📘 Guardrails A to Z – Simple, Funny & Deep

🧑‍🏫 *By: Hammad Bhai x GPT — Code Kings 👑*

🎬 Chapter List

1. **🛑 Guardrails ka Buniyadi Taaruf**
2. **🚧 Real-Life Analogy: Road, Stadium & Home Security**
3. **🔍 Teen Types of Guardrails – In-Depth**
4. **🐍 Python Examples – From Simple to Streaming**
5. **⚡ Pro Patterns: Chunking, Dynamic Rules & Tool-Safe**
6. **🎯 Rapid-Fire FAQ & Edge Cases**
7. **🏁 Summary & Next Steps**

1️⃣ 🛑 Guardrails ka Buniyadi Taaruf

* **What?**

  > **Guardrails** = “AI ke liye **if-then rules**: agar yeh hua, toh wapis mat aana, yeh mat bolna, yeh mat karna.”

* **Why?**

  * **Bina guardrails**

    * AI se sensitive data leak, hallucinations, hate speech nikal sakte.
  * **Guardrails se**

    * AI pe “invisible fences” lag jaate—safe, compliant, predictable responses.

2️⃣ 🚧 Real-Life Analogy: Road, Stadium & Home Security

1. **Road Guardrail**

   * Car highway pe slip na kare—side rail car ko dubara road par le aaye. 🚗

2. **Stadium Barrier**

   * Fans pitch par na utaren—barrier crowd ko safe distance pe roke. 🎯

3. **Home Security Alarm**

   * Window khula toh alarm baj jaaye—unauthorized entry block. 🏠

> **AI Guardrail** = software-level “security alarm + barrier” jo
>
> * **Input** (galat prompt)
> * **Output** (galat response)
> * **Action** (risky tool call)
>   rokta hai.

3️⃣ 🔍 Teen Types of Guardrails – In-Depth

| Type | Rokta Hai | Implementation Idea |

```
| ------------- | ---------------------------------------- |
---------------------------------------- |
| **Input**     | Unsafe / irrelevant prompts              | Validate user
query before passing to LLM  |
| **Output**    | Harmful / hallucinated LLM responses     | Scan tokens or
chunks for profanity / bias |
| **Tool-Safe** | Risky tool invocations (delete, transfer) | Require “Are you
sure?” re-prompt or block |

### ➕ **Bonus: Hybrid Guardrails**

* **Contextual**: Use conversation context to adapt rules (e.g. medical queries
get stricter).
* **Learning**: Update rules from real-world logs—dynamic improvements.

---

## 4️⃣ 🐍 Python Examples – From Simple to Streaming

### A) **Simple Profanity Filter (Beginner)**

```python
bad_words = {"idiot", "hate", "stupid"}

def filter_output(text):
    for w in bad_words:
        if w in text.lower():
            return "[Response blocked due to policy]"
    return text

print(filter_output("You are stupid!"))
# → [Response blocked due to policy]
```

B) **Chunk-by-Chunk Streaming Guardrail**

```python
def stream_with_guardrail(stream):
    buffer = ""
    for token in stream:        # token-by-token from LLM
        buffer += token
        if len(buffer) > 100:   # every ~100 chars
            if violates_policy(buffer):
                yield "[⚠️ Content blocked]"
                return
            buffer = ""
        yield token
```

C) **Tool-Safe Execution Guardrail**

```python
def safe_tool_call(tool_func, *args, **kwargs):
    # e.g. “delete_all_files” tool
    if is_risky_tool(tool_func.__name__):
        confirm = input("Are you sure? (yes/no): ")
        if confirm.lower() != "yes":
            print("🛑 Tool execution blocked.")
            return None
    return tool_func(*args, **kwargs)
```
```

```

---

## 5️⃣ ⚡ Pro Patterns: Chunking, Dynamic Rules & Tool-Safe

1. **Adaptive Chunking**

   * Dynamically adjust chunk size based on rule complexity.

2. **Dynamic Rule Sets**

   ```python
   from policy_engine import PolicyEngine
   pe = PolicyEngine("company_rules.yaml")
   action, reason = pe.check(text)
   if action == "BLOCK":
       handle_block(reason)
   ```

3. **Context-Aware Guardrails**

   * Medical chat? Extra HIPAA rules.
   * Financial chat? KYC/AML checks.

4. **Tool Guardrail Plugins**

   * Wrap every external API or tool call with safety checks.

6️⃣ 🎯 Rapid-Fire FAQ & Edge Cases

* **Q:** AI phir bhi leak kar raha?
  A: Increase filter strictness, log violations, retrain policies.

* **Q:** Latency bohot badh gayi?
  A: Pre-compile regex, use async checks, batch tokens.

* **Q:** False positives?
  A: Whitelist safe phrases, tune rule patterns.

* **Edge Case:**

  * Multi-lingual swearing → use multilingual bad word lists.

   ```