Guardrails Comparison Table

name	Developed by	explanation	Main technologies used	type	merit	Disadvantages
Llama Guard	Meta	A Llama-based model classifies inputs and outputs to improve conversation security.	LLM Base (Llama)	Safeguard Model	Simple to implement (same as how to use the Llama model)	Depends on Llama model performance
NeMo Guardrails	NVIDIA	Colang sets constraints to prevent LLM from crossing boundaries.	Embedding + KNN / LLM	Input/Output Control	Flexible and complex configuration possible	Design and tuning are time-consuming
Guardrails AI	Guardrails AI	Guarantees the format of LLM outputs. Can be used in combination with shared validators.	Varies (depends on guardrails used)	Input/Output Control	You can combine and use various guardrails published by others in the community.	The scope of coverage is limited depending on the guardrails that are combined.