Guardrails are tools and libraries designed to help ensure the safe and reliable deployment of AI models by providing mechanisms for content moderation, monitoring, and guiding model outputs. Here’s a list of notable guardrail systems:

### **1. LlamaGuard**

* **Description**: LlamaGuard is designed to enhance the safety and reliability of language models like LLaMA. It provides mechanisms to restrict and guide model outputs based on specified rules and constraints.
* **Link**: [LlamaGuard GitHub Repository](https://github.com/langchain-ai/llama-guard)

### **2. Nemo Guardrails**

* **Description**: Nemo Guardrails is a conversational AI guardrails system by NVIDIA, integrated with their NeMo framework. It helps in defining and enforcing safety and control policies in AI interactions.
* **Link**: [Nemo Guardrails GitHub Repository](https://github.com/NVIDIA/NeMo-Guardrails)

### **3. TruLens**

* **Description**: TruLens is an open-source toolkit that provides evaluation, debugging, and monitoring tools for ensuring the quality and safety of AI systems, including LLMs.
* **Link**: [TruLens GitHub Repository](https://github.com/trulens/trulens)

### **4. DeepSafe**

* **Description**: DeepSafe is a tool designed to provide guardrails for deep learning models by integrating pre- and post-processing safety checks to mitigate harmful outputs.
* **Link**: [DeepSafe GitHub Repository](https://github.com/DeelAI/deepsafe)

### **5. BlenderBot Guardrails**

* **Description**: Developed for Facebook's BlenderBot, these guardrails help in steering the conversation towards safe and coherent dialogue, preventing the model from generating unsafe or inappropriate responses.
* **Link**: [BlenderBot 2.0: Safety and Control](https://ai.facebook.com/blog/blender-bot-2-an-open-source-chatbot-that-builds-long-term-memory-and-searches-the-internet/)

### **6. AI Fairness 360 (AIF360)**

* **Description**: IBM's AIF360 is an open-source toolkit that helps in detecting and mitigating bias in AI models, providing fairness checks that act as guardrails to ensure ethical AI deployments.
* **Link**: [AI Fairness 360 GitHub Repository](https://github.com/Trusted-AI/AIF360)

### **7. Rasa X**

* **Description**: Rasa X is a platform for managing and improving AI assistants built with Rasa. It includes features for monitoring conversations, improving the assistant's responses, and enforcing guardrails on the interaction.
* **Link**: Rasa X Official Website

### **8. Hugging Face's Inference API Guardrails**

* **Description**: Hugging Face offers an Inference API with built-in safety measures and guidelines to prevent misuse of the models, providing a safe deployment option for LLMs.
* **Link**: Hugging Face Inference API

### **9. BERT Guardrails**

* **Description**: Designed for the BERT model family, these guardrails focus on managing and filtering outputs to prevent unintended harmful responses, applicable in various conversational and QA systems.
* **Link**: BERT Guardrails for Conversational AI

### **10. Fairseq Guardrails**

* **Description**: Facebook AI's Fairseq toolkit includes components for adding safety and ethical considerations to the training and deployment of large language models.
* **Link**: [Fairseq GitHub Repository](https://github.com/facebookresearch/fairseq)

These resources provide a comprehensive set of tools and libraries to implement guardrails in AI systems, ensuring their safe, reliable, and ethical use.