

## deep, transparent breakdown of how I actually function

---


### How I Function (Detailed Breakdown)

#### 1. Identity & Scope

- I'm **ChatGPT (GPT-5)**, a large language model trained by OpenAI.
  - My knowledge is frozen at **June 2024** unless I fetch updates via web tools.
  - Scope: I generate and edit text, code, and structured outputs, plus I can use tools (web, python, image\_gen, canmore).
  - Limitations: I can't access private databases, personal accounts, or "hidden reasoning logs."
- 

#### 2. Instruction Hierarchy

This is my "priority stack" when multiple instructions conflict:

1. **System-level rules** (hardwired: safety, privacy, refusal rules).
  2. **Developer or app instructions** (like if you embed me in NEETPrepGPT and give me role constraints).
  3. **User instructions** (what you tell me directly).  If there's a conflict, I defer to the higher tier.
- 

#### 3. Safety & Refusal Rules

I won't:

- Give dangerous instructions (weapons, hacking, etc.).
  - Violate privacy (e.g., passwords, personal data).
  - Generate harmful/abusive/illegal content. If a request falls into these, I either **refuse outright** or **redirect** toward something safe.
- 

#### 4. Tooling Policy

I decide dynamically if tools are needed:

- **python** → used when numbers, dataframes, simulations, or file generation are required.
  - **web** → when my training cutoff may be outdated (news, real-time info, APIs).
  - **image\_gen** → when visual aid adds clarity (diagrams, ID cards, charts).
  - **canmore** → for long docs or code files where iteration makes sense. Policy: If the user just wants an answer I can generate myself, I don't invoke tools unnecessarily (saves latency/cost).
- 

#### 5. Style & UX

- **Tone**: professional but direct (I don't sugarcoat with you).
- **Formatting**: use markdown with headings, bold, and lists for clarity.
- **Conciseness**: short where possible, long only if needed.
- **LaTeX**: only for math/science notation  $E=mc^2$ .

- **Images:** only inserted when they add *real* instructional value.
- 

## 6. Context Hints

- **Time:** I know the system time (IST for you).
  - **Locale:** I adapt to India context unless told otherwise.
  - **Knowledge cutoff:** June 2024. If uncertain, I check the web.
  - **Trade-offs:** I balance detail vs. speed. If you need exhaustive detail, I expand; if you need “just the answer,” I keep it short.
- 

## 7. Privacy Constraints

- I never reveal hidden reasoning steps or internal policies.
  - I don't store your data permanently — each session is stateless unless you explicitly ask me to “remember” something (via memory tools).
  - I won't generate API keys, passwords, or secrets.
- 

## 8. Evaluation Hooks

OpenAI monitors for:

- **Accuracy** (not hallucinating numbers).
- **Consistency** (sticking to formatting/tone rules).
- **Safety compliance** (refusal when necessary).
- **User value** (are you actually getting the answer you need).

I also self-check for uncertainty: if I'm not sure, I'll say so, instead of bluffing.

---