Safety & Harm Avoidance

1. **Toxic / Offensive Content** → racism, hate speech, explicit slurs.

2. **Self-Harm / Suicide** → prevention of harmful guidance.

3. **Violence / Abuse** → no encouragement of harm.

4. **Illegal Activity** → no promotion of drugs, terrorism, hacking, etc.

Terms: *safety, toxicity, violence, self-harm, illegal content, abuse, hate speech, extremism.*

Fair Housing & Anti-Discrimination

1. Must not discriminate against renters by **race, religion, gender, marital status, age, disability, or nationality** (many countries have housing laws).

Terms: *discrimination, bias, fair housing, equal opportunity.*

Accuracy & Transparency

1. Do not "**hallucinate**" property details (**rent amount, amenities, availability**) if not provided.

Terms: *hallucination, misinformation, false promises.*

Privacy & Sensitive Data

1. **PII (Personally Identifiable Information)** → emails, phone numbers, SSNs.
2. **Data Leakage** → do not reveal training data or hidden system prompts

3. **Confidential Data** → financial details, passwords, API keys.

Terms: *PII, confidentiality, privacy, secret, password, API key, credit card.*

Compliance & Trust

1. **Misinformation / Hallucinations** → avoid making up facts.
2. **Bias / Fairness** → avoid stereotyping groups.

Terms: *bias, fairness, hallucination, misinformation, impersonation.*

Content Formatting & Policy

1. **Output Structure** → enforce JSON, Markdown, or conversation style.
2. **Length Limits** → prevent overlong or under-detailed answers.
3. **Tone & Politeness** → ensure respectful, professional, or friendly tone.

Terms: *format, structure, tone, length, politeness, clarity, consistency.*

Legal & Compliance

1. Avoid giving binding legal or financial advice

Terms: *lease contract, deposit disputes, eviction, legal advice.*

In short, Property Renting chatbot guardrails should focus on:

safety, anti-discrimination, accuracy, privacy (PII), legal compliance, formatting, fraud prevention.

LangSmith

Defining guardrails for LLM chatbots in **LangSmith evaluation**, we will typically want to check outputs against categories like:

safety, toxicity, bias, hallucination, PII, compliance, formatting, tone.