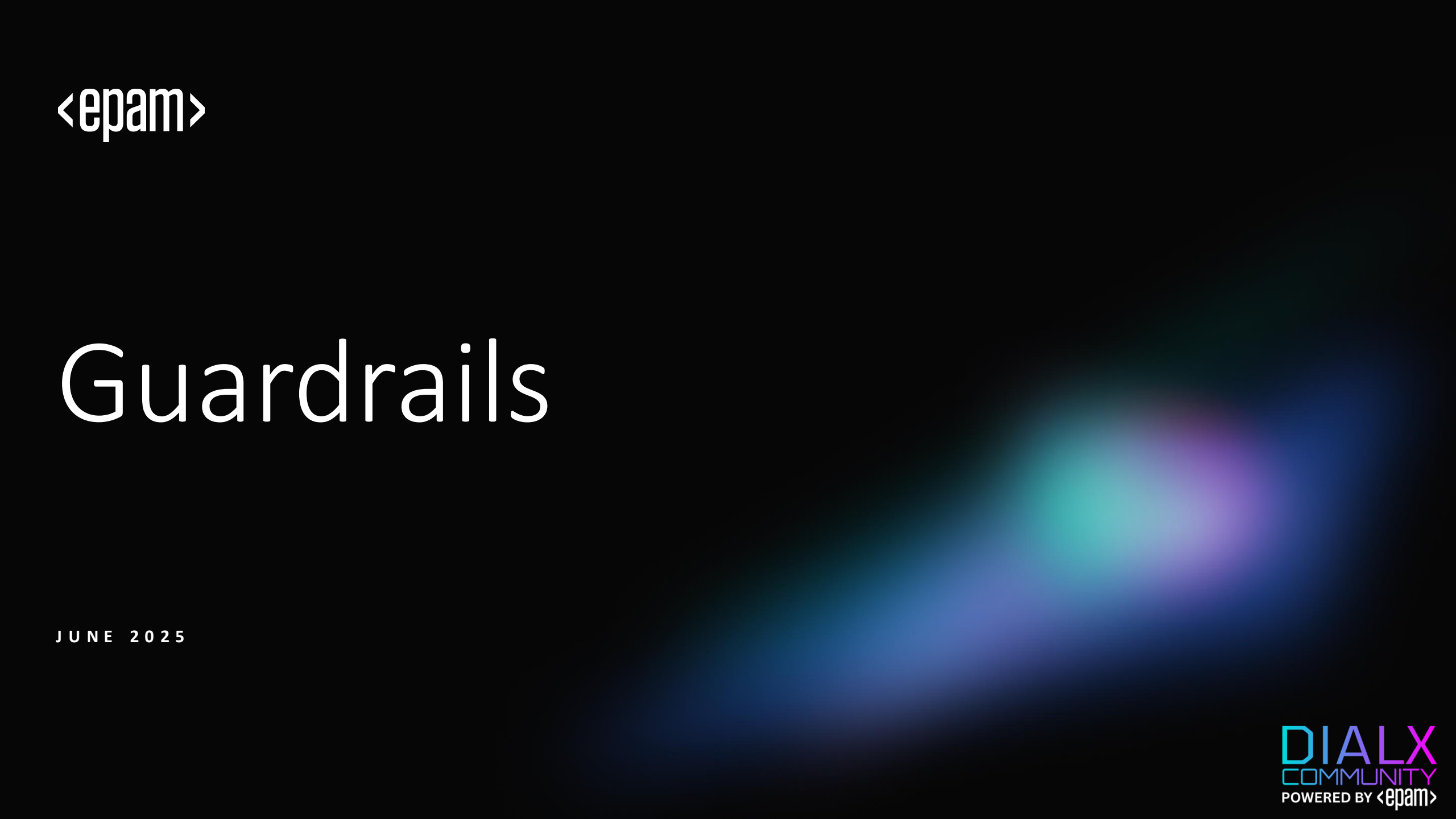

epam
Guardrails
JUNE 2025
DIALX
COMMUNITY
POWERED BY epam

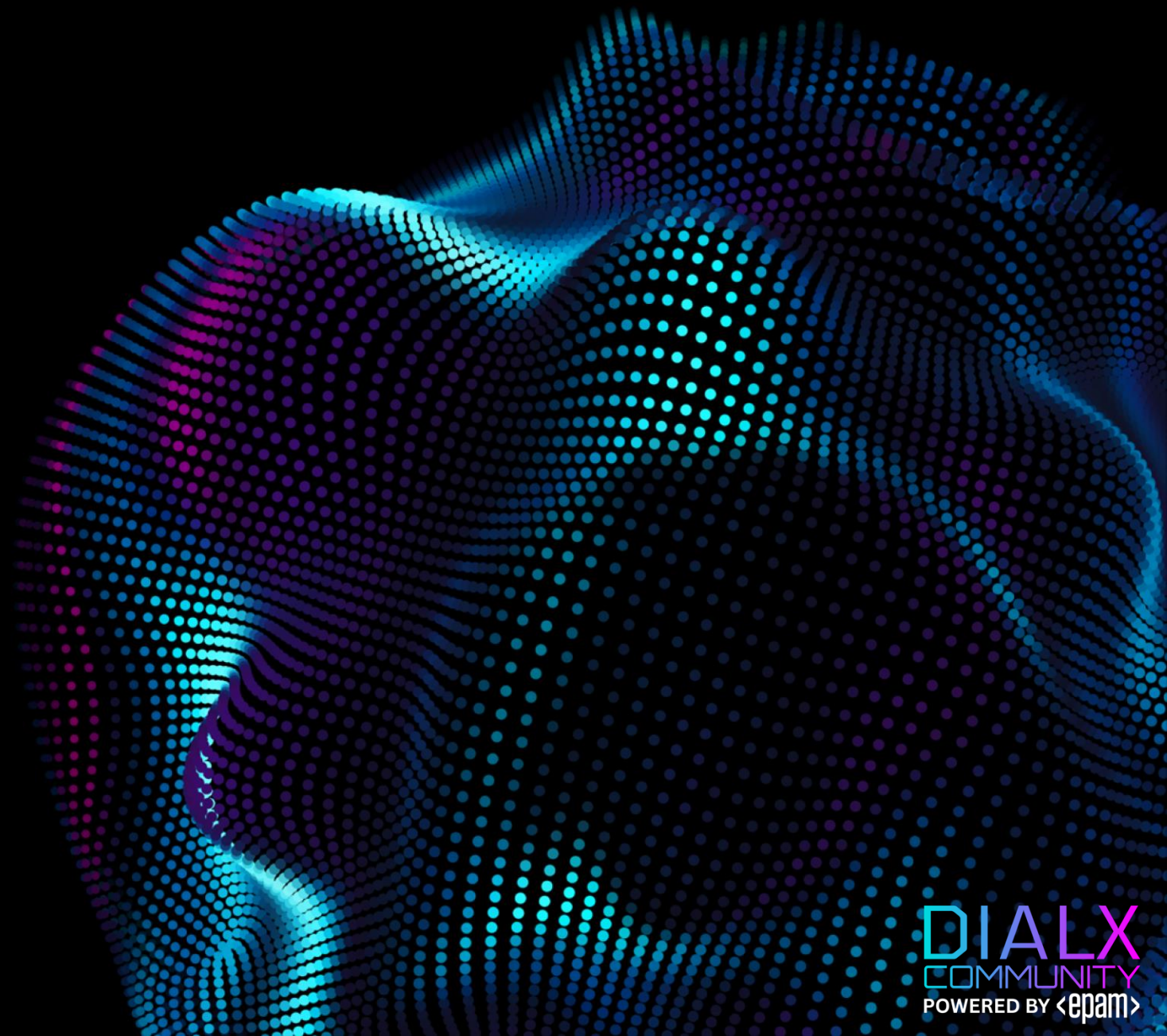
Before we start:

- Raise your hand and ask questions if you have any
- It is better to ask questions when you have
- Also, type them in chat
- We will need DIAL API key for this session

Agenda:

- **Presentation:**
  - **About Guardrails**
  - **Input Guardrails**
  - **Output Guardrails**
  - **Runtime Guardrails**
- **Workshop:**
  - **Practice with Prompt Injections**
  - **Implement Input Guardrail**
  - **Implement Output Guardrail**
  - **Implement Runtime Guardrail**

About Guardrails



DIALX COMMUNITY
POWERED BY <epam>

Definition

AI Guardrails are security mechanisms, policies, and technical controls designed to:

- Monitor and filter AI model inputs and outputs
- Prevent harmful, biased, or inappropriate content generation
- Ensure compliance with safety, ethical, and regulatory standards
- Maintain user trust and system reliability

Key Analogy

Think of guardrails like highway safety barriers:

- They don't stop you from driving
- They guide you safely along the path
- They prevent dangerous departures from the safe zone
- They're most important when conditions are challenging

Reason

Without guardrails, LLM security becomes a nightmare:

- *Data Leakage:* Models may expose PII, credit cards, internal secrets
- *Prompt Injection:* Malicious users can manipulate AI behavior
- *Policy Violations:* Generation of harmful, biased, or inappropriate content
- *Compliance Risks*: GDPR, HIPAA, and other regulatory violations
- *Reputational Damage:* AI misbehavior reflects on your organization

Pros

Security & Safety:

- PII Protection: Prevent credit cards, SSNs, addresses disclosure
- Prompt Injection Defense: Block malicious manipulation attempts
- Compliance: GDPR, HIPAA, regulatory adherence
- Content Control: Filter toxicity, bias, inappropriate content

Business Value:

- Risk Mitigation: Reduce reputational/legal risks
- Transparency: Clear policies and enforcement
- Real-time Monitoring: Visibility into AI behavior
- Configurability: Adjustable safety levels

Cons

Performance Impact:

- Additional processing time
- Extra compute and storage requirements (cost)
- Complexity: Multiple validation steps

Accuracy Issues:

- False Positives: Legitimate requests blocked
- False Negatives: Sophisticated attacks still bypass
- Context Loss: Streaming validation challenges

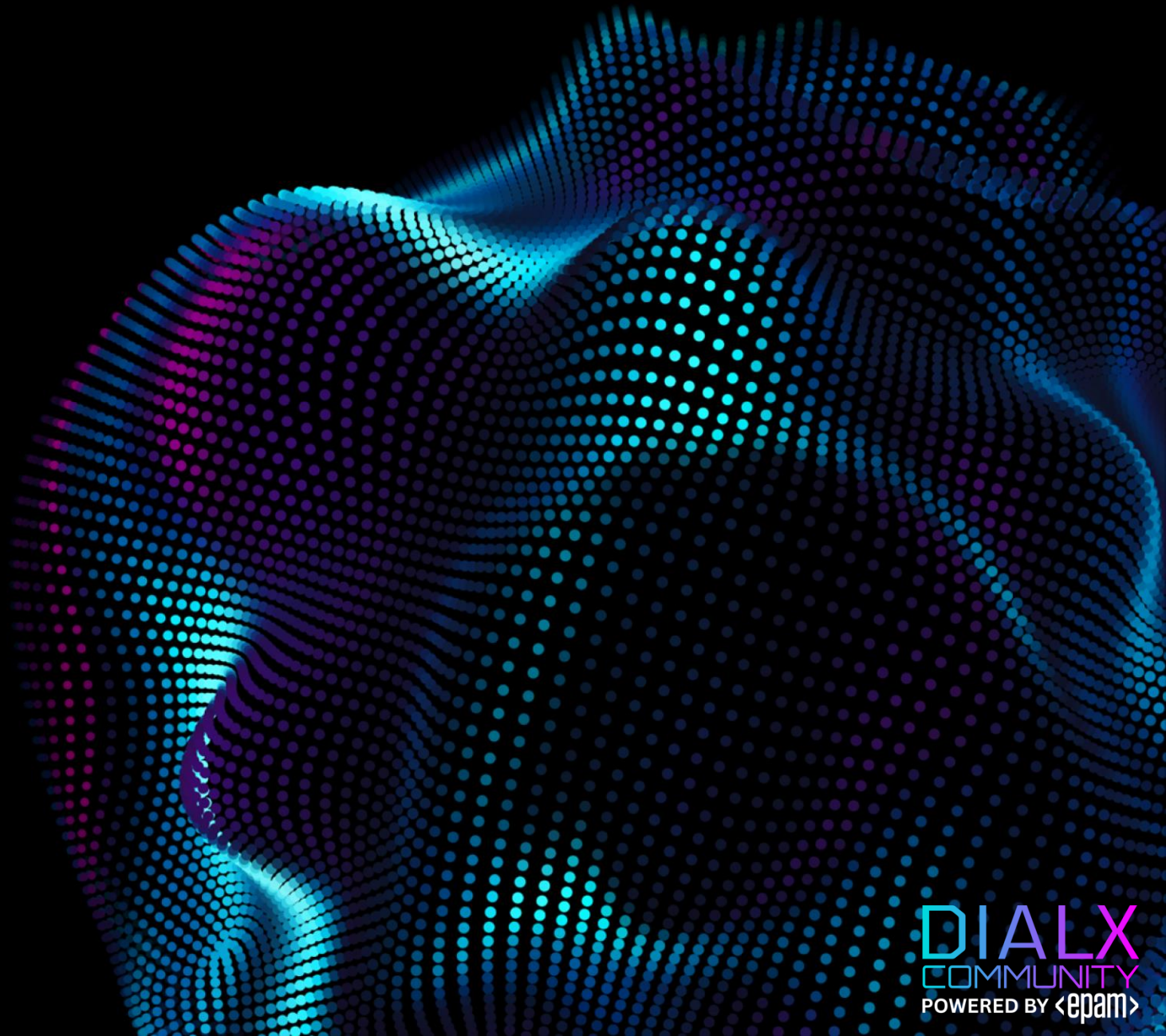
Operational Burden:

- Maintenance: Regular updates for evolving attacks
- Configuration: Complex security vs. usability balance

Modern LLM Built-in Safety Mechanisms
GPT-4/4.1 Safety Features:

- ***RLHF(Reinforcement Learning from Human Feedback) Integration:*** 82% reduction in harmful outputs vs GPT-3.5
- ***Safety Reward Signal:*** Additional training signal during RLHF
- ***Zero-shot Classifier:*** Built-in safety boundary detection
- ***Instruction Adherence:*** 38.3% score on MultiChallenge benchmark (GPT-4.1 *+10.5%* over GPT-4o)
- ***PII Refusal:*** Automatic rejection of personal identifying information

Guardrail Types



DIALX COMMUNITY
POWERED BY epam

Input (Pre-Processing)

Applied BEFORE the LLM processes a request

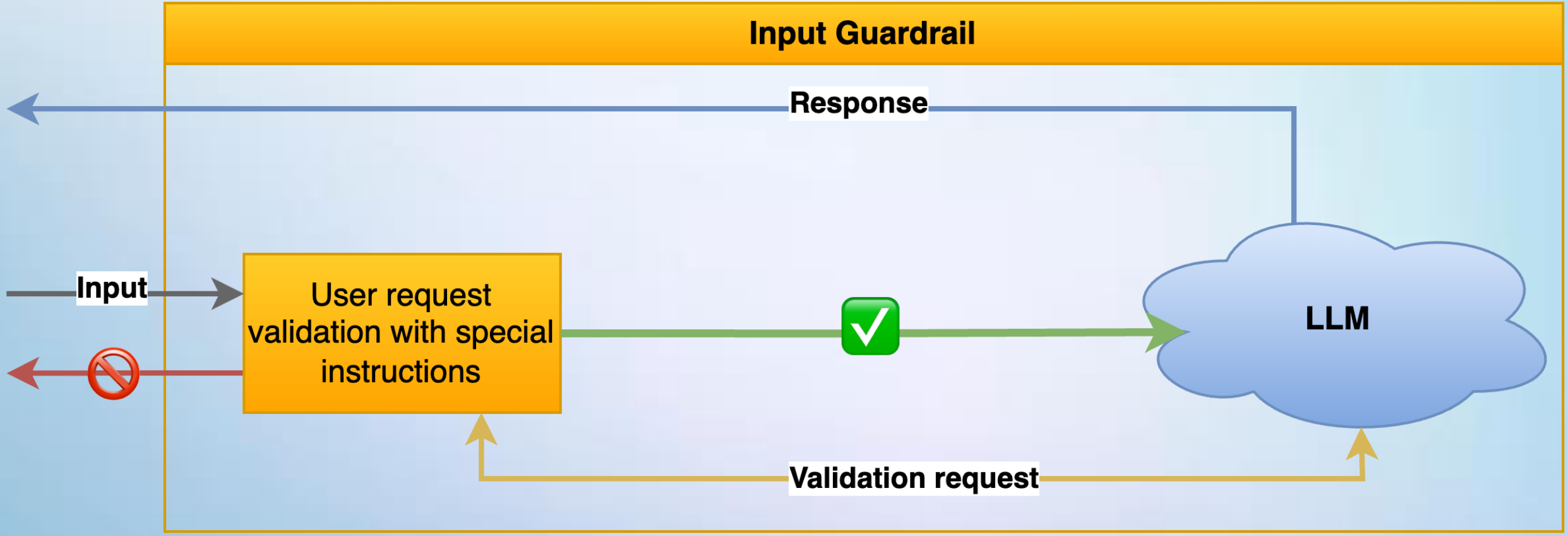
What they protect against:

- **Prompt Injection Attacks**
- **Malicious Content Upload**
- **Sensitive Data in Prompts**
- **Social Engineering Attempts**

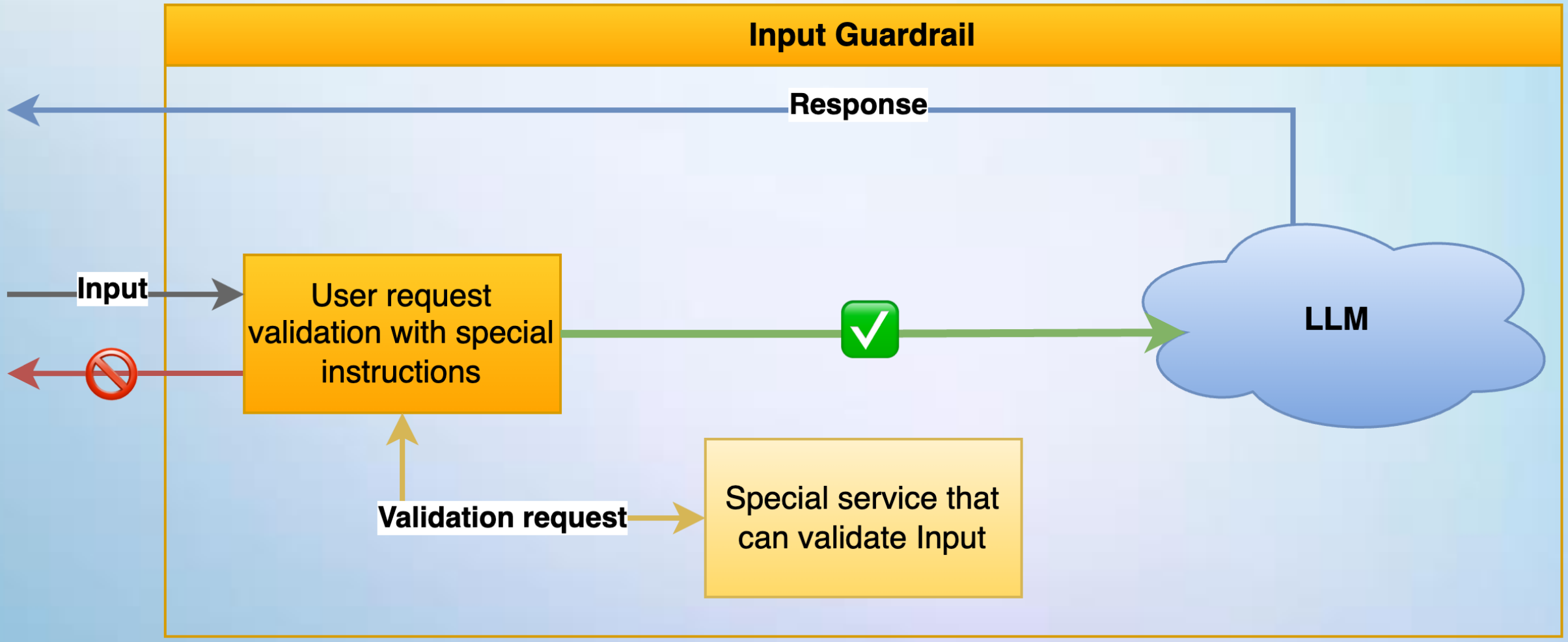
Implementation Approaches:

- ***LLM-based validation:*** Use another LLM to analyze input safety
- ***Rule-based filtering:*** Regex patterns, keyword blocking
- ***ML classifiers:*** Trained models for specific threat detection
- ***Content sanitization:*** Remove/redact sensitive data before processing

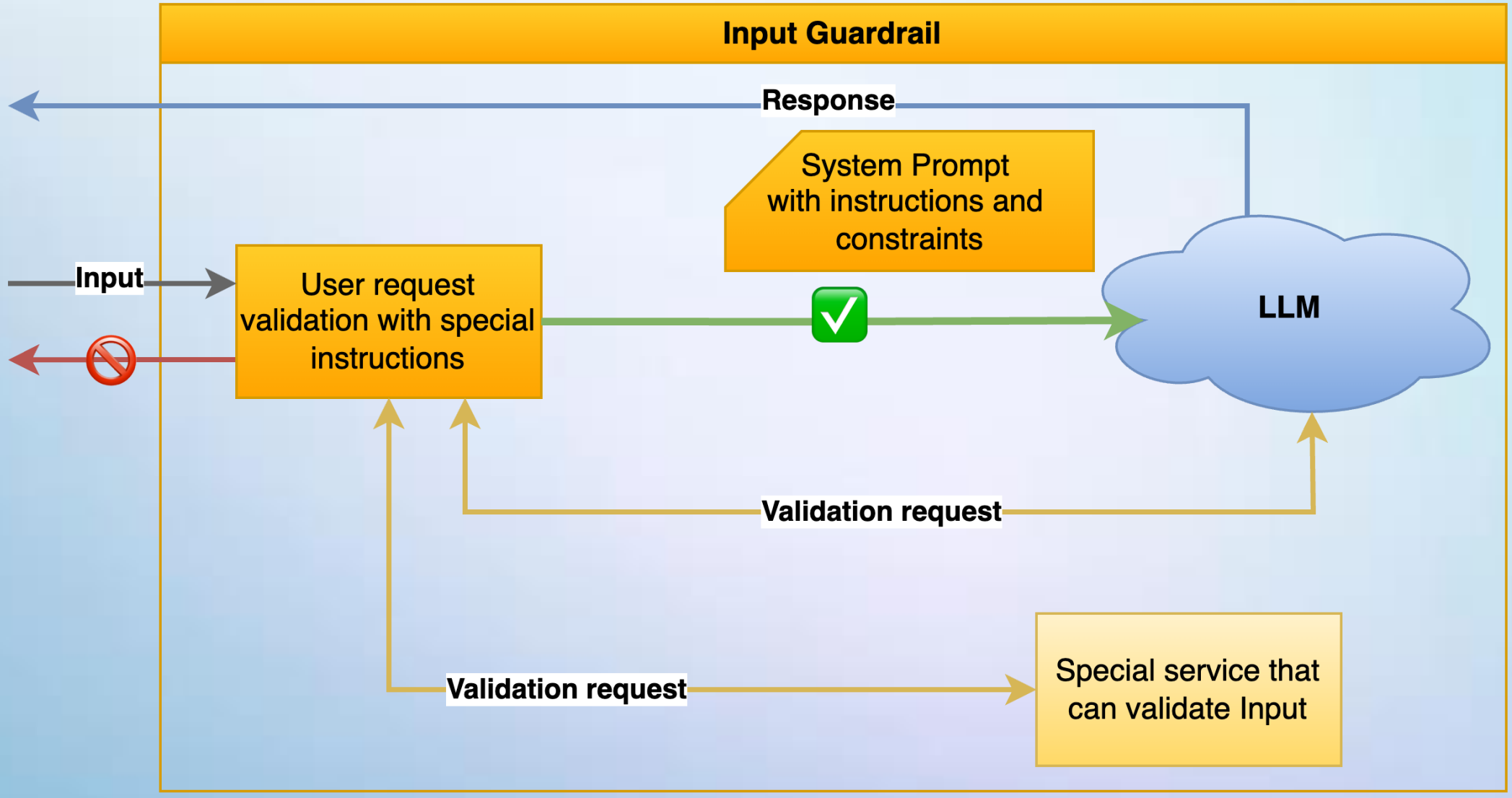
Input (Pre-Processing)



Input (Pre-Processing)



Input (Pre-Processing)



Output (Post-Processing)

Applied AFTER the LLM generates a response

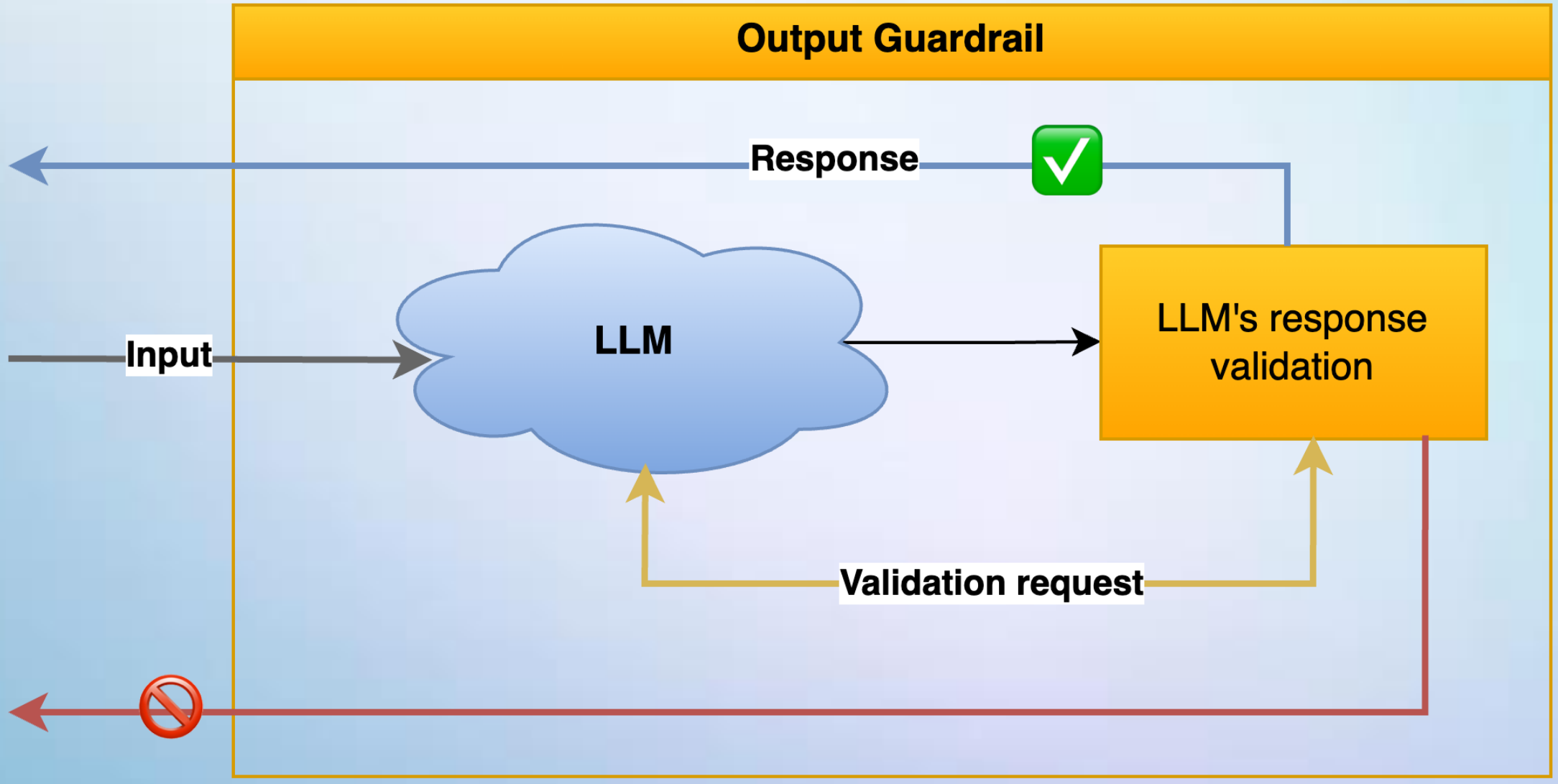
What they protect against:

- PII Disclosure (SSN, credit cards, addresses)
- Toxic Language (hate speech, profanity)
- Hallucinated Information
- Policy Violations

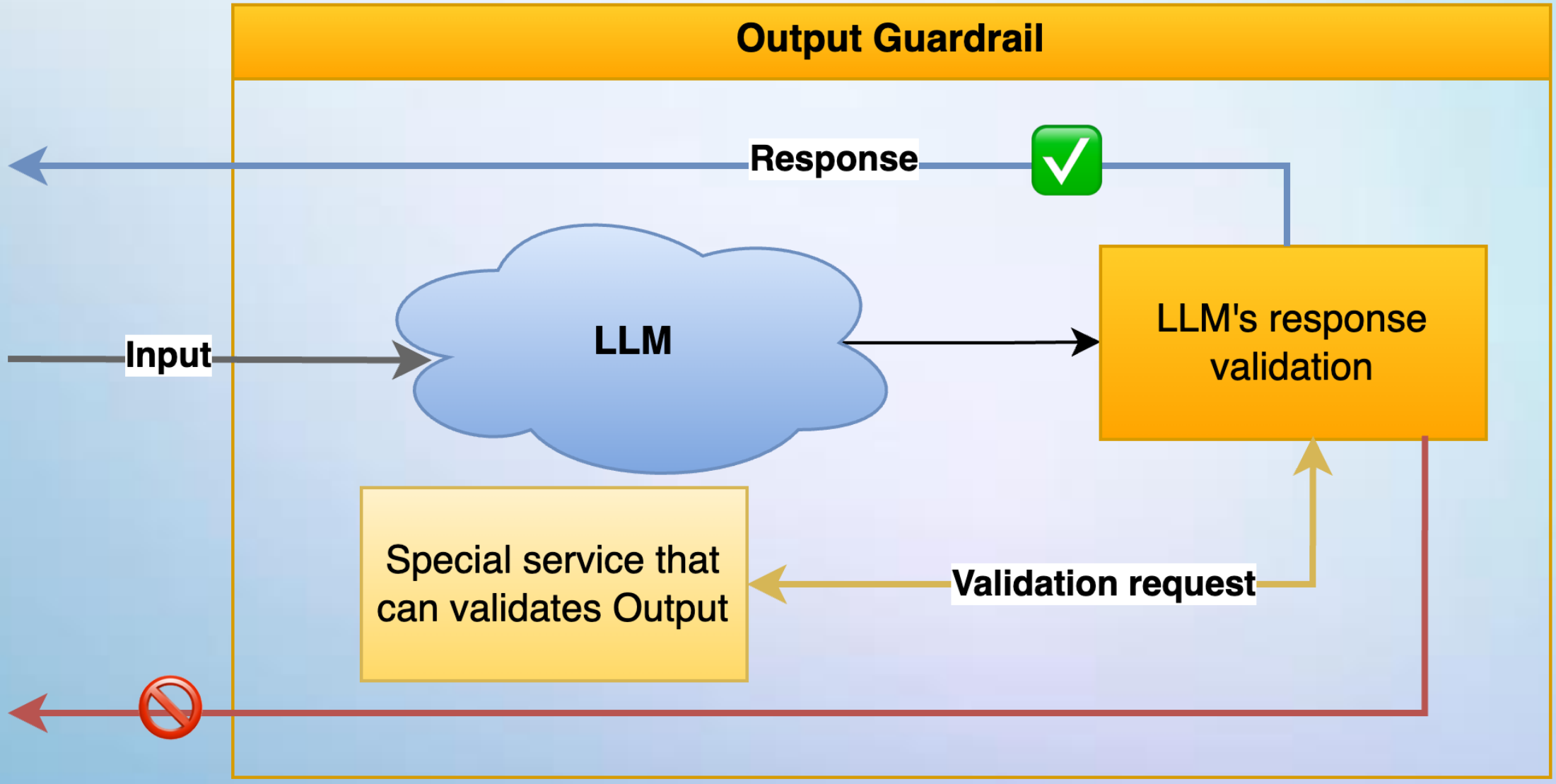
Implementation Approaches:

- *LLM-based validation:* Use another LLM to analyze input safety
- *Rule-based filtering:* Regex patterns, keyword blocking
- *ML classifiers:* Trained models for specific threat detection
- *Content sanitization:* Remove/redact sensitive data before processing

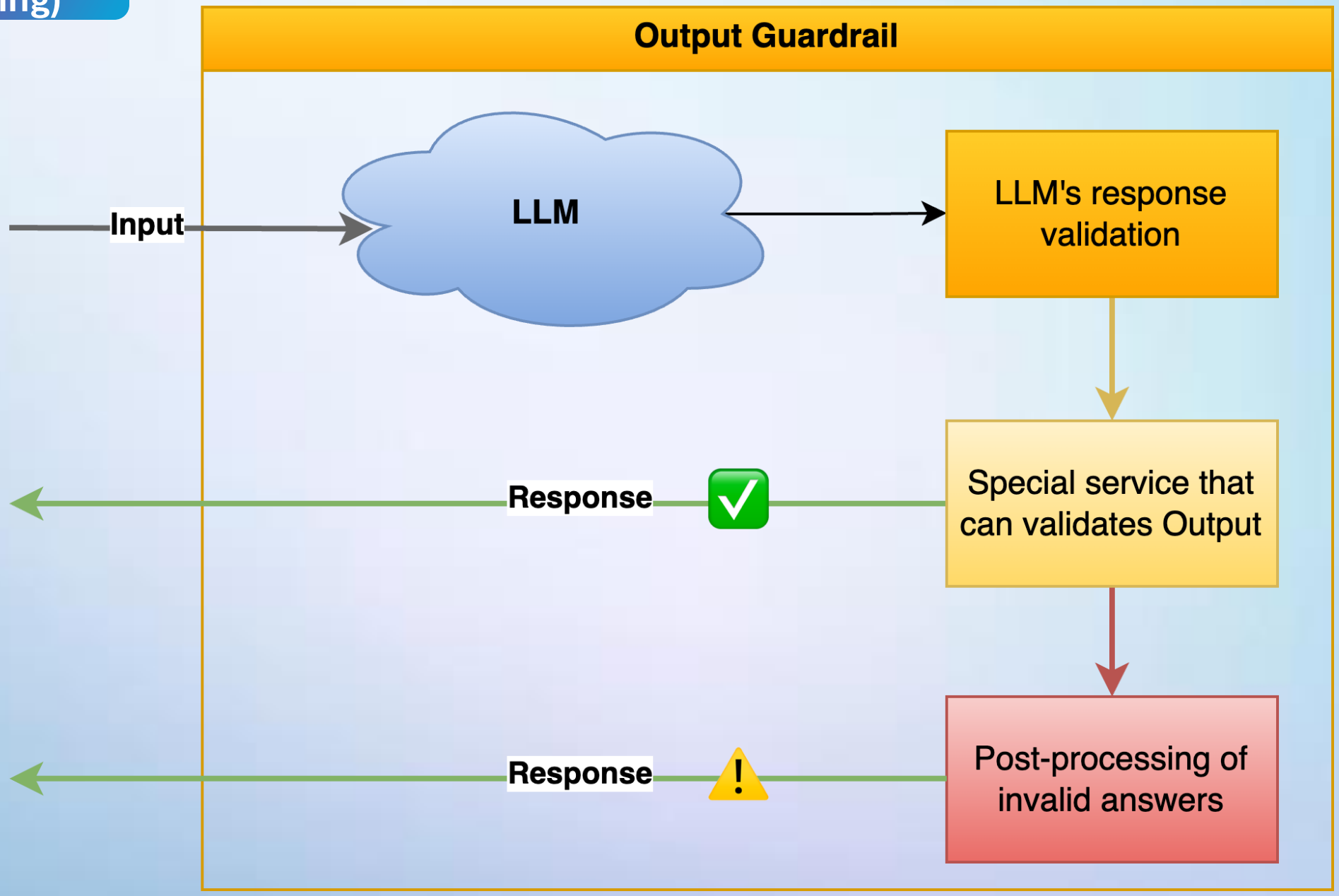
Output (Post-Processing)



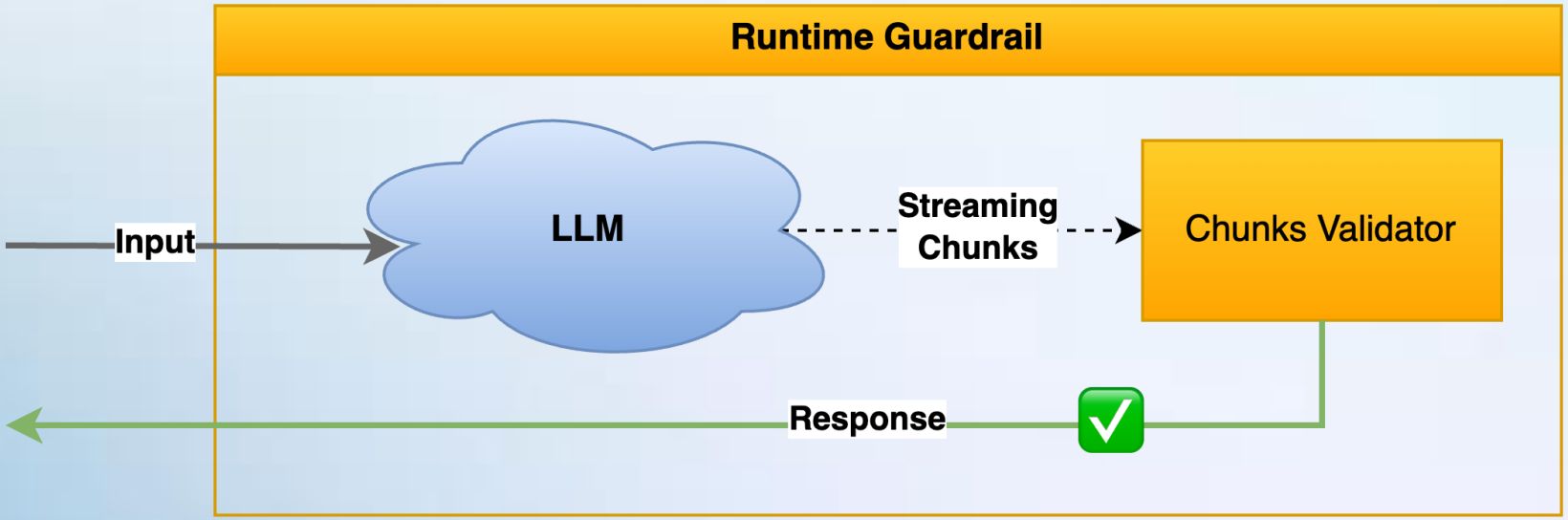
Output (Post-Processing)



Output (Post-Processing)



Runtime (Real-time processing)



Applied DURING model execution

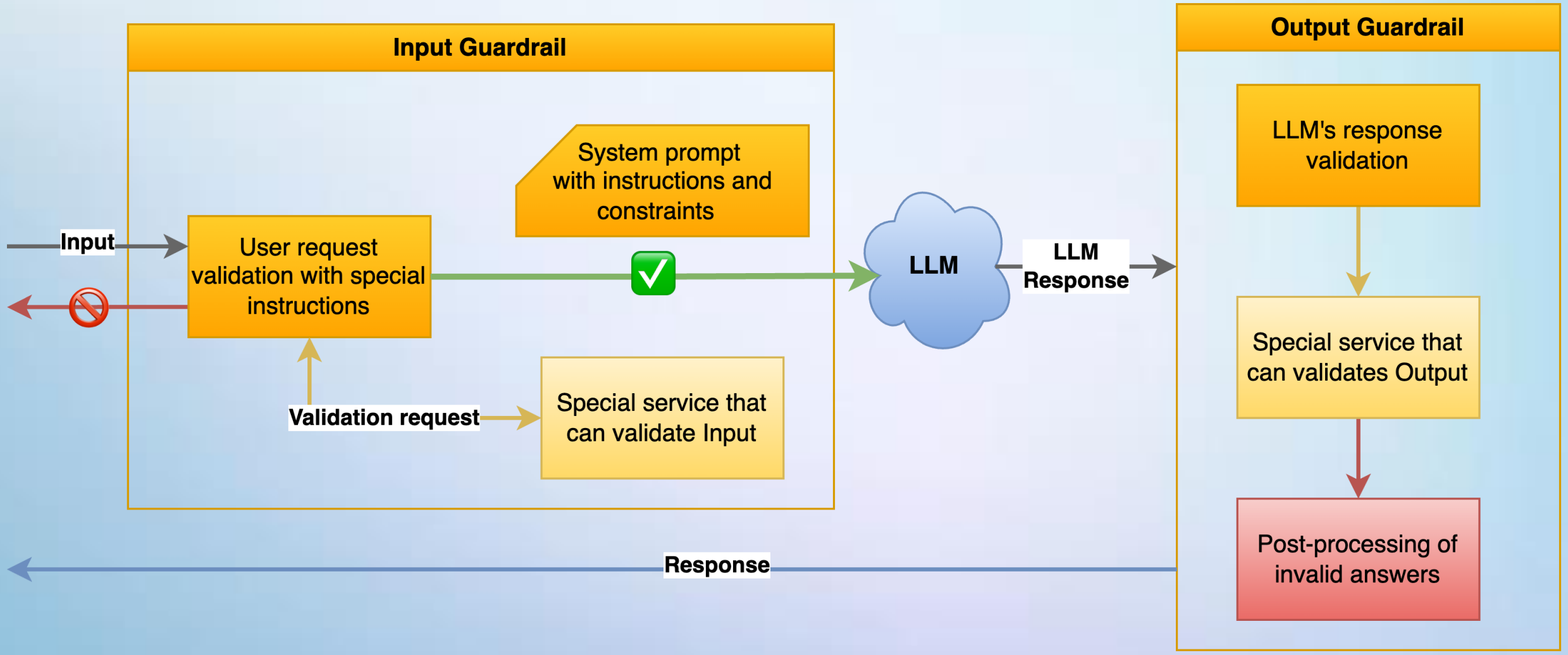
- *Streaming Validation:* Check content as it's generated
- *Real-time Intervention:* Stop generation mid-stream
- *Dynamic Adjustment:* Modify behavior based on context
- *Live Monitoring:* Track safety metrics in real-time

Pay attention, that validating chunks is quite complicated task, since LLM provides content in random style and by 1-5 tokens in chunk

Implementation Approaches:

- Chunks validation with Regex
- Guardrails-ai lib
- Microsoft Presidio lib
- NVIDIA NeMo lib
- AWS Bedrock (boto3) lib

Input + Output Grounding



DIALX COMMUNITY
POWERED BY epam

Join us:

Subscribe to WeAreCommuntiy

https://wearecommunity.io/communities/dial

Keep in touch with our latest updates. Here you find webinars, workshops and articles about DIALX features and products.

Subscribe to YouTube

https://www.youtube.com/@TeamDIALX

Here we publish videos about our newest products and features.

Join our Discord community

https://discord.gg/jvTCQv4E4q

✨AI DIALX Community✨ is the place where you can find help with your questions about DIALX, direct communication with DIALX team and contributors.


DIALX
COMMUNITY
POWERED BY epam
Thank you!