MODULE 01 · AI EDUCATION SERIES

Understanding Modern AI

Models, Access, and How to Use Them

WHAT YOU GET

- Two types of AI models explained: closed (GPT, Claude, Gemini) vs open (Llama, Hermes, etc.)
- Three ways to access AI: subscribe UI, terminal/CLI, or custom bot
- Comparison table + how to choose a mode based on your user profile
- Case study: a real multi-mode AI stack used in daily creative work

Nicky Pandelaki

AI Engineer & Creator · Tangerang, 2026

Introduction

Generative AI (ChatGPT, Claude, Gemini, and the rest) has moved at lightning speed over the last three years. But most people are still unclear on two fundamentals.

- The types of AI models out there (closed source vs open source)
- How to access those models (subscribe UI, terminal/CLI, or custom app)

This module helps you understand the basic AI landscape of 2026 and choose the access path that fits your needs, whether you are a student, developer, creative professional, or business owner.

How to read this module

The module is split into 5 sections. Section 1 covers model types. Section 2 covers the three access modes. Section 3 presents a comparison table. Section 4 walks through a personal stack case study that combines all three modes. Section 5 is a practical guide to picking the right mode.

By the end, you will understand why multi-model and multi-mode is the most flexible and efficient way to extract maximum value from generative AI in 2026.

Section 1: Two Types of AI Models

Before we talk about access, you need to understand model types. There are two major camps in the AI world right now.

1.1 Closed Models

Models whose weights (the numbers inside the neural network that determine how it responds) are kept private by the company that built them. You can only use them through an API or the app they provide. You cannot download them and run them yourself.

Analogy: a fine dining restaurant. You order from the menu, eat on their premises, and pay each visit. The recipe stays inside the kitchen.

- GPT (OpenAI, San Francisco): most popular, strong at chat and general reasoning
- Claude (Anthropic, San Francisco): best at code, longform writing, careful reasoning
- Gemini (Google, Mountain View): best at multimodal (image, video, audio), longest context window

1.2 Open Models (Open Weights)

Models whose weights are released publicly by the company that built them. You can download the model file, run it on your own server or laptop, modify it, and deploy it privately.

Analogy: buying a cookbook. You get the full recipes, cook at home, free forever, improvise however you want.

- Llama (Meta): the most popular open source base model, foundation of many derivatives
- Hermes (Nous Research): a Llama fine-tune that is more steerable with custom personalities
- Mistral (French lab): runs both open and closed line-ups, efficient and fast
- DeepSeek (China): strong at math and code, very cheap, R1 briefly shook Wall Street

1.3 Quick Trade-off

	Closed	Open
General smarts	Smarter (state of the art)	1-2 generations behind
Cost	Per request (token)	GPU upfront, cheap after
Privacy	Data goes through their servers	Data stays on your servers
Control	Limited (their policy)	Full
Setup	Easy (just use it)	Needs tech setup

Section 2: Three Ways to Access AI

Once you understand model types, the next layer is access. There are three main modes to use AI in everyday life and work.

2.1 Mode 1: Subscribe + Use the Built-in Interface

You subscribe to the AI company's website or app. You access it through a browser or mobile app and start chatting. All features are provided out of the box.

- chatgpt.com (OpenAI), around 20 USD per month
- claude.ai (Anthropic), around 20 USD per month
- gemini.google.com (Google), around 20 USD per month for AI Pro

Pros: setup in one minute with no coding, full UI features built-in, flat monthly payment that is predictable.
Cons: locked into their interface, no way to automate custom workflows, every chat is manual.
Best for: everyday use, creative work, quick research, brainstorming, learning.

2.2 Mode 2: API + Terminal / CLI

You get an API key from the company, install a CLI tool or use it from your own code. Every request is billed by token count.

- Claude Code (CLI from Anthropic, run Claude from your terminal)
- gemini-cli (Google CLI), use with subscription or API key
- OpenAI API via curl, Python, or Node
- Cursor (an IDE that uses Claude or GPT for daily coding)

Pros: automation (cron, scripts, triggers), integration into your work pipeline, programmable.
Cons: needs basic dev knowledge, pay per use, you handle errors yourself.
Best for: developers, automated workflows, batch processing, integration into existing pipelines.

2.3 Mode 3: Custom Bot or Your Own Application

Build a custom app that calls the AI API behind the scenes, then design the UI and UX yourself to match the need. Frontend for users, backend that calls the AI plus integrations to other tools like databases, email, calendar.

- A personal multi-tool assistant on Telegram, WhatsApp, or web
- Internal company tools: HR bot, sales agent, knowledge base
- White-label SaaS that uses AI under the hood
- Multi-step agent systems for complex workflows

Pros: UX exactly fits the need, multi-model orchestration, custom branding.
Cons: you have to build and maintain it, you pay for API plus hosting, highest complexity.
Best for: businesses with specific workflows, personal assistants, agent systems, SaaS products.

Section 3: Access Comparison Table

A summary of the trade-offs across the three access modes so you can compare side by side when choosing the right path for your needs.

	Mode 1 (Subscribe UI)	Mode 2 (API/CLI)	Mode 3 (Custom App)
Setup time	1 minute	30 minutes	Days to months
Skill level	Zero (just click)	Basic dev	Full stack
Payment	Flat monthly	Per token	API + hosting + dev
Automation	Limited	Full	Full + integrations
UX control	Zero	Partial	Full
Data privacy	Per their policy	Per their policy	Per your impl
Best profile	Daily personal	Developer workflow	Business / assistant

Rule of thumb

- Want fast top-tier results without any tech setup: subscribe to one closed lab's UI (GPT, Claude, or Gemini)
- Have dev skills and need automation: combine subscribe UI for exploration with API/CLI for production
- Business or team that needs big leverage with custom UX: build a custom app that orchestrates multiple models
- Privacy-sensitive or need deep customization: consider open models + self-hosting (Mode 2 or 3)

Section 4: Personal Stack Case Study

Here is a concrete example of how I use all three modes side by side in daily creative work, covering studio operations and content production.

Mode 1 (Subscribe UI) for quick exploration

Open gemini.google.com for fast brainstorming on creative ideas or topic analysis. Open claude.ai when I need to draft long-form writing or analyze complex documents. Chat manually, close the tab, move on. No setup, instant, predictable cost (monthly subscription).

Mode 2 (API + CLI) for automated production

Claude Code from the terminal for coding internal tools (Telegram bots, automation scripts, production pipelines). Gemini API for batch image generation (Nano Banana) when I need dozens of visuals per project. ModelArk API (Seedance 2.0) for batch video generation. None of this is doable through a subscribe UI because it needs scripts and loops.

Mode 3 (Custom App) for assistants and operations

I run several custom AI agents on Telegram that call Claude (Opus 4.7) under the hood, plus integrations to Google Calendar, Gmail, Drive, Sheets, Instagram, and internal databases. Each agent has its own persona and domain (personal assistant, creative production, HR ops, research and investment analysis). Multi-model behind the scenes: Gemini for image gen, Seedance for video gen, Claude for text orchestration, GPT as a chat fallback.

Key insights from this combination

- Mode 1 for learning + fast idea generation, used almost every day
- Mode 2 for individual technical productivity, accelerating batch automation
- Mode 3 for team leverage and operations, huge ROI once it is built
- Multi-model behind Mode 3 avoids vendor lock-in and gives the best result per task

Section 5: How to Choose Your Mode

It depends on your profile and the reason you are using AI. Here is a practical guide.

Students or beginners

Start with Mode 1. Subscribe to one (Gemini is the cheapest, Claude is best for writing, ChatGPT is the most popular). Use it to learn, research, brainstorm, or help with assignments. Once you are comfortable, explore Mode 2 if you want to automate something repetitive.

Developers or technical professionals

Mode 1 for quick access and brainstorming, Mode 2 for serious work (Cursor for daily coding, Claude Code for complex technical tasks, batch API calls for volume tasks). Explore Mode 3 for side projects or internal team tools.

Businesses, studios, or agencies

Mode 3 is a must for team productivity. Build custom agents that integrate with existing workflows (CRM, calendar, internal databases). The ROI is large because automation scales. Start with the single most repetitive use case, then expand.

Researchers or privacy-sensitive users

Consider open models plus Mode 2 or 3 with self-hosted infrastructure. Data never leaves your servers. Trade-off: upfront cost (GPU) and complexity, but privacy and control are absolute.

Closing

Multi-mode is like multi-model. Same principle: pick the right tool per task. Mode 1 for fast daily exploration. Mode 2 for developer automation. Mode 3 for custom products and assistants that need real leverage.

The key is to stay flexible rather than lock yourself into one path. The AI world of 2026 is moving fast, and the ability to switch between modes will be a real competitive advantage.

Nicky Pandelaki
AI Engineer & Creator