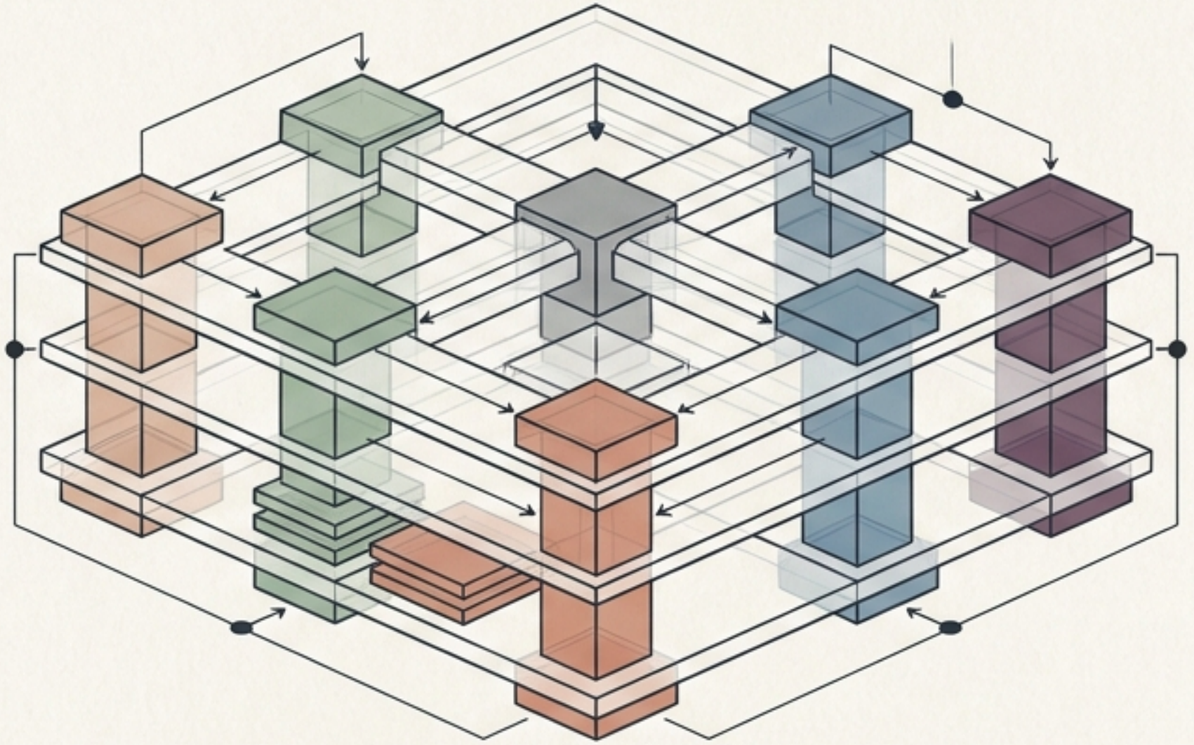
THE MODERN AI ECOSYSTEM

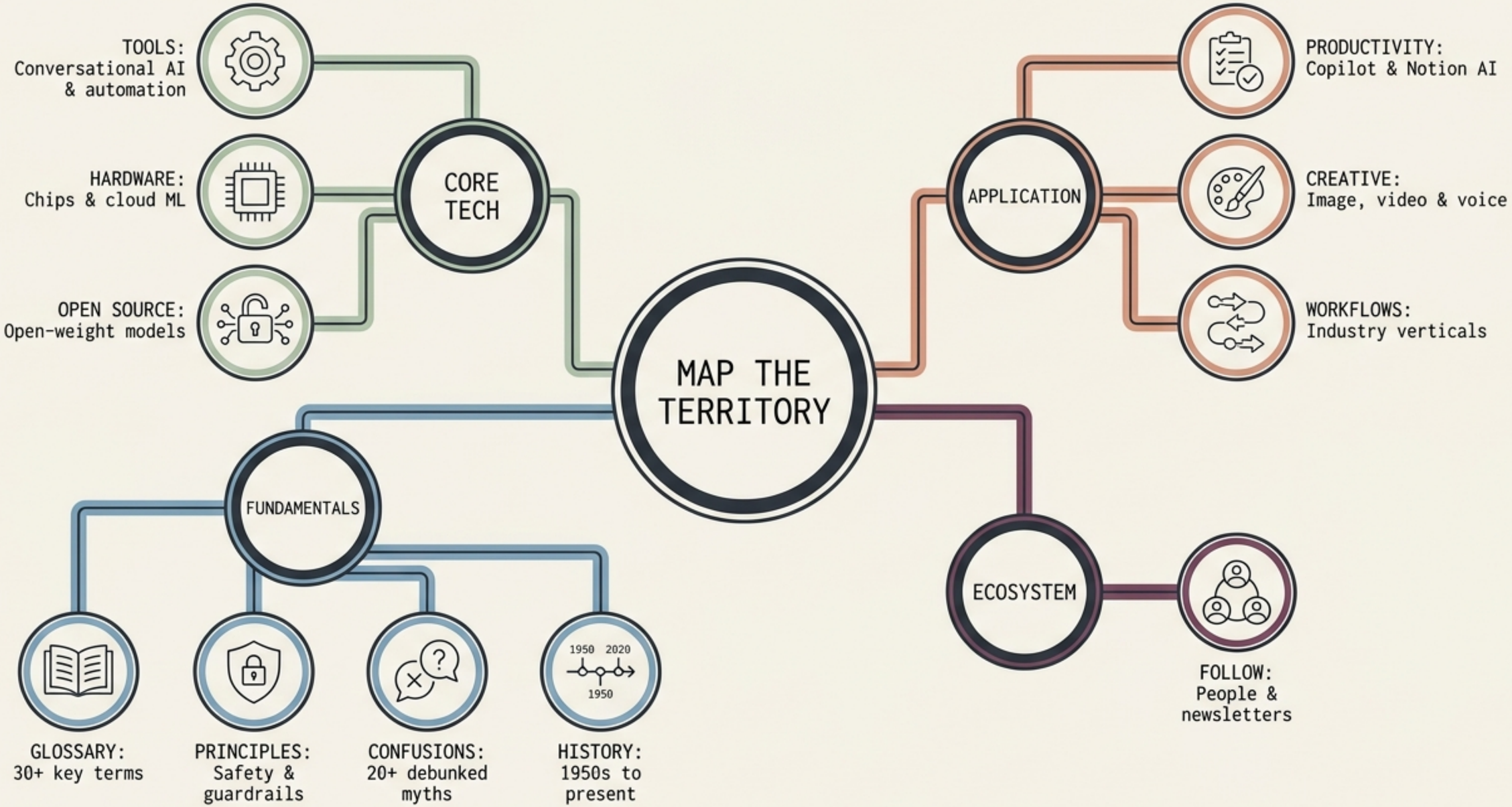
A structural playbook for tools, models, and workflows



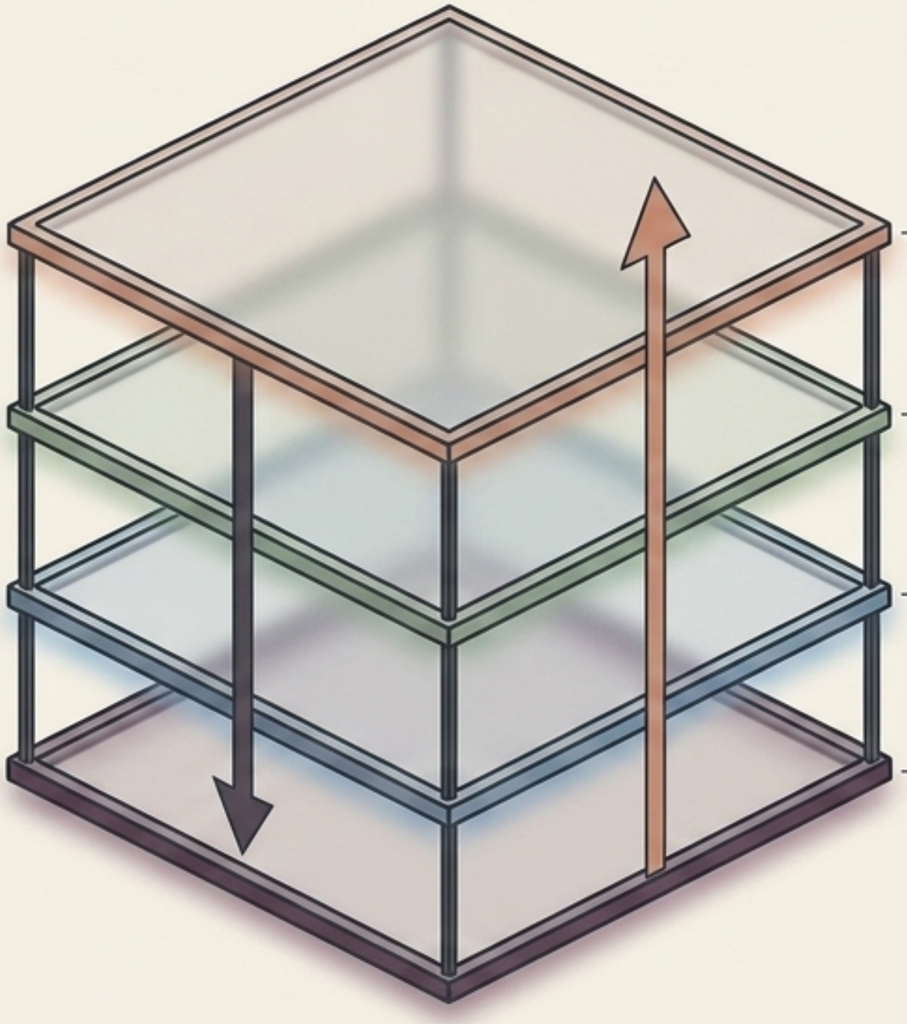
DISTILLED FROM SHUBHAM'S AI PLAYBOOK

THE MODERN STUDIO BLUEPRINT

2/28/2023 | Version 1.9



THE ANATOMY OF A MAJOR AI LAB



LAYER 4: PRODUCTS
The user interfaces. Chat web apps, direct APIs, and enterprise software integrations.

LAYER 3: MODEL TIERS
The engine variations. Segmented from fast & cheap to highly capable deep reasoning.

LAYER 2: MODEL FAMILY
The core algorithmic engine developed by the lab.

LAYER 1: RESEARCH LAB
The foundational organization, capital backing, and core scientific focus.

THE MODERN STUDIO BLUEPRINT

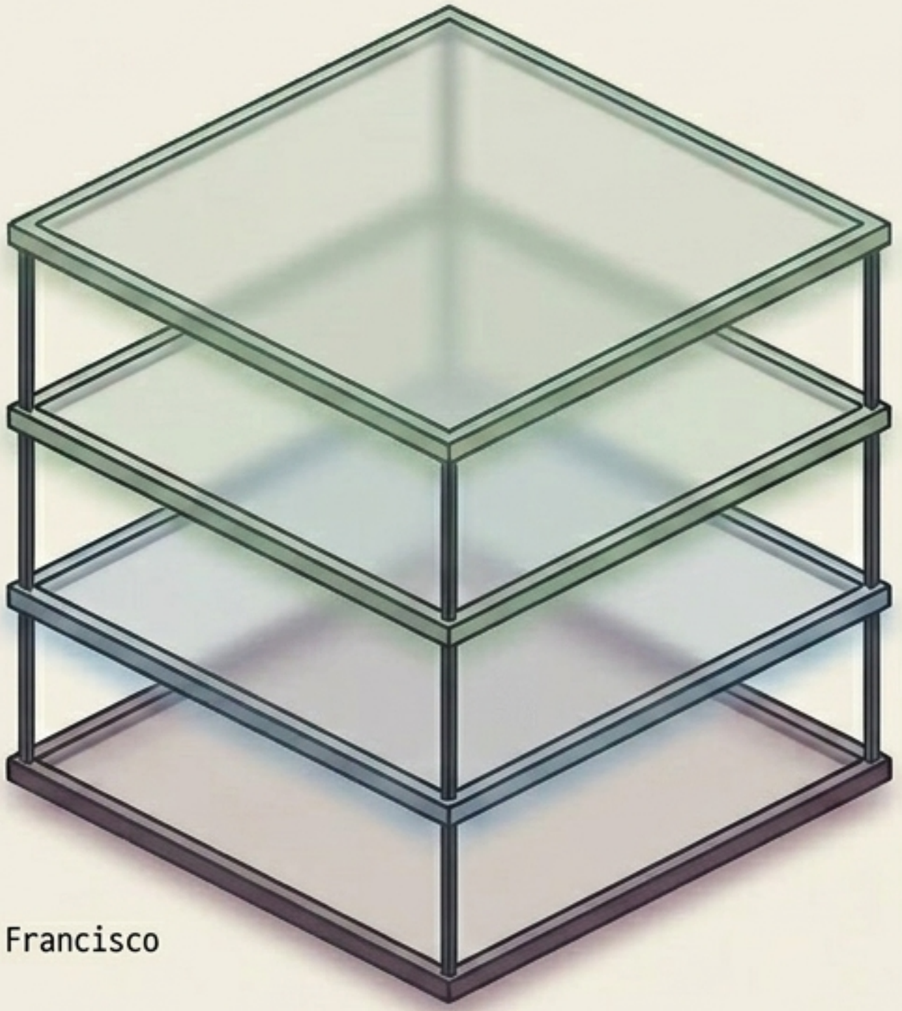
OPENAI FRAMEWORK

LAYER 4: PRODUCTS
ChatGPT (Web/Mobile), API, Codex (GitHub Copilot), DALL-E, Sora, Enterprise

LAYER 3: MODEL TIERS
GPT-4o mini (Fast) -> GPT-4o (Balanced) -> o1/o3/o4 (Deep reasoning)

LAYER 2: MODEL FAMILY
GPT series (General) + o-series (Reasoning)

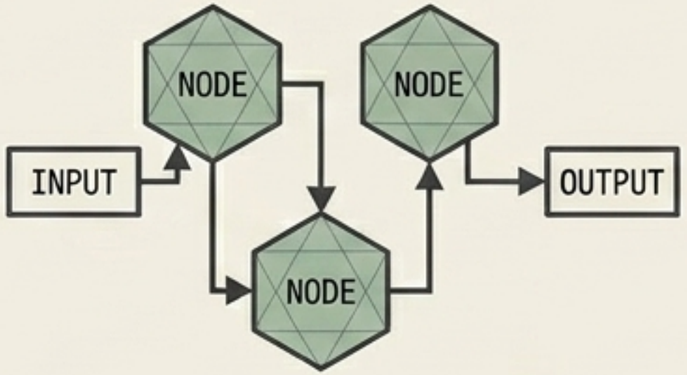
LAYER 1: RESEARCH LAB
The foundational organization,
Founded 2015 · Microsoft-backed · San Francisco



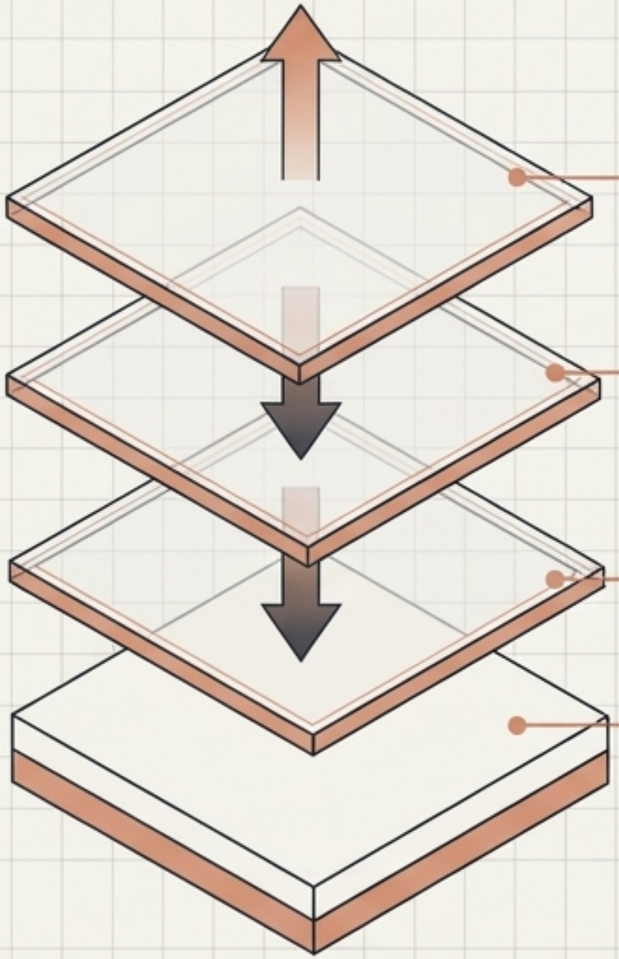
THE THINKING FLOW

STANDARD MODEL: FAST OUTPUT

REASONING MODEL (O-SERIES): HIDDEN CHAIN OF THOUGHT



ANTHROPIC FRAMEWORK



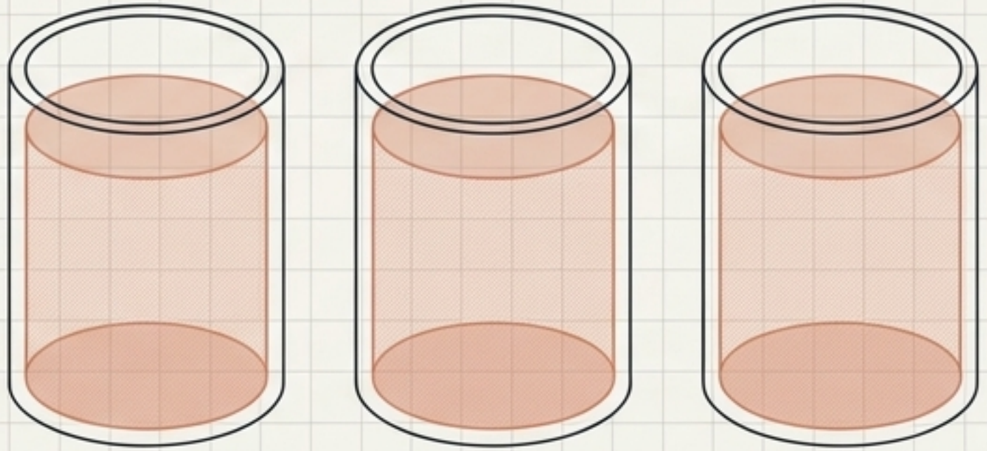
LAYER 4: PRODUCTS
claude.ai, API, Claude Code (Agentic CLI), Enterprise, Powers Cursor/Windsurf

LAYER 3: MODEL TIERS
Haiku (Fast & cheap) -> Sonnet (Balanced/Default) -> Opus (Most capable)

LAYER 2: MODEL FAMILY
Claude

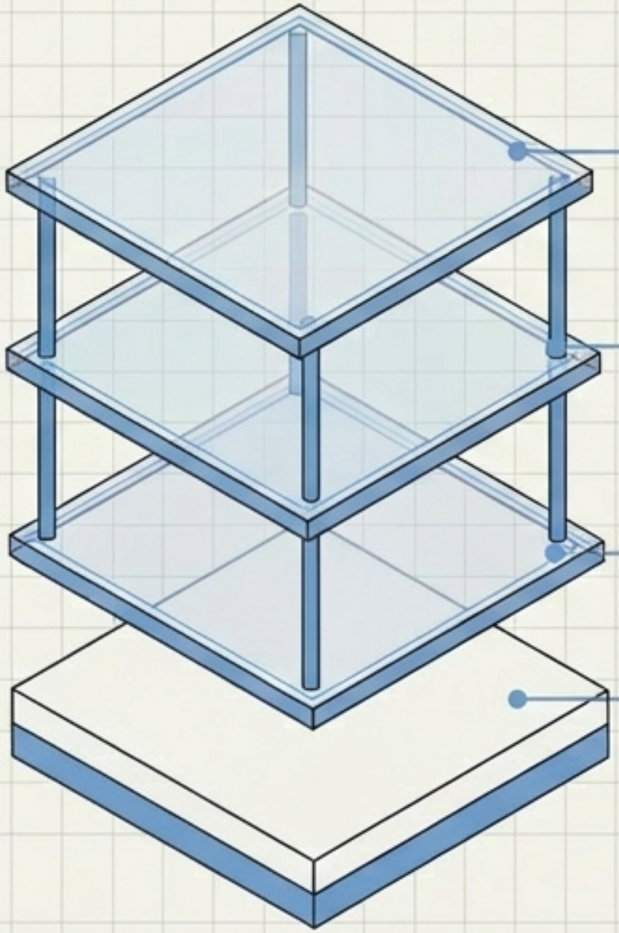
LAYER 1: RESEARCH LAB
Founded 2021 · Safety-focused · San Francisco

CONTEXT CONSISTENCY



Unlike competitors, all Anthropic tiers share the exact same massive 200K token context window.

GOOGLE DEEPMIND FRAMEWORK



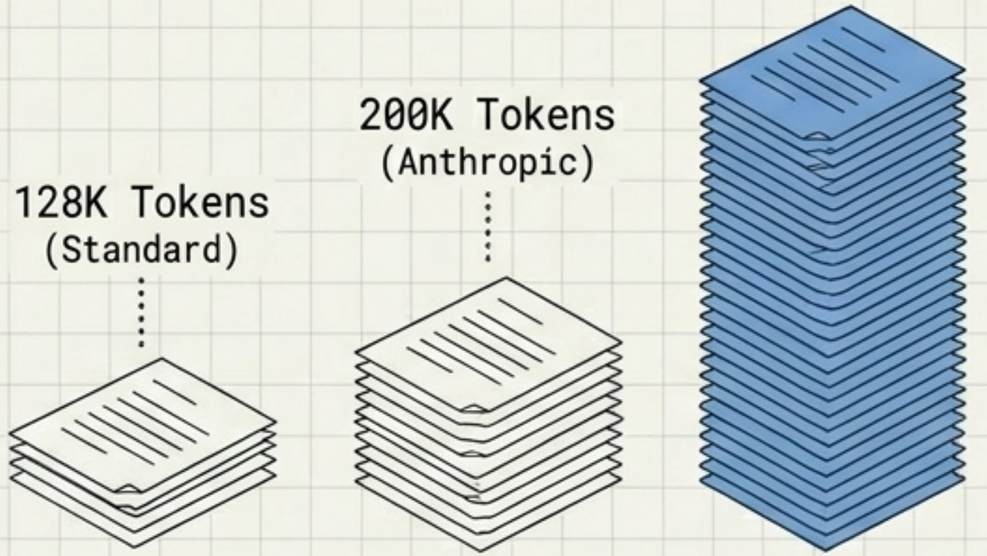
LAYER 4: PRODUCTS
Gemini.google.com, Google AI Studio, Vertex AI, Workspace & Android integration, Imagen

LAYER 3: MODEL TIERS
Flash (Fast & free)
-> Pro (Balanced)
-> Ultra (Most capable)

LAYER 2: MODEL FAMILY
Gemini + Imagen

LAYER 1: RESEARCH LAB
Founded 1998 · Alphabet subsidiary · Mountain View

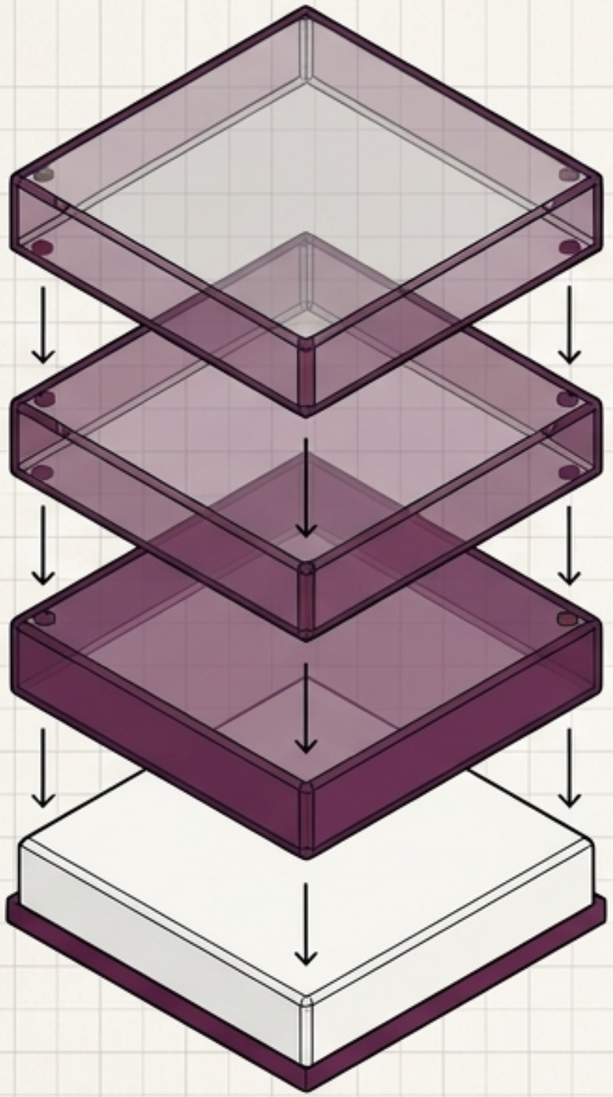
THE CONTEXT SCALE ADVANTAGE



The largest context window of any major model.

THE MODERN STUDIO BLUEPRINT
2/28/2023 | Version 1.9

DEEPSEEK FRAMEWORK



LAYER 4: PRODUCTS

chat.deepseek.com, DeepSeek API (Affordable), Open weights via Ollama/Hugging Face

LAYER 3: MODEL TIERS

V3 (General purpose) -> R1 (Reasoning, matches o1)

LAYER 2: MODEL FAMILY

DeepSeek R1 & V3

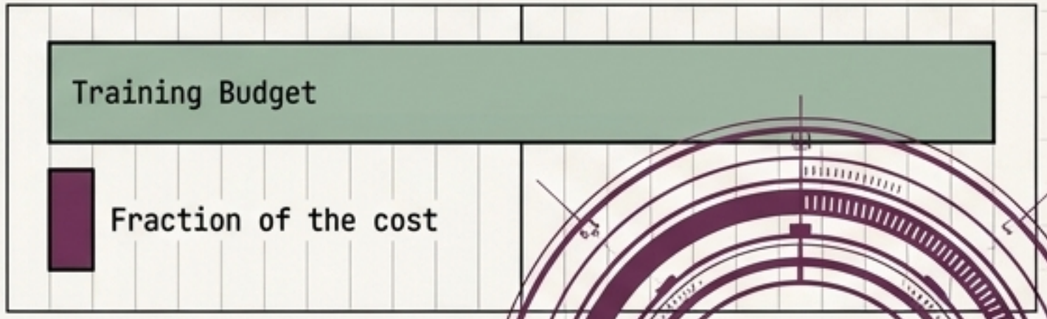
LAYER 1: RESEARCH LAB

Founded 2023 · Hangzhou, China · Open weights

THE OPEN DISRUPTION

Performance

Cost

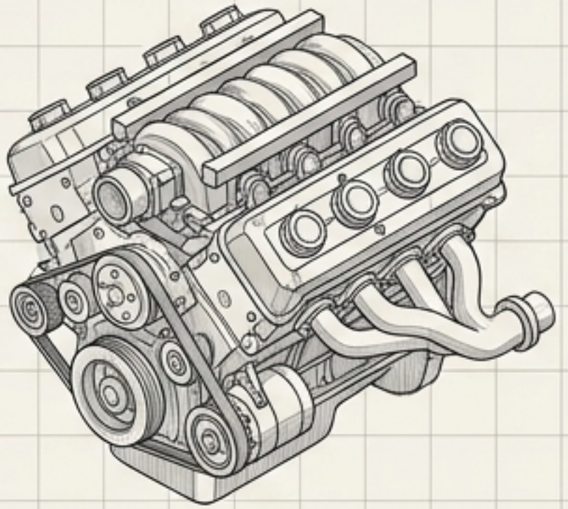


MIT LICENSE

(Free for commercial use)

THE OUTER ORBIT

OPEN WEIGHTS (THE ENGINES)

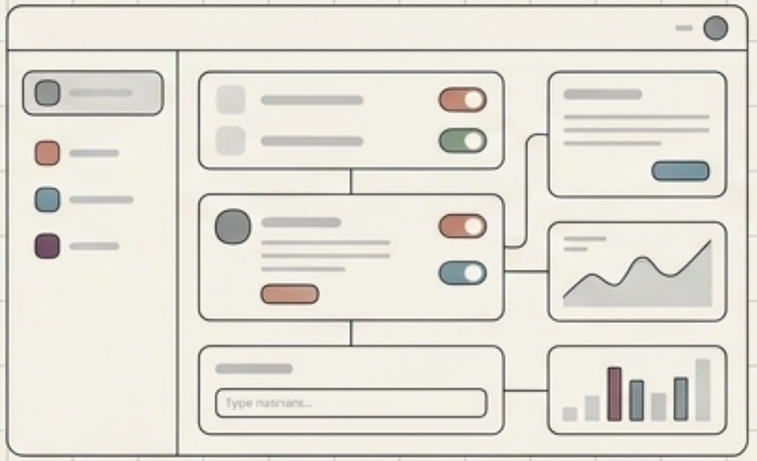


Release model weights publicly; no major consumer product. You run them locally or via hosts.

Models: Meta (Llama), Mistral, Google (Gemma), Microsoft (Phi-4), Yi

Access via: Ollama (Local), Hugging Face, Replicate, Together AI, LM Studio

SITS ON TOP (THE DASHBOARD)

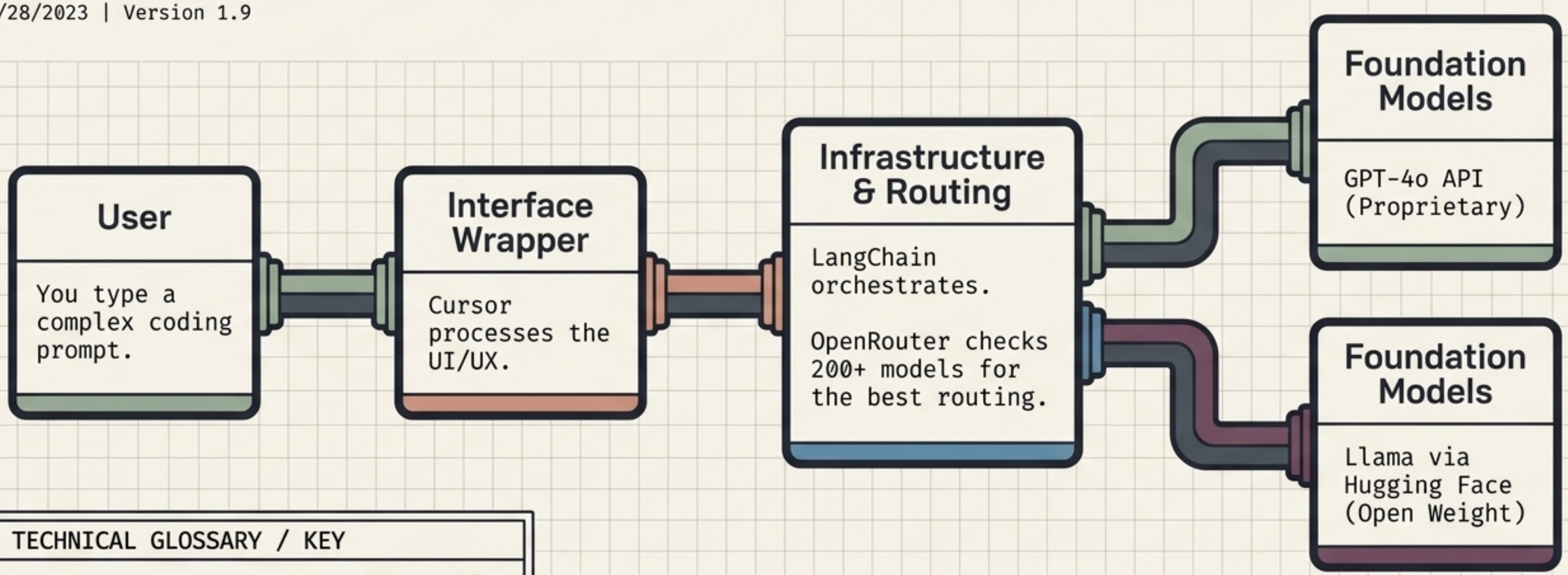


Call foundation APIs and wrap them in highly specialized user experiences.

- **Perplexity:** Claude/GPT-4o + Search Stack
- **Microsoft Copilot:** GPT-4o
- **GitHub Copilot:** Codex/GPT-4o
- **Cursor:** Sonnet or GPT-4o
- **Windsurf:** Claude or Codeium

ECOSYSTEM SYNTHESIS: HOW DATA FLOWS

THE MODERN STUDIO BLUEPRINT
2/28/2023 | Version 1.9



TECHNICAL GLOSSARY / KEY	
OpenRouter:	Route between 200+ models
LangChain:	Chain LLM calls/tools
LlamaIndex:	Connect private data
CrewAI:	Orchestrate AI agents

CORE CAPABILITIES & LIMITS

FEATURES	CHATGPT	CLAUDE	GEMINI	PERPLEXITY	OLLAMA
Context Window	128K	200K	1M	Varies	Varies
Web Browsing	Plugin	claude.ai	Native	Always On	No
Code Execution	Interpreter	Artifacts	Native	No	No
Free Limits	~10-15 msgs	~20-40 msgs	Generous	5 Pro searches	Infinite

ECONOMICS, PRIVACY & USE CASES

FEATURES	CHATGPT	CLAUDE	GEMINI	PERPLEXITY	OLLAMA
Paid Plans	$20/mo Plus	$20/mo Pro	$20/mo One	$20/mo Pro	Free forever
Privacy	Sent to servers	Sent to servers	Sent to servers	Sent to servers	100% Local
Best For	General tasks, plugins	Writing, long docs, safety	Research, very long context	Real-time research with citations	Privacy, offline, free automation

PRO TIP: If you hit free limits often, API access is usually cheaper for heavy use (pay per token, no daily caps).

THE MODERN STUDIO BLUEPRINT
2/28/2023 | Version 1.9

PHASE 1: ESTABLISH THE BASELINE

STEP 1: DAYS 1-7

Pick ONE and stick with it.
Volume over perfection.
Use for emails, summaries, explanations.

- ChatGPT (Widest ecosystem)
- ChatGPT (Widest ecosystem)
- Gemini (Best for Workspace)
- Claude (Best for writing)

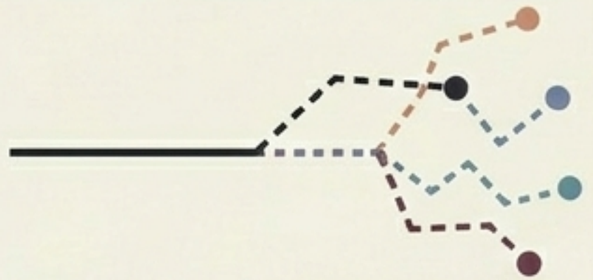
STEP 2: WEEK 2

Learn the Mechanics.
Spend 30 minutes on fundamentals.
Watch 3Blue1Brown's "But what is a GPT?"

TOKEN

4 Tokens

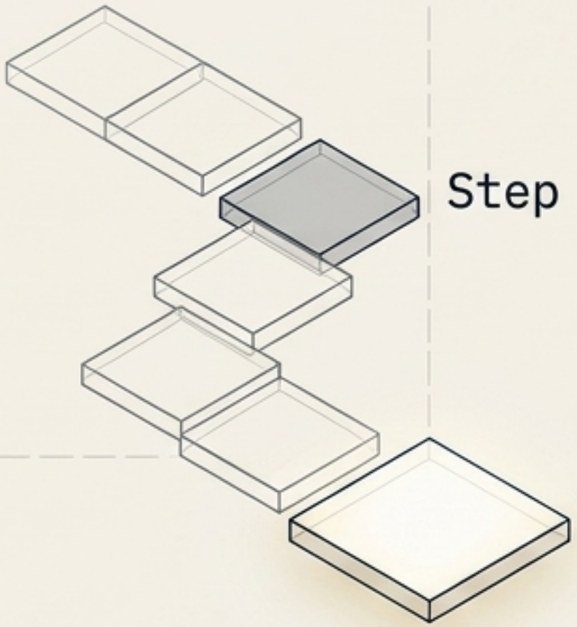
HALLUCINATION



Facts

CONFIDENT BUT INCORRECT OUTPUTS

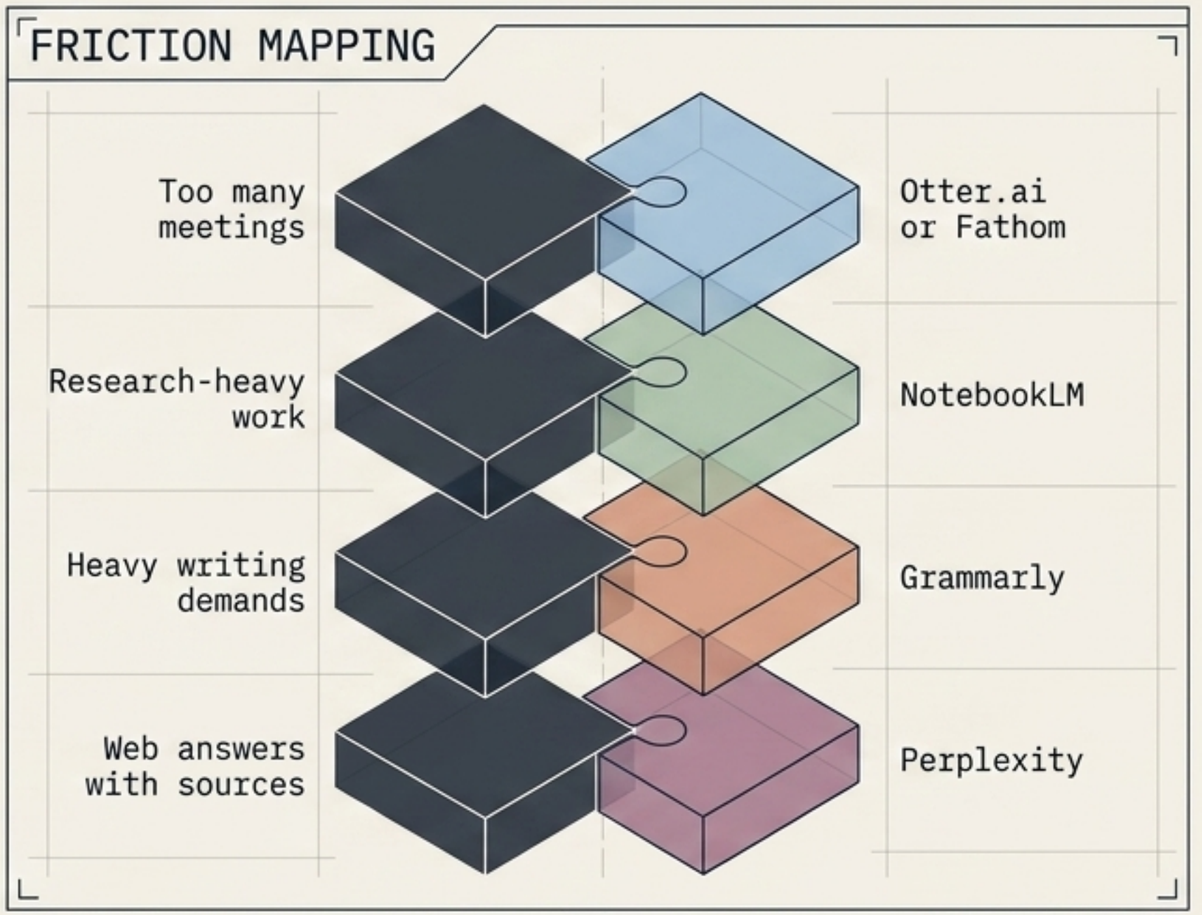
PHASE 2: TARGET SPECIFIC FRICTION



Step 2

STEP 3: WEEKS 3-4

Add ONE specialized tool for your biggest pain point. Don't sign up for everything.



The biggest mistake early adopters make: signing up for 10 tools in week one. Master one foundation model first.

MONTH 2: STAY LOOSELY INFORMED

AI moves too fast to read everything. Optimize for signal.

THE DAILY SCAN

Source: The Rundown AI

Format: 5-minute daily newsletter.

THE DEEP DIVE

Source: The Batch by Andrew Ng

Format: Weekly technical & industry depth.

THE PRACTICAL APPLICATION

Source: Ethan Mollick on X

Format: Real-world, academic-backed takes on using AI at work.

CHOOSE YOUR LANE

Creative Work

Midjourney → Runway → Suno

Building Products

Cursor → Lovable → n8n

Business Automation

Microsoft Copilot → Zapier

Under the Hood

fast.ai → Karpathy's YouTube

Privacy & Self-Hosting

Ollama → LM Studio → Run Llama

Ethics & Safety

Guardrails → Alignment → Principles

Return to Shubham's AI Playbook for deep dives into each lane.