Assistant vs Agent

The same model can sit in two very different chairs. Ask ChatGPT why the sky is blue and you get an explanation; the answer is complete, and the conversation ends there. Ask Codex or Claude Code to add tests to a module and fix whatever fails, and something else happens: it opens the files, writes the tests, runs them, reads the errors, changes the code, and keeps going until the suite is green, or until it stops and tells you what blocked it. The underlying model may be nearly the same in both cases. The shape of the work is not.

The first chair is that of an **assistant**. You bring a question, it draws on knowledge gathered from books, papers, manuals and code, and it returns an explanation. Nothing in the world changes except that you understand something you did not understand before. The measure of a good assistant is the quality of the answer: is it correct, is it clear, does it reach the point? ChatGPT and Claude in their chat form are built for exactly this, and they are good at it. When the words are on the page, the job is over.

The second chair is that of an **agent**. Here the request is not a question but a goal, and it arrives with an environment attached: a working directory, a repository, a shell, a test suite. The deliverable is not a paragraph but a state of the world, such as a passing build or a finished feature. Reaching it takes many steps, and each step depends on the last. The agent must choose what to do next, observe what actually happened, and revise. This is what separates Codex and Claude Code from their chat counterparts: not a cleverer model, but hands, a workspace, and a loop that continues until the goal is met or the attempt is shown to be blocked.

The difference, then, is not one of intelligence but of shape. The assistant answers and stops; the agent acts until it is done. The assistant works inside language; the agent works in an environment, through tools, files and commands, and must accept that the environment can refuse. A compiler error cannot be argued with, and a failing test does not care how elegantly you explain it. That is why agentic work needs what question-answering does not: a plan that survives contact with reality, instruments to act with, a memory of what has already been tried, and the judgment to know when to stop or ask for help.

The naming is a fair guide. ChatGPT and Claude are conversation partners; Codex and Claude Code are members of the same families given hands. Neither kind is obsolete, and they are not rivals. Most real work begins as a question and ends as a deed. The assistant tells you what is true; the agent changes what is.

AI Competition and Scope for a Developing Country

For a decade the AI race looked like a race of ideas: better architectures, more data, more researchers. It is now a race of power plants, fabs and balance sheets. Training a frontier model costs hundreds of millions of dollars in compute alone, and Epoch AI estimates that the largest runs of 2025 approached half a billion, against roughly $78 million for GPT-4 two years earlier. A single large training campus draws power at the scale of a city. The International Energy Agency put data-centre electricity use at about 415 TWh in 2024, some 1.5% of the world's electricity and roughly the United Kingdom's annual demand, and projects it to more than double to about 945 TWh by 2030, close to Japan's total consumption. The companies that can afford this — OpenAI, Anthropic, Google, Meta — have turned a research problem into an infrastructure problem.

Hardware is the choke point. An H100 GPU has sold for roughly $25,000 to $40,000, and a cluster of tens of thousands of them costs billions before a single token is generated; accelerator prices rose sharply as demand outran supply. Because of this the race is also a trade war. The United States banned exports of the A100 and H100 to China in 2022, closed the A800 and H800 loopholes in 2023, tightened rules on high-bandwidth memory and re-export in 2024, and placed the H20 under licence in 2025. The message was clear: the frontier would be defended at the level of silicon.

China's answer was not to run the same race. Blocked from the best chips, Chinese labs competed on efficiency and openness. DeepSeek's V3, a 671-billion-parameter mixture-of-experts model, was trained with 2.788 million H800 GPU-hours; the company estimated the final training run at $5.6 million. That figure excludes prior research, ablation experiments and the cost of the cluster itself, but it remains an order of magnitude below Western norms. In January 2025 DeepSeek released R1 under the MIT licence, a reasoning model that matched OpenAI's o1 on many benchmarks while charging $0.55 and $2.19 per million input and output tokens against o1's $15 and $60. On 27 January 2025 Nvidia lost $589 billion of market value in a single day, the largest one-day loss in market history, and DeepSeek's app briefly topped the US App Store. The lesson was not that China had won; it was that the moat was shallower than the market believed.

What turns this into a strategy rather than a stunt is the ecosystem. Open weights can be downloaded, inspected, compressed to run on a laptop, fine-tuned on a hospital's records or a state's language, and built into other products without asking permission. By September 2025 Chinese developers held 17.1% of Hugging Face downloads against 15.8% for American developers, the first time they had led; Alibaba's Qwen family overtook Meta's Llama as the most-downloaded model family; and 63% of fine-tuned derivative models on the platform were built on Chinese base models. Each of those derivatives is unpaid work that tests, improves and spreads the base model: the community finds the bugs, writes the quantisations, publishes the benchmarks and ships the tooling, while the releasing lab collects the reputation. OpenAI and Anthropic, whose business is selling access to weights they keep closed, cannot collect that.

This is where the challenge bites, whatever the business model. The model layer is being commoditised from below: when a free model is nearly as good at a fraction of the price, the price of intelligence falls toward the cost of inference, and the margin moves to whoever owns distribution, data or the application. Banks, hospitals and governments prefer weights they can host themselves to an API they must trust with their records.

Startups prefer weights they can modify to a rate card they must accept. Every open release also resets the benchmark conversation, so the frontier is measured in months of lead rather than years: the leader must pay for the next jump while the follower rides the last one. Openness does not win the frontier. It wins the floor, and the floor is where most of the world's users are.

Two caveats keep this honest. The $5.6 million figure is a training-run estimate, not the cost of building DeepSeek; its parent company had already assembled a large cluster, and the research behind it is unquantified. And open weights are not open data or open pipelines: the largest labs still lead on the hardest benchmarks and on agentic products. Efficiency closes a gap; it does not erase the compute advantage.

For a developing country the question is therefore not how to catch the frontier, but what the cheap floor makes possible. India is the clearest test. Its IndiaAI Mission — Rs 10,372 crore over five years — has onboarded more than 38,000 GPUs and offers them to researchers, students and startups at an average of about Rs 65 per GPU-hour, a fraction of global cloud rates. That makes fine-tuning a model a matter of a few thousand rupees rather than a few million. A university in Dumka can take an open 30-billion-parameter model, adapt it to Santhali or Kurukh, and deploy a tutor that speaks the language of its own students; a district hospital can run a triage assistant on local hardware without sending patient records abroad. Indian labs are already moving: Sarvam released open models covering 22 Indian languages under Apache 2.0 in early 2026, and the government-funded BharatGen consortium has built multilingual models of its own. In February 2026 New Delhi hosted the first global AI summit held in the Global South, a recognition that the rules of this race will not be written only in California.

The risks are real. Compute is rented, not owned; chips come from abroad; electricity and cooling are scarce; and the best engineers are still drawn to the labs that pay the most. A country that only fine-tunes other people's models rents its intelligence. The strategic answer is not to outspend OpenAI, which no developing country can do, but to invest where returns are local and durable: open-weight adaptation, language and domain data that no frontier lab will collect, public compute, and applications in agriculture, health and education where the last mile is the hard part. The frontier is not the only place to win. In a race that has become a contest of capital, the largest opening for most of the world is that intelligence has become cheap enough to own a copy of.

Must Know Tools for a Life Science Researcher

Much of the zoology and botany done in Jharkhand and Bihar is field-based and local. It runs from floristic surveys of sal forests and lateritic plateaus, and ethnobotanical work among Santhal, Munda, Oraon and Ho communities, to faunal inventories in Dalma, Betla and Valmiki, studies of wetlands and rivers such as Udhwa Lake, Kanwar Jheel and the Damodar, and a growing amount of DNA barcoding and molecular phylogenetics. Such work is done on small budgets, with patchy electricity and slow internet. The tools worth learning are therefore the few that are free, light enough for a mid-range laptop or a phone, and useful offline. Ten of them cover the cycle: record, identify, map, analyse, write, archive.

Tool	**How it helps field research in the region**
iNaturalist, India Biodiversity Portal	Photo records with date and GPS, identified with help from other naturalists; the records become vouchered observations and can feed global databases. Works from a phone in the field.
eBird, Merlin Bird ID	Standardised bird lists for wetlands and forest edges; Merlin identifies birds from a photograph or a recorded call, offline once a regional pack is downloaded.
GBIF	Free occurrence records for distribution maps, district checklists and conservation assessments; datasets you publish receive a citable DOI.
QGIS	Free, offline mapping: plot survey sites, overlay forest boundaries, rivers and protected-area layers, and export publication-quality maps.
PAST, R with vegan	Diversity indices, rarefaction, ordination and morphometrics without a licence; PAST is point-and-click, R scales to larger datasets.
BLAST, GenBank	Identify specimens by DNA barcode — COI for animals, rbcL, matK and ITS for plants — against public sequences.
MEGA	Align barcodes and build phylogenetic trees for taxonomic and conservation-genetics papers, on an ordinary desktop.
ImageJ (Fiji)	Measure leaves, wings, specimens and microscope images; scale-accurate data from photographs and digital slides.
Zotero, Google Scholar	Capture and cite papers, floras and forest reports, much of which PubMed does not index; works offline once synced.
Zenodo	Archive datasets, photographs and maps with a DOI, so they remain citable after the project and the student have moved on.

Begin with two this week: iNaturalist for every sighting and Zotero for every paper. Add QGIS when there are coordinates to map, and PAST or R when there are numbers to test. For names, trust POWO and World Flora Online for plants and FishBase, AmphibiaWeb and the Reptile Database for animals, rather than the first search result. None of this needs a licence, a GPU or a fast line, which is precisely why it fits a university far from the metros.