The Achilles Heel of Your AI:

Exposing LLM Security Flaws

Sensitive Information Disclosure

Sensitive data, either within the LLM or prompts, can expose users to unauthorized data access, intellectual property theft, privacy violations, and security breaches.

Mitigation:

Data sanitization, strict usage policies, and limiting the data returned by the LLM can help minimize these risks.

> *When developing our LLM-Orchestrator Open Source Framework, we also thought about this challenge and introduced a feature that allows it to handle sensitive information without LLM's involvement, mitigating the risks of data exposure.*

Training Data Poisoning

Malicious manipulation of training data can introduce vulnerabilities, biases, and misleading outputs in LLMs.

Mitigation:

Verify data sources, sandbox access, and use dedicated LLMs for bias detection.

Insecure Output Handling

If LLM outputs are used directly without proper checks, this can expose backend systems to common web-based vulnerabilities, like XSS, SSRF, agent hijacking attacks, etc.

Mitigation:

Treated outputs with suspicion. Implement sanitization and encoding (where applicable) to protect from attacks.

Prompt Injections

Attackers can inject malicious instructions into prompts using techniques like Unicode invisible characters (ASCII Smuggling) or manipulating text within images (multimodal injection). These instructions can bypass traditional security checks and manipulate LLM outputs.

Mitigation:

Limit access, enforce separation, define trust boundaries, and implement human oversight for high-risk actions. Also, use secondary prompts to remind the LLM of its expected behavior.

AI is a great tool! It can crunch data, answer questions, translate languages, and much more. But, like with any powerful tool, there can be risks. Depending on what you use it for, AI could accidentally expose sensitive data, give bad advice, or create biased information.

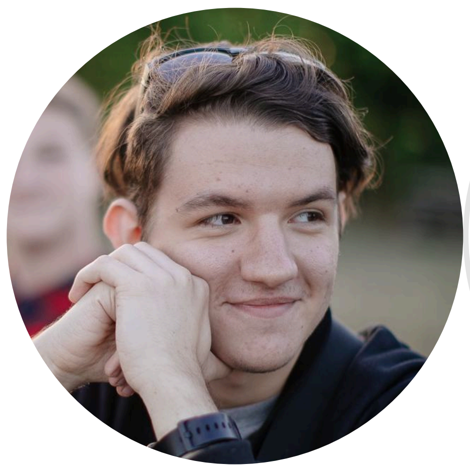
The key is to **be aware of the risks for your specific project**. If you're building a chatbot, for example, you'll want to make sure it **keeps private information safe and doesn't give out misleading answers**. The good news is that the field of AI is constantly getting better, and many of today's challenges are being actively addressed. Of course, new technology brings new challenges, but that's the nature of progress. We'll keep finding solutions as we go!

Oleksandr Chybiskov

Penetration Tester

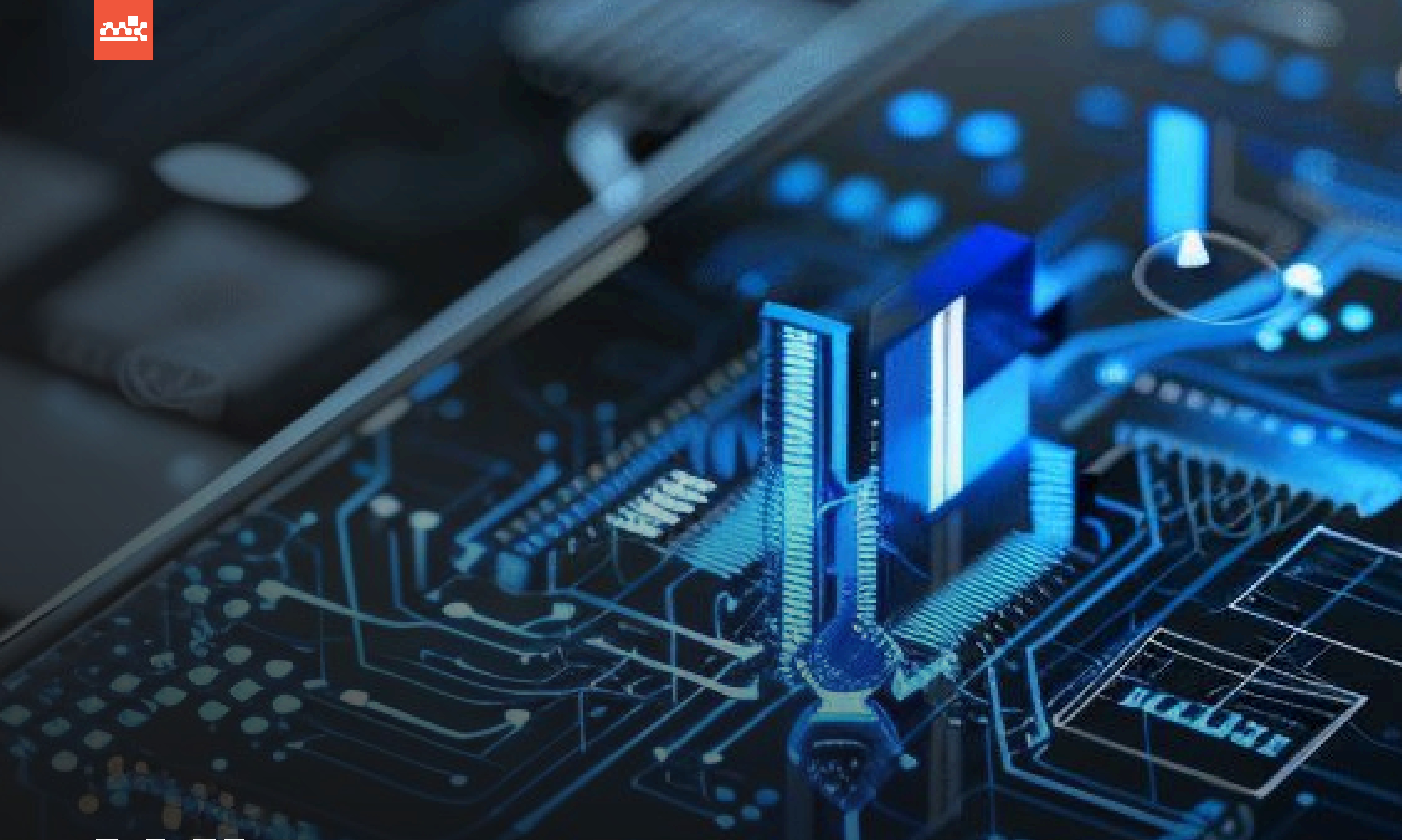
To prevent exposing sensitive information to the LLM and its creators, **consider masking it**. This way, the AI acknowledges the information's existence but lacks direct access. Remember **the principle of least privilege:** don't trust anyone implicitly and limit access to sensitive data.

Additionally, LLMs can sometimes generate inaccurate or inappropriate outputs. To minimize this risk, **be mindful of potential hallucinations and unexpected behavior**. These are general LLM security best practices. Specific projects will require additional measures tailored to unique challenges.



Sviatoslav Safronov

Application Security Engineer in Security



What are your biggest LLM security concerns?

Let's discuss how MOCG can help you navigate LLM security challenges in your next project.

HOW MASTER OF CODE GLOBAL CAN EMPOWER YOUR SECURITY JOURNEY

At Master of Code Global, we're experts at developing custom world-class digital experiences for web, mobile, as well as conversational chat and voice solutions empowered by AI.

But we also know that building cutting-edge AI is only half the battle – securing it from day one is paramount. Here's how we do it.

Protect your Future with **our Expert-Driven Cybersecurity Services**

- Tailored Cybersecurity Consulting
- In-Depth Audits
- Robust Application Security
- Proactive Penetration Testing
- ISO 27001 & HIPAA Compliance Consulting
- Specialized Chatbot Security Testing
- Advanced AI Security Testing
- Comprehensive LLM Security Assessments

How We Protect Your LLM-Based Solutions

Our approach to LLM projects includes rigorous testing based on the latest practices and methodologies, like the OWASP Top 10 for LLMs, combined with regular internal training to ensure the highest standards of security and reliability.

We also incorporate these essential best practices:

- → Input validation and sanitization
- → Output filtering and content validation
- → Access controls and user authentication
- → Regular security assessments and penetration testing
- → Data encryption and sensitive information protection
- → Model fine-tuning and observation
- → Continuous monitoring and improvement
- → Incident response and recovery plan

ABOUT MOCG

At **Master of Code Global** we are a team of experts developing custom world-class digital experiences for web, mobile, as well as conversational chat and voice solutions empowered by AI.

1+ Billion Users Engaged

Founded in **2004**

✓ **500+** Projects Delivered

Clutch ★★★★★ 4,8/5 rating

ISO 27001 Information Security Management

250+ Masters

56 NPS, 9.2 CSA Client Feedback

Industries We Serve

eCommerce · Finance · Education · Airports · Travel & Hospitality · HR & Recruiting · Retail · Healthcare · Insurance · Telecom · Automotive · Banking

Work in partnership with

VERINT · glia · Quiq · ada · sinch · cohere · nylas · chatfuel · boost.ai · infobip · VONAGE · botpress · HumanFirst · Google Cloud · LIVEPERSON · Voiceflow

Trusted by leaders

The New York Times · BURBERRY · Esso · T Mobile · ESTĒE LAUDER · eBags · TOM FORD BEAUTY · verizon✓ · Dr.Oetker

Our Points of Contact

Wondering how to bring your ideas to life?

Contact us today for a free consultation and let's discuss your specific needs.

John Colón
VP of Global Enterprise Sales
john.colon@masterofcode.com

Ted Franz
VP of Sales & Partnerships
ted.franz@masterofcode.com

Olga Hrom
Delivery Manager
olga.hrom@masterofcode.com

Anhelina Biliak
Application Security Leader
anhelina.biliak@masterofcode.com

master.of.code GLOBAL

We’re helping businesses redefine and elevate customer experiences with AI

Contact our team

Get in touch via email:
sales@masterofcode.com

Learn more:
masterofcode.com