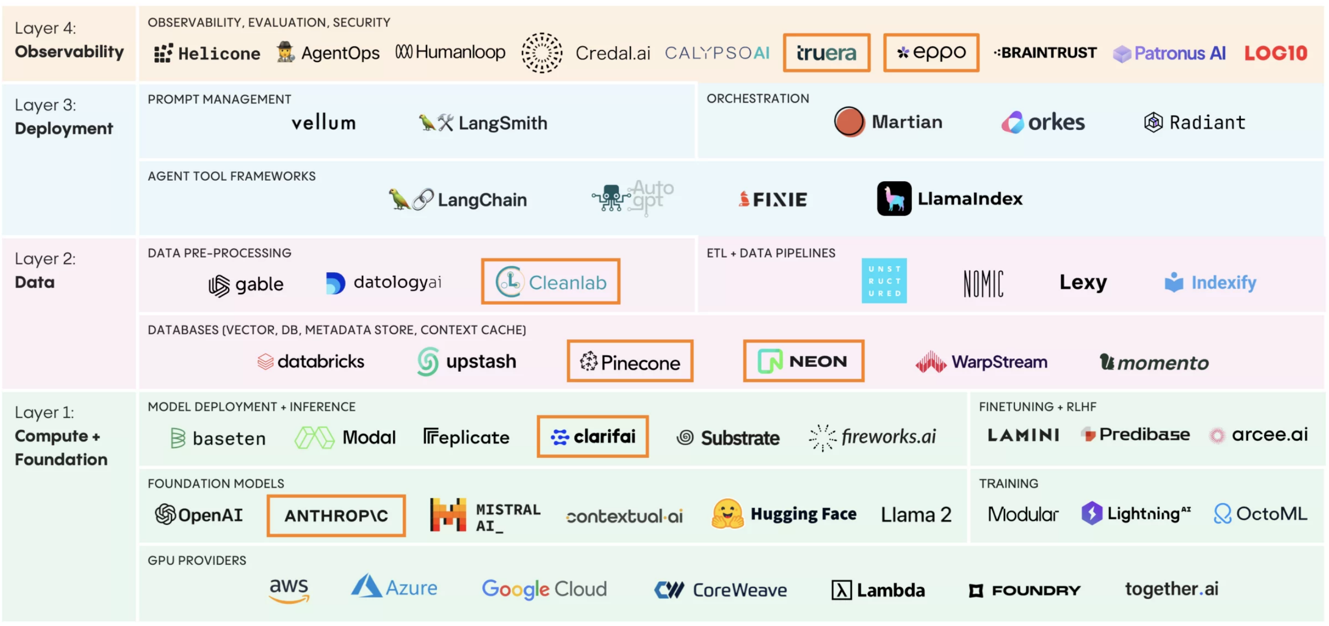
In 2023, enterprises spent over $1.1 billion on the modern AI stack—making it the largest new market in generative AI and a massive opportunity for startups.

At Menlo Ventures, we define the key layers of the modern AI stack as:

- **Layer 1: Compute and foundation models.** The compute and foundation model layer contains the foundation models themselves, as well as the infrastructure to train, finecontext wherever they may exist within enterprise data systems. Core components include data pre-processing, ETL and data pipelines, and databases like vector databases, metadata stores, and context caches.

- **Layer 3: Deployment.** The deployment layer contains the tools that help developers manage and orchestrate AI applications, and includes agent frameworks, prompt management, and model routing and orchestration.
- **Layer 4: Observability.** The final layer of the modern AI stack contains solutions that help monitor run-time LLM behavior and guard against threats, including new categories for LLM observability and security solutions.

Modern AI Stack: The Emerging Building Blocks for GenAI



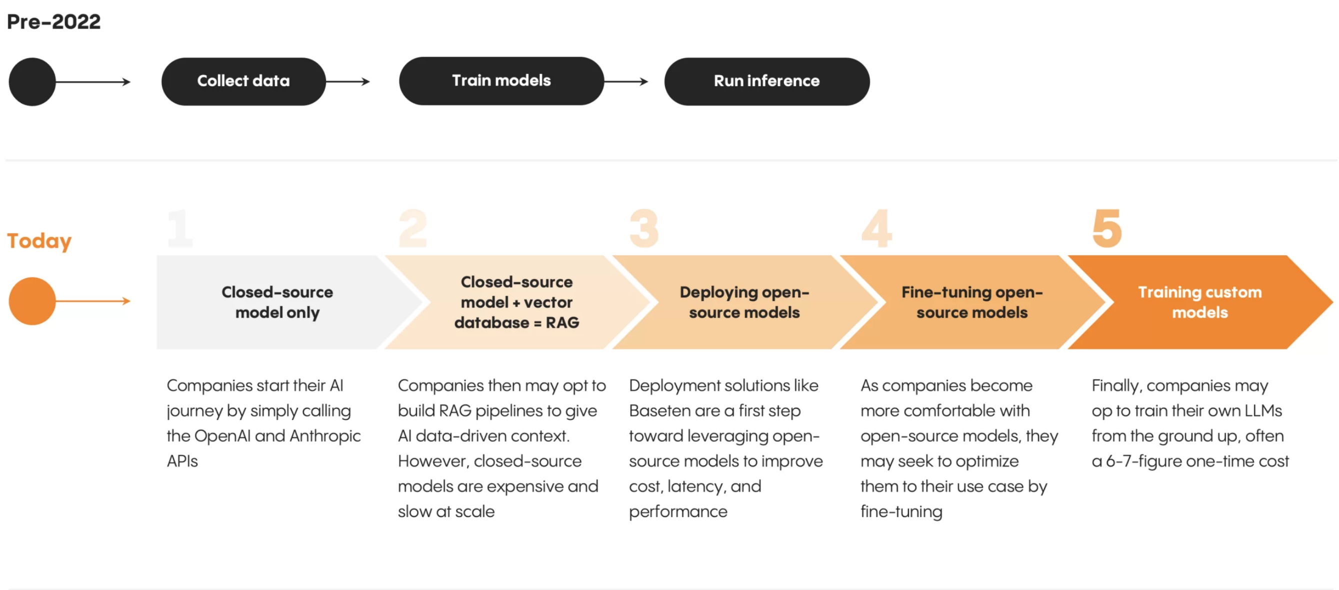
Layer	Category	Companies
Layer 4: **Observability**	OBSERVABILITY, EVALUATION, SECURITY	Helicone, AgentOps, Humanloop, Credal.ai, CALYPSOAI, truera, eppo, BRAINTRUST, Patronus AI, LOG10
Layer 3: **Deployment**	PROMPT MANAGEMENT	vellum, LangSmith
	ORCHESTRATION	Martian, orkes, Radiant
	AGENT TOOL FRAMEWORKS	LangChain, Auto gpt, FIXIE, LlamaIndex
Layer 2: **Data**	DATA PRE-PROCESSING	gable, datologyai, Cleanlab
	ETL + DATA PIPELINES	UNSTRUCTURED, NOMIC, Lexy, Indexify
	DATABASES (VECTOR, DB, METADATA STORE, CONTEXT CACHE)	databricks, upstash, Pinecone, NEON, WarpStream, momento
Layer 1: **Compute + Foundation**	MODEL DEPLOYMENT + INFERENCE	baseten, Modal, replicate, clarifai, Substrate, fireworks.ai
	FINETUNING + RLHF	LAMINI, Predibase, arcee.ai
	FOUNDATION MODELS	OpenAI, ANTHROPIC, MISTRAL AI_, contextual.ai, Hugging Face, Llama 2
	TRAINING	Modular, Lightning AI, OctoML
	GPU PROVIDERS	aws, Azure, Google Cloud, CoreWeave, Lambda, FOUNDRY, together.ai

© 2024 Menlo Ventures

□ Backed by Menlo Ventures

The New AI Maturity Curve

Both the market structure and technology that define the modern AI stack today are rapidly evolving. However, key components and leaders in these categories are already emerging.


Pre-2022
Collect data
Train models
Run inference
Today
1
Closed-source model only
2
Closed-source model + vector database = RAG
3
Deploying open-source models
4
Fine-tuning open-source models
5
Training custom models

Companies start their AI journey by simply calling the OpenAI and Anthropic APIs

Companies then may opt to build RAG pipelines to give AI data-driven context. However, closed-source models are expensive and slow at scale

Deployment solutions like Baseten are a first step toward leveraging open-source models to improve cost, latency, and performance

As companies become more comfortable with open-source models, they may seek to optimize them to their use case by fine-tuning

Finally, companies may op to train their own LLMs from the ground up, often a 6-7-figure one-time cost

© 2024 Menlo Ventures

Before LLMs, ML development was linear and “model-forward.” Teams seeking to build AI applications needed to start with the model—which often involved months of tedious data collection, feature engineering, and training runs, as well as a team of PhDs, before the system could be productionized as a customer-facing end product.

LLMs have flipped the script, shifting AI development to be “product-forward” and enabling teams without ML expertise to incorporate AI into their products. Now that anyone can access the OpenAI or Anthropic APIs to harness the world’s most powerful models instantly, companies can actually start with the product, rather than the model.

It’s easy enough to incorporate simple API calls into a product, but as AI stacks mature, dev teams look to customize their AI experience through enterprise- or customer-specific data. Teams start with prompt-level optimizations like retrieval augmented generation (RAG), but eventually move toward model-level optimizations, such as model routing, fine-tuning, or quantization, driven by considerations like performance, cost, and latency.

AI builders collectively evolved from traditional ML to the new AI maturity curve over the past year, locking in new building blocks as essential infrastructure for production AI systems at

- **Phase 1: Closed-source models only**. In the early days of 2023, dollars and engineering efforts were primarily focused on the foundation models themselves, with only relatively simple customizations on top (e.g., prompt engineering, few-shot learning). Leading closed-source model providers like OpenAI and Anthropic* gained early traction in this phase, cementing them as the earliest winners of the modern AI stack.
- **Phase 2: Retrieval-augmented generation**. In the next phase of the new maturity curve, enterprises have focused on the data layer as the center of gravity for their AI application efforts (as opposed to the model layer). The popularization of RAG in particular necessitated more robust data layer infrastructure like the vector database Pinecone* and data pre-processing engine Unstructured. Most enterprises and startups are currently at this stage.
- **Phase 3: Hybrid model deployment**. The third phase is the latest evolution in the AI maturity curve as leading companies like Typeface* and Descript start complementing their closed-source model usage with open-source for high-volume, domain-specific tasks. With this tailwind, model deployment vendors like Modal, Baseten, and Fireworks are starting to see significant traction.
- **Phase 4 and beyond: Custom models**. Although few companies have reached the sophistication or need to build their own models, down the road, we see use cases for large enterprises that want to reach deeper into the stack. Supporting them will be companies like Predibase and Lamini, which provide the tools for memory-efficient fine-tuning (including 4-bit quantization, QLoRA, and memory paging/offloading).

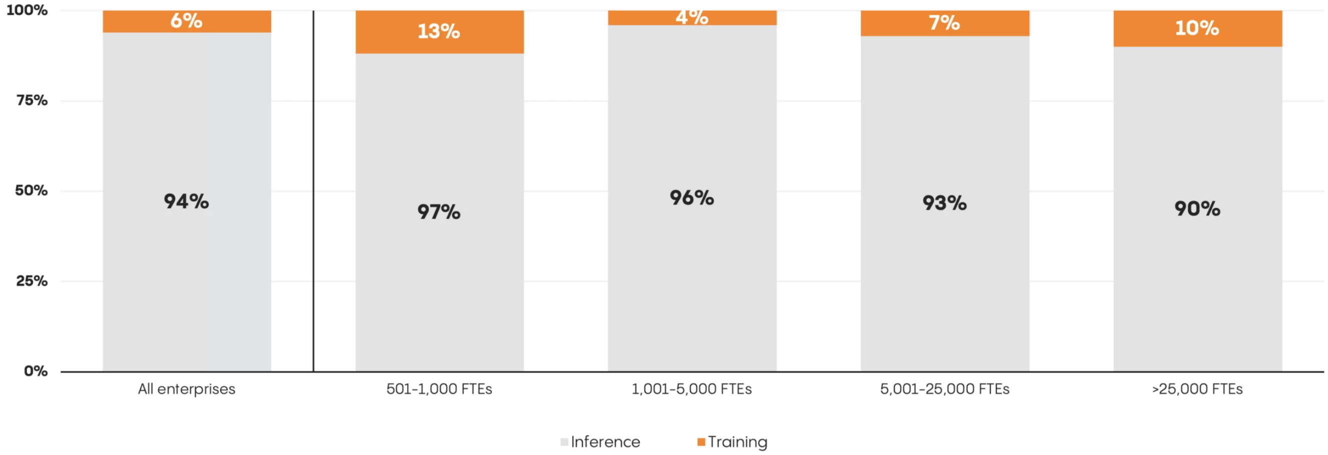
Four Key Design Principles for the New AI Infrastructure Stack

The AI revolution not only spurred demand for a new infrastructure stack, but actively reshaped how enterprises approach application development, R&D spend, and team composition. In the following section, we'll outline four key design principles for the new paradigm.

1. The Majority of Spend Is for Inference vs. Training

parameter LLM trained specifically on financial data released in March 2023, was heralded as an example of the flood of enterprise and domain-specific LLMs to come.

Estimated Spend by AI Adopters on Training vs. Inference

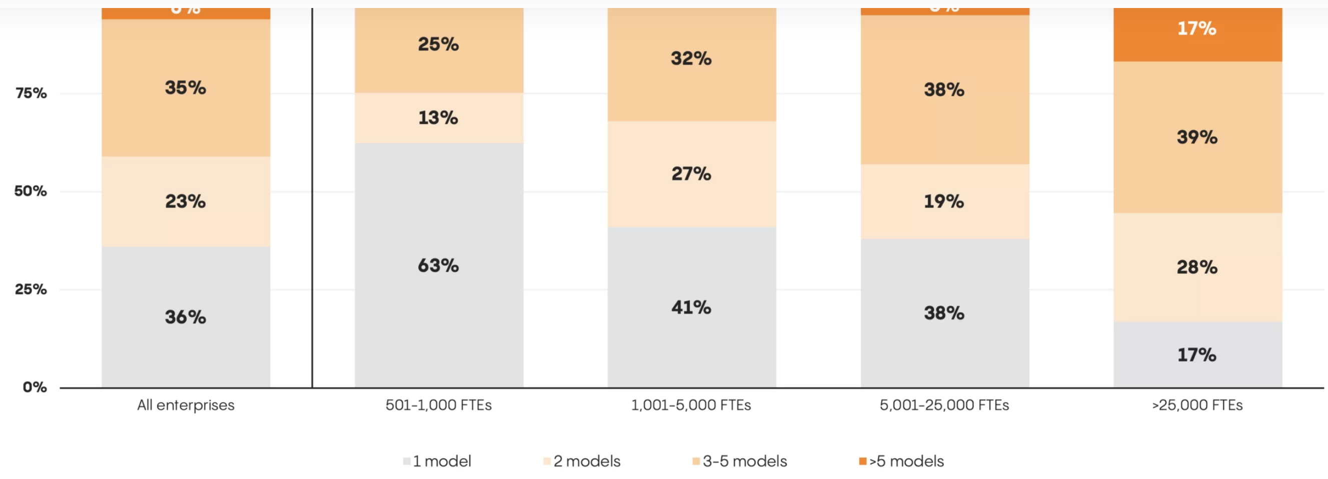


© 2024 Menlo Ventures

The expected deluge never materialized. Instead, Menlo Ventures' recent enterprise AI survey indicates that almost 95% of all AI spend is on run-time vs. pre-training. Only for the largest foundation model providers like Anthropic is this ratio flipped. At the application layer, even sophisticated AI builders like Writer are spending upwards of 80% of their compute on inference as opposed to training.

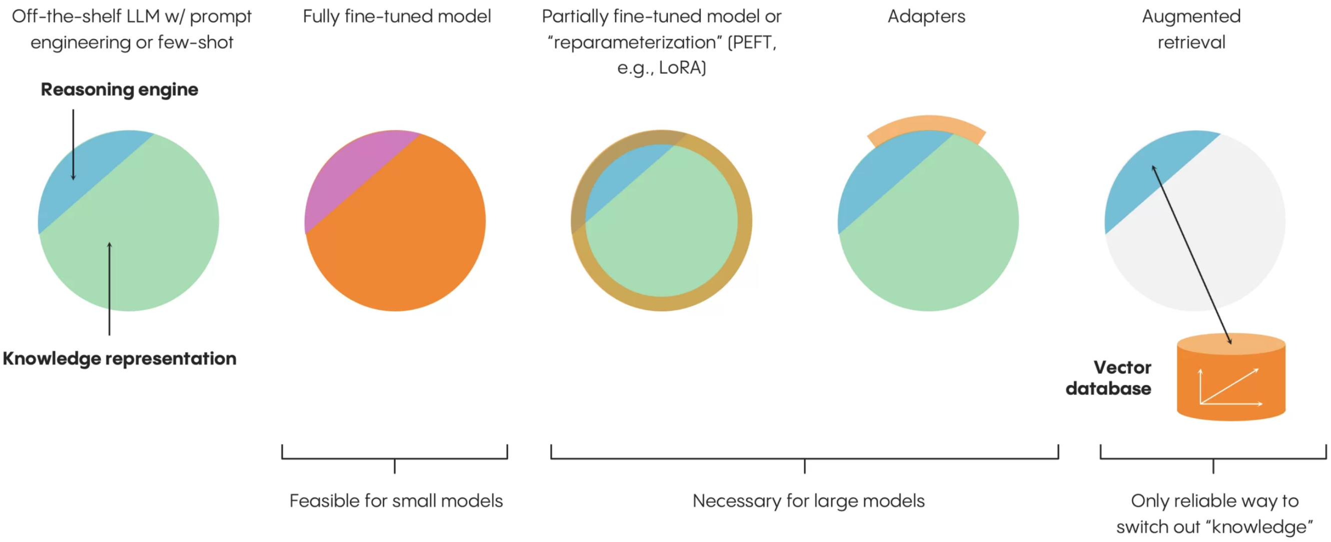
2. We Live in a Multi-Model World

A single model won't "rule them all." According to Menlo's Enterprise AI report, 60% of enterprises use multiple models and route prompts to the most performant model. This multi-model approach eliminates single-model dependency, offers higher controllability, and cuts costs.



3. RAG Is the Dominant Architectural Approach

LLMs are excellent reasoning engines, but have limited domain- and enterprise-specific knowledge. To create useful AI experiences, teams are quickly deploying knowledge augmentation techniques—starting with retrieval augmented generation, or RAG.


Off-the-shelf LLM w/ prompt engineering or few-shot
Fully fine-tuned model
Partially fine-tuned model or "reparameterization" (PEFT, e.g., LoRA)
Adapters
Augmented retrieval
Reasoning engine
Knowledge representation
Vector database
Feasible for small models
Necessary for large models
Only reliable way to switch out "knowledge"

RAG endows the base models with enterprise-specific "memory" via a vector database like Pinecone. This technique is far outpacing other customization techniques like fine-tuning, low-rank adaptation, or adapters in production today, which primarily work at the model layer as opposed to the data layer. Moving forward, we expect this trend to continue, and new pieces of the data plane—including data pre-processing engines (like Cleanlab*) and ETL pipes (like Unstructured)—to coalesce in runtime architectures.

4. All Developers Are Now AI Developers

Worldwide, there are 30 million developers, 300,000 ML engineers, and only 30,000 ML researchers. For those innovating at the very forefront of ML, our references estimate there may only be 50 researchers in the world that know how to build a GPT-4 or Claude 2-level system.

In the face of these realities, the good news is that tasks that used to require years of fundamental research and sophisticated ML expertise can now be accomplished in days or weeks by mainstream developers engineering data systems on top of powerful pre-trained LLMs.

Products like Salesforce's Einstein GPT (a generative AI copilot for sales) and Intuit Assist (the generative AI-powered financial assistant) were built primarily by lean teams of AI engineers: traditional full-stack engineers working on the data plane of the modern AI stack, as opposed to data scientists, ML engineers, or even ML researchers working at the model layer.

What's Next

The modern AI stack is rapidly evolving, and as we look forward to its continued progression this year, we see a number of developments emerging:

Next-Gen AI Applications Pilot More Advanced RAG

count-based chunking of documents and inefficient indexing and ranking algorithms. As a result, these architectures often suffer from problems like:

- **Context fragmentation**. In many academic benchmarks, the correct answer is in one place in documentation, but this is almost never the case in production codebases
- **Hallucinations**. LLMs degrade in performance and accuracy in multi-step reasoning tasks
- **Entity rarity**. “Sparse retrieval” (e.g., word-matching algorithms) sometimes work better than “dense retrieval” based on embeddings in one- or zero-shot scenarios
- **Inefficient retrieval**. High latency and costs

To address these problems, next-generation architectures are exploring more advanced RAG applications, folding in novel techniques like chain-of-thought reasoning, tree-of-thought reasoning, reflexion, and rules-based retrieval.

Small Models Become a Larger Share of the Modern AI Stack

As AI application builders increase their sophistication and focus deeper in the modern AI stack, the next phase of the maturity curve points towards a proliferation of fine-tuned, task-specific models for certain areas where larger closed-source models prove unwieldy or expensive. Infrastructure for building ML pipelines and fine-tuning will become critical in this next phase as enterprises create their own task-specific models. Quantization techniques like those offered by Ollama and ggml will help teams enjoy the full-speed boost that smaller models offer.

New Tools for Observability and Model Evaluation Emerge

For much of 2023, logging and evaluation were either not done at all, done by hand, or done with academic benchmarks that served as a starting point for the majority of enterprise applications. Our research suggests that close to 70% of AI adopters are using humans to review outputs as their main evaluation technique. That’s because the stakes are high: Customers expect and deserve high-quality outputs, and enterprises are smart to be promising new approaches like Braintrust, Patronus, Log10, and AgentOps.

Architectures Move Toward Serverless

As with the rest of enterprise data systems, we see the modern AI stack moving towards serverless over time. Here, we distinguish between "ephemeral machine"-type serverless (e.g., Lambda functions) vs. true scale-to-zero serverless (e.g., Neon's* architecture for Postgres).

For the latter, abstracting away the infrastructure relieves developers from the operational complexity of running applications, enables more rapid iteration, and allows enterprises to enjoy significant resource optimization by only paying for compute vs. availability. The serverless paradigm will be applied to all parts of the modern AI stack. Pinecone has embraced this approach with its latest architecture for vector compute. Neon has done the same for Postgres, Momento for caching, and Baseten and Modal for inference.

At Menlo, we're very active investors across all layers of the modern AI stack, including Anthropic, Pinecone, Neon, Clarifai, Cleanlab, Eppo, and Truera—as well as the companies building with these tools like Abnormal Security, Aisera, Eve, Genesis Therapeutics, Lindy, Matik, Observe.ai, Sana, Typeface, and Vivun. As the stack continues to evolve, we're looking to partner with infrastructure builders who will define its next critical building blocks. If you're building in this space, shoot us a note.

Matt Murphy (matt@menlovc.com)
Tim Tully (tim@menlovc.com)
Grace Ge (grace@menlovc.com)
Derek Xiao (derek@menlovc.com)
Katie Keller (katie@menlovc.com)

*Backed by Menlo Ventures