THE AI CAPABILITY STACK

From autocomplete to *autonomous.*

A two minute crash course in how modern AI actually works, and what we are building with it at Cal. **Eight small steps, one big idea.**

L1 BASE	L2 LOOK UP	L3 TOOLS	L4 MEMORY	L5 AGENT	L6 TEAM	L7 SKILLS	L8 TRUST

AHEAD SINCE 2018 · SWIPE TO LEARN →

AI 101 · LEARN THE LINGO

Talk the talk in *30 seconds.*

L1	**Base model** LLM predicts the next word.
L2	**Retrieval** RAG looks things up.
L3	**Tools** TOOL USE presses the buttons.
L4	**Memory** CONTEXT remembers you.
L5	**Agent** AGENTIC does a task solo.
L6	**Orchestration** MULTI-AGENT a crew of agents.
L7	**Skills** MCP plugs into anything.
L8	**Guardrails** GOVERNANCE safe to trust.

Now you can say **"RAG"** and **"agentic"** at dinner and mean it. We take each one slow next. **Swipe if you're in.**

LAYER 01 / 08

L1 LLM base

Predicts the next word, over and over. No memory, no actions, and its knowledge is frozen the day training stops.

> IN PLAIN ENGLISH
>
> *It is autocomplete on steroids. Nothing more, nothing less.*

REAL EXAMPLES

Claude | ChatGPT | Llama on Ollama

[■] WHAT CAL DOES HERE

LIVE — Model-agnostic. Claude, GPT, Gemini, Llama, even on-device.

BUILDING — A private on-device mode so borrower data never leaves the machine.

LAYER 02 / 08

L2 + Retrieval

Looks things up before it answers, pulling in fresh facts and your own documents it was never trained on.

IN PLAIN ENGLISH

> *You just handed it Google. The nerds call it RAG.*

REAL EXAMPLES

Perplexity | Copilot | LlamaIndex

[■] WHAT CAL DOES HERE

LIVE — AskCal reads **1,424 lender guideline docs** and cites every source it uses.

BUILDING — Auto-refresh, so lender guidance is never a day out of date.

LAYER 03 / 08

L3 + Tools

Now it can do things, not just talk. Run code, search the web, look up a record, use another app.

> IN PLAIN ENGLISH
> *Now it can press the buttons, not just describe them.*

REAL EXAMPLES

Claude + MCP | OpenAI tools | Gemini

[■] WHAT CAL DOES HERE

- LIVE — Agents pull live lender data, sync the CRM, and run quality checks.
- BUILDING — Two-way sync that writes updates back into the loan system, not just reads.

LAYER 04 / 08

L4 + Memory

Remembers past chats, your preferences, and what worked, instead of starting from zero every single time.

IN PLAIN ENGLISH

Now it remembers you from last time.

REAL EXAMPLES

Claude Projects | ChatGPT Memory | Mem0

WHAT CAL DOES HERE

LIVE — The loan file itself is the memory. Credentialed data, not scattered PDFs.

BUILDING — A borrower memory that carries context across every touch and every loan.

LAYER 05 / 08 · THE BIG LEAP

L5 = AI Agent

Put it all in a loop aimed at a goal. It plans a step, does it, checks the result, and repeats. On its own.

> IN PLAIN ENGLISH
>
> *It can go do a whole task while you grab coffee. This is the leap everyone means by "agent."*

REAL EXAMPLES

Claude Code | Claude Agent SDK

[■] WHAT CAL DOES HERE

- LIVE — Nine autonomous agents, plus Drop: co-branded listing alerts that send themselves.
- BUILDING — An agent that chases every open item on a loan all the way to done.

LAYER 06 / 08

L6 + Orchestration

One agent becomes a team. A manager splits the job up, others work in parallel, and the results merge back together.

IN PLAIN ENGLISH

One turns into a team that divides up the work.

REAL EXAMPLES

LangGraph | CrewAI | AutoGen

WHAT CAL DOES HERE

LIVE Those nine agents already run as a crew.

BUILDING A full loan pod that runs the busywork from application to closing.

LAYER 07 / 08

L7 + Skills & protocols

Reusable know-how and shared wiring, so any agent and any tool can plug into each other instead of being hand-built.

IN PLAIN ENGLISH

Shared plug-ins, so every tool speaks the same language.

REAL EXAMPLES

Agent Skills | MCP | A2A protocol

[■] WHAT CAL DOES HERE

LIVE We built the **first MCP server for mortgage**, the wiring that lets AI tools connect.

BUILDING Open Cal skills, so any loan system or lender can plug straight in.

LAYER 08 / 08

L8 + Guardrails

Audit trails, testing, permissions, and safety limits. The boring layer that makes AI safe for a real business.

> IN PLAIN ENGLISH
>
> *Seatbelts, cameras, and receipts, so a company can rely on it.*

REAL EXAMPLES

LangSmith | MLflow | Galileo

WHAT CAL DOES HERE

- **LIVE** A 75-case eval gates every release near 92 percent.
- **BUILDING** A live compliance check that blocks a bad answer before it ever ships.

WHERE THIS CAME FROM

From frozen text to *working agents.*

’17
’19

THE SPARK

The Transformer, then GPT-2. Raw scale alone produced general text. Impressive, but frozen and inert.

CAPABILITY · L1

’20
’22

EVERYONE MEETS AI

GPT-3, then ChatGPT put a chatbot in front of the whole world. Look-ups began fixing “confidently wrong.”

CAPABILITY · L1 → L2

’23
’24

TOOLS AND MEMORY

Function calling and code execution let models act. Memory arrived, and the first real agents showed up.

CAPABILITY · L3 → L4 → EARLY L5

’25
now

THE AGENT ERA

Agents went from demo to production. Teams of them, shared standards, and a whole safety industry appeared.

CAPABILITY · L5 → L8

I CALLED IT IN 2018

They said I was *crazy.*

My whole pitch was one line: **blockchain turns the mortgage into data.** I had no money, no network, and a room full of people telling me I was an idiot who would never be right.

FAST FORWARD TO 2026

$7.4B

Figure built exactly that. The blockchain lender went public in 2025, and that is its market cap today.

I was right. Just early, and broke. So I taught myself to build it: first Cal in 2022, the whole stack by 2025, out past the frontier now.

LAYER 09 · THE FRONTIER, RIGHT NOW

L9 = the frontier

This one is arriving right now, some of it brand new this week. Systems that improve themselves. Agents that hire other agents. Verified data that moves from one company to the next.

> IN PLAIN ENGLISH
> *The bleeding edge of what AI can do, and it moves every single week.*

A month ago this was research papers. This week, real products started shipping. We are already out here at the edge, with pieces live.

LAYER 10 / CLASSIFIED

L10 = [redacted]

While everyone else is still meeting agents, we are already building the layer after the frontier: [redacted] that turns the loan file into [redacted] nobody can fake.

> **THE TELL**
> *When the whole industry runs on it, remember who was early.*

We are not putting the receipts on a slide. **Not yet.** But you can be one of the first to see it.

WANT IN EARLY? DM ME "CAL."

SO WHAT ARE WE BUILDING?

The whole stack, on the *mortgage file.*

You just saw Cal across **all eight layers.** Some live today, more building. Not a chatbot bolted on the side, but the loan file turned into data an agent can read, check, and act on.

L1 BASE	L2 LOOK UP	L3 TOOLS	L4 MEMORY	L5 AGENT	L6 TEAM	L7 SKILLS	L8 TRUST

The agent does the busywork. The loan officer stays the human on top. **Software that makes them better, not software that replaces them.**

WHY THIS SCALES

The guidelines no human *will ever read.*

Every lender has dozens of loan products, each with its own fine print. That is **tens of thousands of pages** no human can read, so brokers default to the lender they know. The borrower rarely gets the best one.

CAL READS ALL OF IT

It matches a borrower to the genuinely best product, across every lender, in seconds.

LENDERS GET SEEN

They win business by being on the system, and push live pricing specials straight to it.

OFFICERS GET IT LIVE

The right product at the best price, in real time, at the point of sale.

One spine connecting every lender to every loan officer. **That is the part that scales.**

THE ENDGAME

Then lenders *bid for the loan.*

Once Cal runs the file through its own automated underwrite and weighs the credit risk, lenders compete for it in real time. The loan goes to the best offer, not the most familiar one.

WHAT COMES OUT OF THE PRICE

100+ bps

The cost of the roles automation makes obsolete comes out of the loan, and goes back to the borrower.

Cheaper loans, better matched, at scale. A real dent in housing affordability.

WHY THIS IS HARD TO COPY

The moat is not the *model.*

NOT A PILE OF PDFS

Every answer is cited, guideline-backed, and evaluated nightly. Solid enough to hand a borrower.

THE OPERATOR WHO CODES

Fifteen years at the desk, a top 1% loan originator in the nation, now built into the software. You cannot fake the domain.

IT COMPOUNDS

Every question and nightly refresh sharpens the brain. A newcomer starts at zero.

TRUST IS THE GATE

Built for the 2026 mortgage-AI rules before most of the industry finished reading them.

Everyone has the same models. Almost no one can run one in mortgage **and defend the answer.**

WHERE WE ARE RIGHT NOW

Live today, *building fast.*

ASKCAL

An AI copilot, live with early loan officers.

DROP

Co-branded listing alerts that send themselves.

THE BRAIN

Cited, guideline-backed, and evaluated nightly. Built to cover every lender.

The file is *data now.*

Chris Black · Building **Cal** · askcal.io · DM me “Cal”