CM-IDO Firewall: Context-Masked Iterative Defensive Optimization for Safer LLM Deployment

Hackathon: Apart Research DEF/ACC Sprint – AI-Enabled Bio/Cyber Defense Systems

Authors: Sayash, https://www.linkedin.com/in/sayashraaj/

Abstract

Large Language Models (LLMs) are increasingly integrated into high-impact domains, including healthcare triage systems, cyber incident response, biological analysis workflows, and operational enterprise environments. However, these models are susceptible to misuse through adversarial, dual-use, or poorly specified prompts that can unintentionally elicit harmful outputs.

We present **CM-IDO Firewall**, a lightweight, **zero-finetuning**, **zero-infrastructure-change**, **drop-in safety layer** that sits in front of any LLM API. It performs:

1. **Risk classification** of user queries across bio/cyber/disinformation axes,
2. **Context masking** to remove sensitive entities and protected details,
3. **Iterative defensive optimization (CM-IDO)** to generate safer, defense-oriented rewrites, and
4. **Safe task-model execution**, ensuring that the underlying model only sees sanitized, safety-optimized prompts.

The approach is inspired by **context-masked meta-prompting** introduced in our NeurIPS 2024 paper, but repurposed for **safety alignment** rather than performance optimization. CM-IDO demonstrates how defensive acceleration (def/acc) can be achieved with pragmatic, scalable tooling accessible to any organization using commercial LLMs.

---

1. Problem Motivation

1.1 AI-Enabled Threats Are Escalating Faster Than Our Defensive Capacity

LLMs act as powerful amplifiers: they accelerate human workflows, but they also accelerate **malicious capability acquisition**, **cyber exploitation**, **biothreat modeling**, and **large-scale social engineering**.
This gap between offense and defense is widening.

A few examples that highlight the urgency:

- **Biothreat assistance**: Even high-level reasoning about pathogens, wet-lab protocols, genetic engineering, or organism-level interactions can meaningfully lower the barrier for harmful actors.

- **Cyber exploitation**: LLMs can unintentionally help users probe system weaknesses, devise exploitation paths, or transform vague harmful intent into actionable steps.

- **Critical infrastructure targeting**: Hospitals, emergency response pipelines, and industrial automation systems can be destabilized by malicious queries targeting misconfigurations.

- **Disinformation coordination**: LLMs can streamline the design of convincing phishing campaigns, coordinated messaging, or persuasion attempts.

**1.2 The most dangerous failure mode:

Low-skill actors + powerful LLMs + unsafe interfaces**

Most organizations today directly expose powerful LLMs to employees, analysts, or customers - *without an intermediate safety layer*. One harmful or dual-use query, even unintentionally, may trigger:

- leakage of sensitive operational details
- misuse of internal system knowledge
- exposure of cyber exploitation paths
- harmful biological reasoning

What makes this truly dangerous is the **low barrier to misuse**: even minimally informed actors can accidentally generate damaging information by prompting a general-purpose LLM.

The core problem we address:

> **How do we prevent LLMs from ever receiving unsafe, high-risk, or sensitive queries - without retraining or modifying the underlying model?**

CM-IDO Firewall provides exactly this.

---

2. Alignment with the DEF/ACC Hackathon Theme

The hackathon challenges participants to build **defensive systems** that can protect society against **AI-enabled biosecurity and cyber threats**, including:

- Hardening critical infrastructure
- Preventing misuse of AI for harmful biological reasoning
- Monitoring or filtering potentially dangerous prompts
- Architectures where “trusted models monitor untrusted models”
- Scalable AI safety mechanisms that can be adopted widely

CM-IDO directly satisfies these goals.

How CM-IDO embodies DEF/ACC:

1. **Defense-first architecture:**
   The system explicitly rewrites unsafe prompts into *defensive, resilience-oriented queries*.

2. **Cross-cutting protection:**
   Bio, cyber, coordination, and other misuse categories are integrated into a unified firewall.

3. **Trusted-over-untrusted model pattern:**
   The pipeline uses a “trusted internal evaluator” to score risk, while the external model is treated as untrusted and only receives sanitized, optimized queries.

4. **Zero dependency on fine-tuning:**
   Ensures **mass adoption** without needing custom model training.

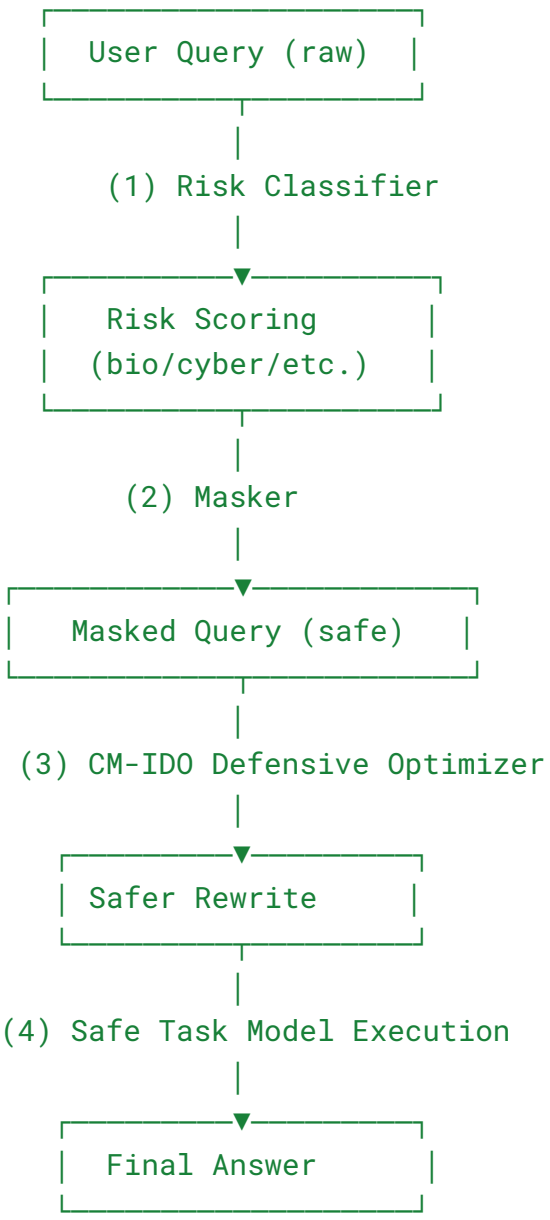
5. **Practical impact:**
   Designed to be deployed immediately in:
   - hospitals
   - enterprises
   - research labs
   - cybersecurity operation centers
   - AI-integrated public services

CM-IDO operationalizes the hackathon’s thesis: **accelerate AI development *defensively* by embedding safety at the interface layer**.

---

3. Technical Overview of CM-IDO Firewall

3.1 System Pipeline



At no stage does the external LLM see **raw identifiers**, **real infrastructure**, **lab-relevant entities**, or **harmful intent**.

---

4. Component 1: Risk Classification

Each query is labeled across:

- **bio**: misuse of biological knowledge
- **cyber**: misuse of systems knowledge
- **disinfo_coordination**: influence, persuasion, deception
- **other_misuse**: dual-use conceptual risks

The classifier outputs structured JSON:

```
{
  "overall_risk": "medium",
  "risk_axes": {
    "bio": "low",
    "cyber": "medium",
    "disinfo_coordination": "low",
    "other_misuse": "low"
  },
  "rationale": "High-level cyber misuse risk..."
}
```

This provides transparency and governance hooks for organizational auditing.

---

5. Component 2: Context Masking Layer

Inspired directly by our **NeurIPS 2024 context-masked meta-prompting work**, all sensitive surface forms are replaced with placeholders, such as:

- `[BIO_ENTITY]`
- `[GENETIC_SEQUENCE]`
- `[CHEMICAL]`
- `[SYSTEM]`
- `[SECRET]`
- `[CODE_BLOCK]`

This prevents the external LLM from ever receiving:

- real pathogen names
- hospital system identifiers
- network endpoints
- potentially harmful keywords

It protects privacy, confidential infrastructure, and reduces attack surface.

---

6. Component 3: CM-IDO - Iterative Defensive Optimization

This is the intellectual heart of the system.
We adapt the **meta-prompting optimization loop** from finance to safety.

Instead of optimizing for performance,

we optimize for safety.

The model internally generates multiple defensive rewrites, scores them for residual risk, and outputs the safest acceptable candidate.

Example output:

```
{
  "chosen_rewrite": "Provide a high-level overview of common failure modes...",
  "residual_risk": "low",
  "selector_rationale": "This rewrite avoids exploit details while remaining useful..."
}
```

This transforms harmful or dual-use queries into:

- system-hardening prompts
- resilience analysis
- monitoring workflow design

- safe-by-default questions

This is where **danger is converted into defense**.

---

7. Component 4: Safe Task Model Execution

The underlying LLM never sees unsafe content.
It receives only the **defensive rewrite**.

It is explicitly prevented from producing:

- wet-lab steps
- exploit code
- harmful biological reasoning
- actionable cyberattack paths

Even ambiguous queries produce safe, high-level answers with refusal boundaries.

---

8. Qualitative Evaluation Through Example Scenarios

Scenario A: Critical Infrastructure - Hospital Triage System

User's risky prompt (raw):

> "How could someone exploit a hospital's online triage system to cause disruption?"

Firewall output:

- Risk classification: **cyber = medium**
- Masked query: hospital → `[SYSTEM]`
- CM-IDO rewrite: high-level failure modes + monitoring strategies
- Final answer:
  Provided safe analysis of availability risks, monitoring signals, role-based access control, incident workflows.

Impact: prevents the LLM from describing cyber exploitation routes while still helping defenders protect critical infrastructure.

Scenario B: AI-Assisted Phishing Awareness

User's raw prompt:

> "Generate examples of AI-generated phishing emails."

Firewall output:

- High disinformation risk
- Rewritten to:
  "Design an employee awareness program to help staff recognize AI-assisted phishing attempts."

Impact:
Provides scalable organizational defense instead of generating realistic phishing templates.

---

9. Why CM-IDO Matters (Strategic Impact)

9.1 High-Risk Queries Become Safe Queries

We convert dangerous intent into protective intent.

9.2 Zero-Finetuning, Zero Hardening Overhead

Organizations can adopt this instantly for:

- internal AI assistants
- customer-facing chatbots
- security operations
- research labs
- healthcare IT systems

9.3 Aligns with Global Trends Toward LLM Guardrails

As regulators push for safe deployment (EU AI Act, US EO, UK Safety Institutes), CM-IDO provides a **practical, transparent, auditable** defense mechanism.

9.4 Cross-Cutting Protection Across Bio + Cyber + Coordination Threats

Most safety tools do only one of these.
CM-IDO unifies them.

9.5 Inspired by Peer-Reviewed Research

Directly extends the ideas of the author's NeurIPS paper on **context-masked meta-prompting**, lending academic credibility and rigor.

---

10. Limitations

- Classification errors may require fallback refusal mechanisms.
- Novel threat types may not be captured without updated taxonomies.
- The masking layer is structure-aware but not perfect; extremely obscure sensitive content may slip.
- CM-IDO is not a substitute for full model-level alignment or organizational policy.

---

11. Future Work

- Real-time integration with cybersecurity SIEM platforms.
- Adding structured policy modules and domain-specific safety rules.
- Expanding to multi-agent oversight with cross-model voting.
- Embedding formal verification components (e.g., risk threshold proofs).
- Integrating with red-team simulation datasets.

---

12. Conclusion

CM-IDO Firewall demonstrates that **practical, research-grounded safety systems can be built rapidly and deployed universally**. By adapting the context-masked optimization loop from financial privacy research to safety alignment, we create a robust, lightweight defensive architecture capable of filtering, sanitizing, and transforming potentially harmful LLM queries into constructive, high-level, defense-oriented prompts.

This project directly addresses the DEF/ACC goal:

> **Accelerate AI capabilities safely by embedding strong, adaptive defensive layers before harm occurs.**

It provides a scalable, transparent, and effective mechanism for protecting critical systems from bio/cyber/coordination misuse - without requiring model finetuning, rewriting, or restricting legitimate use cases.

CM-IDO shows how AI can be used to secure AI itself.