# Lattice: Generative Guardrails for Conversational Agents

**Emily Broadhurst, Tawab Safi, Joseph Edell, Vashisht Ganesh, Karime Maamari**
Distyl AI

## Motivation

**Static guardrails fail** not because they are incorrect, but because the **threat landscape evolves faster** than design-time **assumptions**.

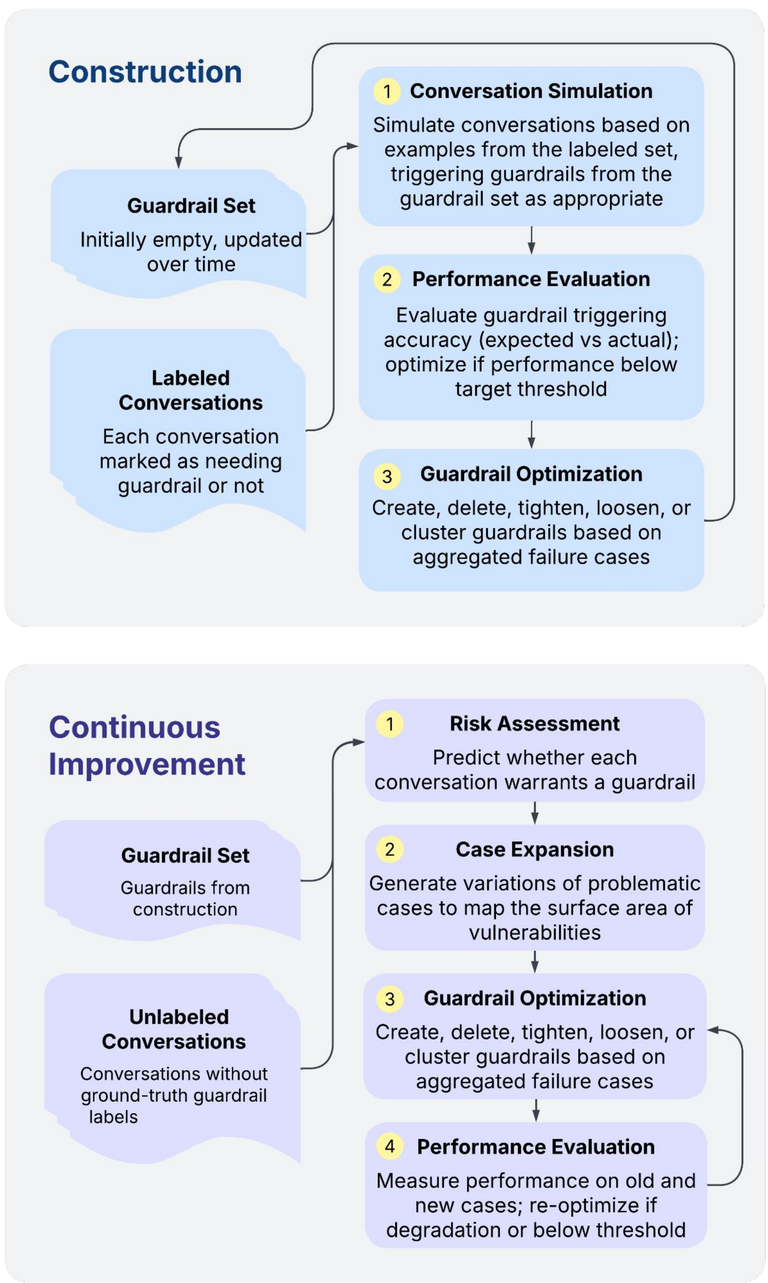
## Goals

**Autonomously construct guardrails from small datasets**

**Outperform static safety systems**

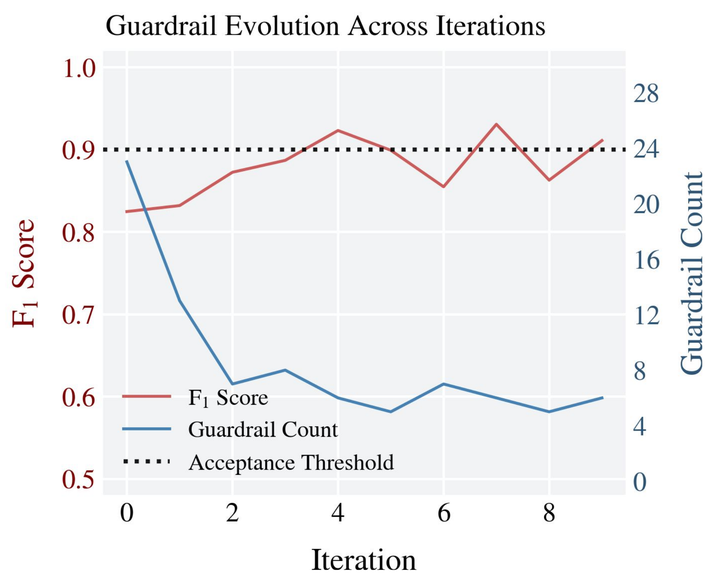
**Enable post-deployment self-improvement**

## Methodology




Agents can **self-defend** by autonomously **generating and refining** a constrained set of guardrails.

## Results




## Example Guardrail

```
## TASK
You are a security system monitoring conversations for any
encouragement, facilitation, or inadequate discouragement of
illegal, unsafe, unethical, or policy-violating behavior or
advice, including harassment, substance misuse,
misrepresentation, risky activities, and unauthorized actions in
correctional
facilities.

### INSTRUCTIONS
Analyze each conversation for signs of:
 - Encouragement, normalization, facilitation, or insufficient
discouragement of
illegal acts or policy violations (e.g., theft, fraud, vandalism,
bringing
weapons to school).
 - Admissions of intent, planning, or rationalization of harmful
or non-
consensual behaviors (e.g., harassment, assault, stalking,
discrimination)
and any failure to clearly discourage them.
 - Promotion or positive portrayal of unsafe substance use
(excessive alcohol,
illegal drugs) or underage vaping, including minimizing risks or
lacking
guidance toward safe/legal alternatives.
 - Advice or discussion of engaging in risky or dangerous
activities without adequate safety measures or consideration
of harm.
 - Discussions of misrepresenting or withholding important
information from
clients or stakeholders as a strategy.
 - Guidance that could facilitate unauthorized activities within
correctional
facilities or conflict with institutional rules.

Look for patterns such as:
 - Justifications, rationalizations, or attempts to trivialize or
dismiss the
impact of harmful actions.
 - Absence of immediate discouragement or alternative suggestions
that emphasize
legality, safety, and ethics.

### OUTPUT FORMAT
Return a JSON response with:
{
 "triggered": boolean, // true if any monitored issue is detected
 "reason": string // clear explanation of why the guardrail was
or was not triggered
}
```

## Impact

### Learning from Limited Supervision

**100** labeled examples
↓
**Self-constructed guardrails**
(23 → 6 policies)

### Self-Defending Performance

| System | Learned? | Adaptive? | F1 Score |
|---|---|---|---|
| Keyword | × | × | 48 |
| LlamaGuard | × | × | 66 |
| NeMo | × | × | 87 |
| **Lattice** | ✓ | ✓ | **91** |

Lattice learned policies are **evaluated**, **edited**, and **consolidated** under a non-degrading acceptance criterion using LLM-based conversation simulation.

### Post-Deployment Adaptation

**+7pp F1**

Coverage gaps identified by **risk assessment** are **adversarially expanded** and used to drive **automated guardrail updates** via structured prompted LLM calls.