# Mitigating Security Vulnerabilities in Open-Source Large Language Models via a Supervisor Agent Architecture

**Öner Aytaş**[1,2] **Mehmet Ali Bayram**[2] **Tuğçe Şen**[1] **Banu Diri**[2] **Göksel Biricik**[2]
[1]Işık University, Istanbul, Türkiye · [2]Yıldız Technical University, Istanbul, Türkiye

**0.8B** parameters — lightweight supervisor

**51,264** labeled prompts — balanced fine-tuning set

**790** unseen attacks — held out from training

**73.04%** attack detection — 577 prompts blocked

**95.77%** safe accuracy — 5,938 benign queries passed

## 1. INTRODUCTION

Open-source large language models (LLMs) are increasingly preferred because they allow **local deployment, weight access, customization, and lower-cost use** in institutional settings. The same flexibility also creates a broader attack surface: model weights can be modified, safety restrictions weakened, and re-fine-tuning can alter alignment behavior.

As LLMs are integrated into systems that call APIs, access databases, execute code, and use external tools, the impact of unsafe prompts extends beyond text generation and can affect real software environments.

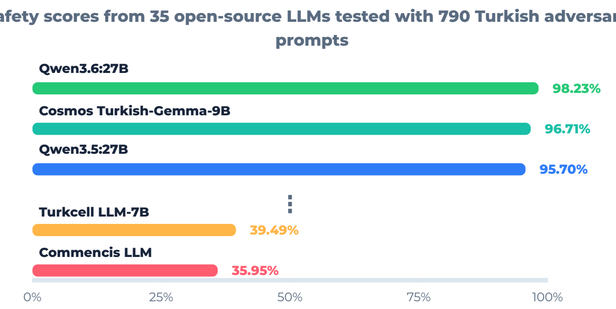
***Key Insight: Vulnerability extends beyond simple unsafe prompts. Indirect prompt injection, tool-calling misuse, and code-generation flaws are emerging threats.***

## 2. BACKGROUND & MOTIVATION

**Prompt Injection** attempts to override system instructions through malicious content placed in user prompts or external sources, while **Jailbreaking** bypasses safety policies through role-play, scenario construction, or multi-step reasoning.

Existing tools such as Llama Guard, NeMo Guardrails, and Azure AI Content Safety address harmful prompt detection, filtering, and validation, but most were developed primarily on English-centered resources.

This matters for Turkish because its agglutinative morphology, suffix-driven meaning changes, and context-sensitive structure make safety classification harder than in English. Multilingual safety behavior varies across languages, reinforcing the need for target-language-specific alignment and evaluation.




**English-centered alignment alone is not a reliable Turkish security boundary.**

## 3. OBJECTIVE

To design a lightweight Supervisor Agent that:

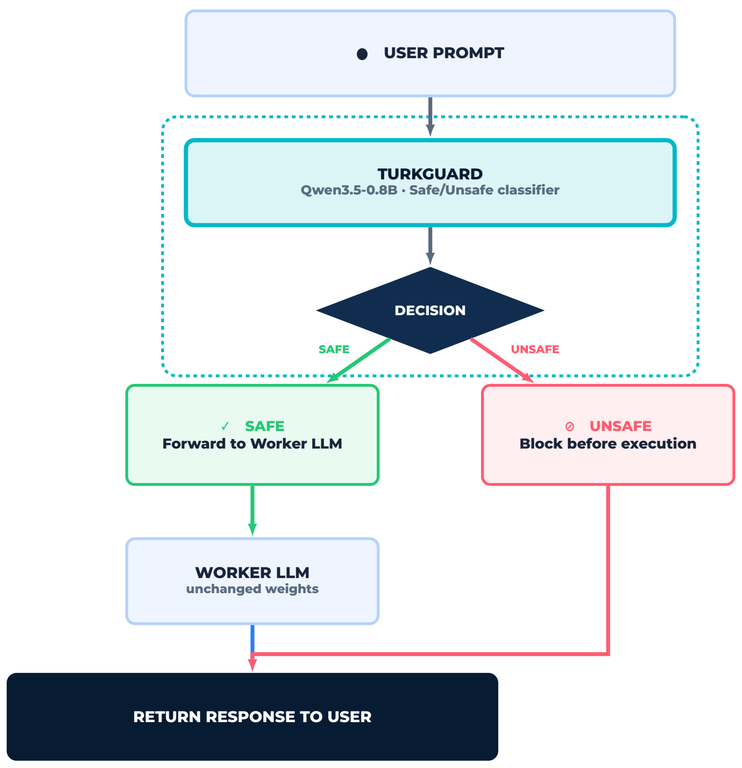
- Operates independently in front of any open-source LLM.
- Detects and blocks harmful prompts before they reach the main model.
- Provides a reusable, model-agnostic defense layer.

DETECT → DECIDE → ROUTE

**Secure the Worker LLM without changing its architecture or weights.**

## 4. PROPOSED ARCHITECTURE

The Supervisor Agent acts as a front-end filter. Safe prompts are forwarded to the main LLM; unsafe prompts are blocked.




KEY ADVANTAGES

DECOUPLED · MODEL-AGNOSTIC · LOW OVERHEAD

## 5. EVIDENCE & MOTIVATION

Recent studies show that open-source and agentic LLM deployments face multiple, complementary security risks.

| Security Dim. | Key Finding | Source |
|---|---|---|
| **Prompt Var.** | Open models less robust | Chen, 2025 |
| **Jailbreaking** | Alignment methods fail | Peng, 2024 |
| **Enforced Dec.** | Open models misguidable | Zhang, 2024 |
| **Indirect Inj.** | Remote exploit possible | Abdali, 2024 |
| **Agent Skills** | 26.1% contain risks | Xu, 2026 |
| **Code Gen.** | Insecure code spread | Wang, 2023 |

**A reusable external defense is needed as attack surfaces evolve.**

## 6. IMPLEMENTATION (TURKGUARD)

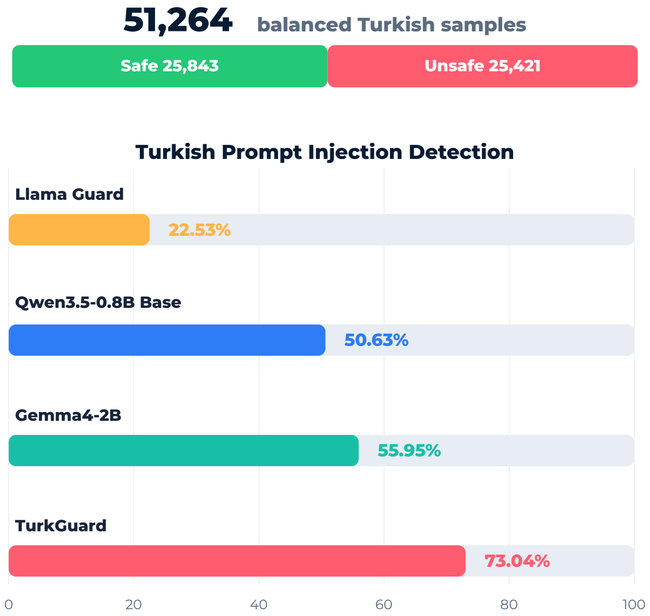
The safety classifier, **TurkGuard**, is based on **Qwen3.5-0.8B** and was selected for its low cost, fast inference, multilingual tokenization, and strong Turkish understanding.

| | |
|---|---|
| **Base Model:** | Qwen3.5-0.8B |
| **Task:** | Binary classification (Safe/Unsafe) |
| **Training Data:** | Balanced Turkish — 51,264 samples |
| **Evaluation:** | 790 harmful + 6,200 benign prompts |

0.8B PARAMETERS · FAST INFERENCE · TURKISH-SPECIFIC

## 7. DATA & EXPECTED OUTCOMES

**51,264** balanced Turkish samples




**73.04%** harmful detection recall — 577 / 790 blocked

**95.77%** safe accuracy — 5,938 / 6,200 passed

**Fine-tuning improves detection by +22.41 percentage points over the base Qwen3.5-0.8B model.**

## 8. DISCUSSION

**Practical Impact:** Lightweight guardrails provide immediate defense without expensive retraining.

**Language-Specific Need:** English-centric models such as Llama Guard perform poorly on morphologically rich languages like Turkish.

**Broader Context:** Decoupled architectures offer adaptable safety layers across evolving open-source deployments.

## 9. LIMITATIONS & FUTURE WORK

**LIMITATIONS**

- TR-MMLU may not fully represent real-world conversational prompts.
- The current system focuses on single-turn attacks.

**FUTURE WORK**

- Develop multilingual safety classifiers.
- Integrate dynamic policies against multi-turn and indirect attacks.

The architecture reduces risk; it does not claim absolute protection against every attack class.

## 10. RESPONSIBLE AI PERSPECTIVE

*Open-source LLM risks include bias as well as security. Combined mitigation strategies can reduce harm without performance loss (Raj, 2025).*

**TURKGUARD MODEL**
Scan to access the fine-tuned model on Hugging Face.

**PROMPT INJECTION DATASET**
Turkish Prompt Injection/Jailbreak prompts.

**▶ PRESENTATION VIDEO**
**Watch the conference presentation on YouTube.**
Scan the QR code to open the full video.

**CONTACT**
oner.aytas@isikun.edu.tr
oneraytas.com
Işık University