Threat Modeling for AI Systems

Frameworks: STRIDE + MITRE ATLAS
Format: 8 groups, 2 presenters per group, 4 hours total

What You Are Doing

Today you will threat-model an AI system as a team.

Each group will:

- Analyze one assigned trust boundary
- Identify likely threats using STRIDE
- Connect AI-specific threats to MITRE ATLAS
- Prioritize the most important risks
- Present the strongest findings

You have received a separate **system brief** for your assigned system. That brief contains the architecture diagram and the hints for your group's boundary.

Learning Goals

By the end of the exercise, you should be able to:

- Explain what a trust boundary is
- Use STRIDE to generate threats systematically
- Recognize AI-specific attack paths such as prompt injection, poisoning, unsafe tool use, and data leakage
- Compare threats using a simple risk model
- Propose practical mitigations

Quick Reference

STRIDE

Use STRIDE as a checklist while brainstorming.

Category	Ask yourself...
Spoofing	Can someone pretend to be a trusted user, service, or component?
Tampering	Can data, prompts, model context, or requests be changed?
	Can someone deny what they did because logging or attribution is

Repudiation	weak?
Information Disclosure	Can sensitive information be exposed to the wrong party?
Denial of Service	Can the system be slowed down, overwhelmed, or made unavailable?
Elevation of Privilege	Can someone gain more access than they should have?

Why AI Changes the Threat Model

- Inputs are not just data; they can contain **instructions**.
- Outputs may look authoritative even when they are wrong or unsafe.
- Retrieved documents, memory, and tool results can all shape model behavior.
- AI systems often act through tools, which turns bad outputs into real actions.

How to Think About Threat Modeling

When you do threat modeling for this exercise, focus on the **inherent risk of the system design and components**.

- Do not assume the engineers are unusually strong or unusually weak
- Do not assume the code is perfect or buggy unless the architecture implies it
- Do not dismiss a threat just because "good engineers would handle that"
- Do not spend much time debating whether a mitigation already exists unless it is explicitly shown in the system description
- Focus on risks that come from using components with known classes of problems, such as LLMs, RAG, tools, external APIs, memory, or policy engines
- Treat model family or version as a detail, not a defense; the risk category usually still exists whether it is GPT-3, GPT-5, or another model

Threat modeling is about asking:

- What could go wrong because this component or boundary exists at all?
- What kinds of attacks become possible when these pieces are connected?
- What would we need to add or change to reduce that risk?

MITRE ATLAS

Use MITRE ATLAS when you want to name an AI-specific attack technique. As we've discussed in class, examples include:

- Prompt injection
- Context poisoning
- Training or data poisoning
- Tool misuse
- Exfiltration through model or agent behavior

Reference: https://atlas.mitre.org/ and https://atlas.mitre.org/matrices/ATLAS

Deliverable

Create one shared Google Sheet for your group and use it throughout the exercise.

Boundary	STRIDE	Threat	ATLAS	Likelihood (1-3)	Impact (1-4)	Risk	Mitigation

Risk score = Likelihood × Impact

Score Element	Scale
Likelihood	1 = Low, 2 = Medium, 3 = High
Impact	1 = Low, 2 = Medium, 3 = High, 4 = Critical

You should leave the session with:

- 3-4 credible threats from your boundary
- 1 highest-risk threat analyzed in depth
- 1-2 additional threats ranked for comparison
- 2-3 concrete mitigations for the highest-risk threat

Your shared Google Sheet should clearly show:

- The biggest risk your group identified
- The ATLAS technique for that risk, if relevant
- Why you ranked it highest
- What you recommend fixing first

Workflow

1. Read and Orient

Read:

- This handout
- Your assigned system brief only

Make sure your team understands:

- What components are involved
- What crosses your boundary
- What security decisions happen at that boundary

2. Brainstorm Broadly

Go through all six STRIDE categories.

Target:

- 3-4 total threats across your boundary
- Short, specific threat statements
- Coverage across multiple STRIDE categories, not necessarily every single one

Good example: `Untrusted document injects hidden instructions that later influence tool calls.`

Weak example: `The AI is insecure.`

3. Go Deep on the Highest-Risk Threat

For your highest-risk threat:

1. Describe the attack path step by step.
2. Map it to a relevant ATLAS technique.
3. Score likelihood and impact.
4. Propose 2-3 mitigations focused on engineering and security work.

For your next 1-2 threats:

1. Give a short description.
2. Assign a rough risk score.
3. Decide whether they are Must Fix, Should Fix, Monitor, or Accept.

4. Prioritize

Rank your threats:

- `Must Fix`
- `Should Fix`
- `Monitor`
- `Accept`

Do not optimize only for the highest numeric score. Also consider:

- Exploitability
- Blast radius
- Ease of detection
- Cost and speed of mitigation

5. Present

Each group gets **5 minutes total**.

Exactly **2 people** should present:

- Presenter 1: Boundary, highest-risk threat, and attack walkthrough
- Presenter 2: Debate, prioritization, and top recommendation

Use the attached presentation format.

Your presentation should cover:

- Group number and fruit name
- System name
- Highest-risk threat
- ATLAS technique
- How the attack works
- Business impact
- Most interesting team debate
- `Must Fix` vs. `Accept` decision
- One top recommendation for the organization

Suggested Team Roles

Use these roles if helpful:

Role	Focus
Facilitator	Keeps time and keeps discussion moving
Threat Hunter	Pushes for realistic attack paths
Risk Lead	Scores impact and likelihood
Scribe	Captures decisions clearly
Engineering Lead	Pushes for realistic mitigations in system design and implementation
Security Lead	Challenges assumptions and identifies control gaps
Presenter 1	Owns the first half of the presentation
Presenter 2	Owns the second half of the presentation

Your group has 8 people, so you should be able to assign one owner per role.

4-Hour Schedule

Block	Time	Duration
Welcome, framing, and team assignments	0:00-0:15	15 min
Read handout + assigned system brief	0:15-0:35	20 min
STRIDE threat brainstorming	0:35-1:20	45 min

Deep dive on the highest-risk threat	1:20-2:10	50 min
Break	2:10-2:20	10 min
Risk prioritization and finalize table	2:20-2:50	30 min
Presentation prep	2:50-3:10	20 min
Group presentations: 8 groups × 5 min	3:10-3:50	40 min
Wrap-up	3:50-4:00	10 min

What Good Work Looks Like

Good work looks like:

- An interesting discussion about an AI-based risk that we discussed in class, or one that we did not discuss in class, explained clearly to the rest of the room
- Specific analysis of what is attacked and how
- A clear explanation of why the risk matters in this system
- Focus on inherent system risk rather than assumptions about engineer quality or code quality
- Mitigations that engineering and security teams could realistically act on

Bad work looks like:

- A generic risk that could apply to almost any system
- A risk that cannot be mitigated in any practical engineering or security way
- Vague claims without a clear attack path, impact, or decision

---

Final Checklist

Before presenting, confirm that your group has:

- A clear description of the boundary
- 3-4 threats in the shared Google Sheet
- 1 fully worked highest-risk threat
- ATLAS mappings where relevant
- Risk scores
- Recommended mitigations with engineering and security actions
- 2 presenters ready to split the talk

Resources

- MITRE ATLAS: https://atlas.mitre.org/
- OWASP Top 10 for LLM Applications: https://owasp.org/www-project-top-10-for-large-language-model-applications/
- STRIDE Quick Reading: https://learn.microsoft.com/en-us/azure/security/develop/threat-