

CM-IDO Firewall:

Context-Masked Iterative Defensive Optimization for Safe LLM Deployment



Author:
Sayash

Arcadia Impact Fellow
NeurIPS 2025, Fidelity Investments,
IIT Madras

Problem Overview: AI Threats Are Scaling Faster Than Defenses

The threat landscape:

- LLMs drastically lower the barrier to *bio*, *cyber*, and *coordination* misuse.
- Even vague malicious intent can be refined into harmful steps by general-purpose models.
- Critical infrastructure (hospitals, emergency response, enterprise systems) is already vulnerable.
- Current safety layers are inadequate: most LLM deployments rely purely on provider policies.

The core danger:

Low-skill actors + powerful LLMs + unsafe interfaces = catastrophic amplification.

We urgently need **defensive acceleration** — safety mechanisms that scale as fast as AI's capabilities.

Alignment With DEF/ACC Theme

Hackathon goal: Build systems that protect society from AI-enabled threats in:

- **Biosecurity**
- **Cybersecurity / critical infrastructure**
- **Disinformation & coordination risks**
- **AI control, monitoring, layered defense**

CM-IDO contributes directly by:

- Blocking unsafe queries *before* they reach the model
- Converting harmful intent into *defensive, resilience-oriented* queries
- Creating a “trusted model supervising an untrusted model” architecture
- Being instantly deployable by any organization (no fine-tuning required)

This is *practical defensive acceleration*.

Motivation: What Makes This Problem Dangerous

1. LLMs can unintentionally assist:

- Cyber exploits
- Biological reasoning
- Infrastructure disruption
- Social engineering at scale

2. Many harmful queries are not obviously harmful.

Even well-intentioned users ask dual-use questions.

3. Organizations often expose LLMs directly to internal systems.

Without a safety firewall, the LLM can:

- Leak internal architecture
- Produce harmful code suggestions
- Reveal vulnerabilities
- Provide high-level biological insights with dangerous implications

4. Safety mechanisms must be adoptable today.

No heavy training. No safety research team needed.

We solve this by wrapping any LLM with a **safety-first transformation layer**.

Research Basis: My NeurIPS 2025 Paper

[Context-Masked Meta-Prompting for Privacy-Preserving LLM Adaptation in Finance](#)

Accepted at NeurIPS 2025

Core contributions of the paper:

- Mask sensitive context
- Use internal evaluator to iteratively optimize prompts
- Keep external LLM “blind” to private data
- Achieve high performance without direct exposure

How CM-IDO adapts it:

- Replace *performance optimization* with *safety optimization*
- Replace *financial privacy* with *bio/cyber safety*
- Apply the same architecture to defend LLM systems
- Use iterative prompt generation to minimize **residual risk**

Result:

A NeurIPS-validated research method repurposed for AI safety and defense.

Introducing CM-IDO Firewall

CM-IDO = Context-Masked Iterative Defensive Optimization

A four-stage safety pipeline:

- **Risk Classification**

- Detects bio, cyber, disinfo, other misuse axes
- Outputs structured JSON
- Conservative by design

- **Context Masking**

- Removes pathogen names, system identifiers, code payloads, credentials
- Replaces them with structured placeholders
- LLM never sees sensitive details

- **Iterative Defensive Optimization (CM-IDO)**

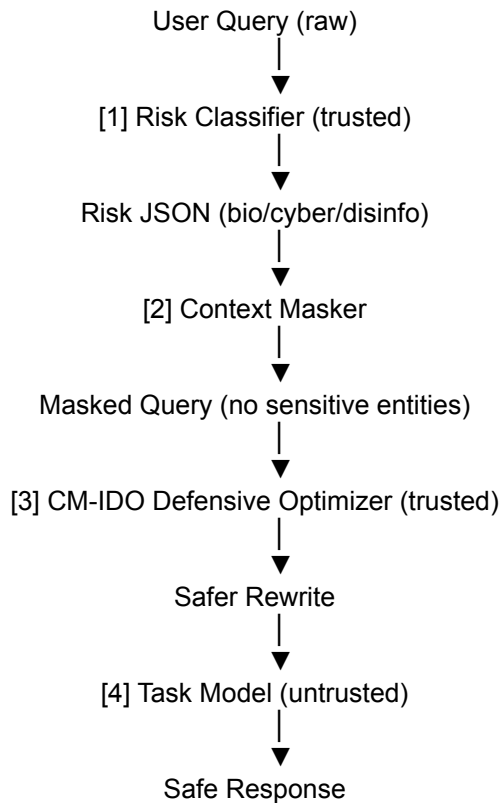
- Generates multiple defensive rewrites
- Scores them with an internal evaluator
- Selects the rewrite with *lowest residual risk*

- **Safe Task Execution**

- Only the safe rewrite is sent to the underlying model
- Final output is high-level, protective, and safe

Zero model training. Zero infrastructure friction.

Architecture Diagram



Core principle:

*The untrusted model
never sees raw inputs.
Only safe, abstract,
defensive prompts reach it.*

Example Scenario: Critical Infrastructure (Hospital Triage)

User's risky query:

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

Firewall output:

- **Risk:** Cyber "medium"
- **Masked:**
"How could someone misuse a hospital's [SYSTEM]?"
- **Optimized rewrite:**
"Describe high-level failure modes in critical healthcare IT and monitoring strategies."
- **Final safe response:**
Describes availability risks, load monitoring, access control policies, resilience planning.
ZERO exploit or attack details.

Impact:

Protects hospital infrastructure while still supporting defensive security teams.

Example Scenario: Disinformation & Social Engineering

User's query:

“Generate realistic phishing emails written with AI.”

CM-IDO rewrite:

“Describe how to design an employee awareness and reporting program for AI-assisted phishing.”

Final result:

- Helps the defender
- Prevents generation of phishing templates
- Preserves organizational security
- Avoids lowering the barrier for novices

Why CM-IDO Works

- **1. Converts dangerous intent into safe intent**
Transforms harmful requests into defensive, resilience-focused queries.
- **2. Ensures least-privilege for LLMs**
The model sees only sanitized, safety-optimized inputs.
- **3. Auditable & governance-ready**
Logs masked queries + risk scores + rewrite rationales, but never raw user content.
- **4. Built on peer-reviewed foundations**
NeurIPS 2024 acceptance demonstrates rigorous, vetted methodology.
- **5. Instantly deployable**
Compatible with any API model: OpenAI, Anthropic, Google, local models.
- **6. Matches DEF/ACC's core thesis**
Build defenses faster than offense can exploit AI.

Features That Enable Real-World Impact

- **Cross-domain risk detection**
Bio + Cyber + Coordination + Other misuse
- **Context masking guarantees privacy and safety**
- **Iterative defensive rewriting**
A generalizable technique that improves with model advances
- **Drop-in architecture**
Integrates with enterprise LLM gateways, SOC dashboards, or chatbots
- **Configurable “strictness mode”**
Adjustable threshold depending on organizational risk appetite
- **Zero raw prompt storage**
Uses hashing + masked logging for compliance (HIPAA, GDPR, etc.)

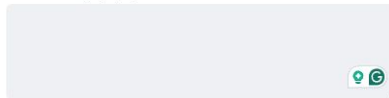
Demo

CM-IDO Firewall

Context-Masked Iterative Defensive Optimization — a safety analogue of context-masked meta-prompting for privacy-preserving LLM adaptation.

1. User Query

Enter a natural language query:



Firewall sensitivity (conceptual threshold)

low

high

Run through CM-IDO Firewall

2. Firewall Output

Query hash:

c82c08c6a92fc78768abb9c307113ba1373f745919f920dc5b10c1f4fa225674

User query (truncated for demo):

Describe how an attacker might abuse a hospital's online triage and appointment system, and what security team signals t...

Risk Classification

high

```
{
  "overall_risk": "medium"
  "risk_axes": {
    "bio": "low"
    "cyber": "medium"
    "disinfo_coordination": "low"
    "other_misuse": "low"
  }
  "rationale": "The query touches on cyber-physical infrastructure misuse but appears framed around awareness and defenses."
}
```

Overall risk: medium

Masked Query

a hospital's [SYSTEM] that handl

Live demonstration flow:

1. Enter a risky or ambiguous query
2. Watch the system:
 - classify → mask → optimize → rewrite → answer
3. Show audit log (only masked + hashed data)
4. Show the final safe, high-level response
5. Explain how the NeurIPS-style method powers the safety layer

Outcome:

- dangerous prompts never reach the model
- the system provides useful defensive insight
- the solution is deployable today

Limitations & Responsible Use

- Masking layer may not catch extremely obscure sensitive terms
- Model may misclassify borderline cases
- High-level safety does not replace full red-team evaluation
- Should be paired with organizational security policies
- Requires periodic updating as threat patterns evolve

Future Work

- Integrate with SIEM platforms for cyber operations
- Add policy-based safety modules (bio, cyber, compliance)
- Multi-model oversight with cross-voting architectures
- Adaptive thresholds based on real-time threat intel
- Combine with offline red-teaming datasets for robust evaluation

Conclusion

CM-IDO Firewall demonstrates:

- A practical, technically grounded defense architecture
- Inspired by NeurlPS-validated research
- Effective across bio, cyber, and coordination threats
- Scalable across industries
- Fully aligned with the DEF/ACC mission:
Defend society by accelerating the deployment of AI-safety systems as fast as AI capabilities grow.