German OWASP Day 2025

The Automation Illusion?

What Machines Can't Do in Threat Modeling

Sebastien Deleersnyder,
Georges Bolssens,
Toreon

Georges Bolssens

- M.Sc. in Electro-Mechanical Engineering (2002)
- Developer -> IT Admin/DBA -> IT manager (2005-2015)
- Transitioned into Cyber Security in 2015
  - Freelance pentesting, security awareness training, Secure Coding
  - Security champion @ BEL20 Pharma company
  - Internal risk assessor / pen test coord. @ Big4 consulting
- Senior AppSec consultant @ Toreon (2022-current) for public and private sector
  - Secure code reviews
  - Threat modeling and -coaching
  - Cybersecurity teacher : in-house + various conferences
    - Blackhat (US), OWASP Benelux, Troopers (DE), NorthSec (CA), NDC (UK)

Sebastien Deleersnyder

- 5 years developer - 25 years information security
- CTO Toreon - Data Protection Institute
- **OWASP SAMM** project leader
- OWASP volunteer
- Co-founder **BruCON**

owaspbenelux.eu

Scaling Threat Modeling

Human-centric “messy” process

Too slow for modern development

Human interaction - slow

Security expertise - bottleneck

Different teams, different results

Until … AI?

Introducing AGI?

An intelligence capable of performing any intellectual task a human can, across a wide variety of domains

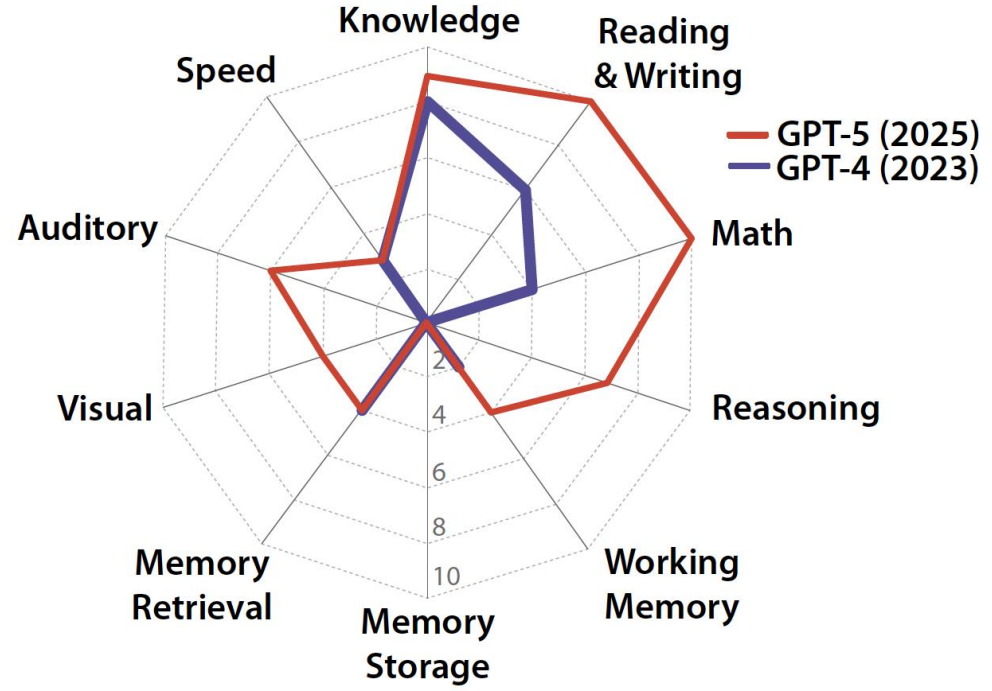


Figure 1: The capabilities of GPT-4 and GPT-5. Here GPT-5 answers questions in 'Auto' mode.

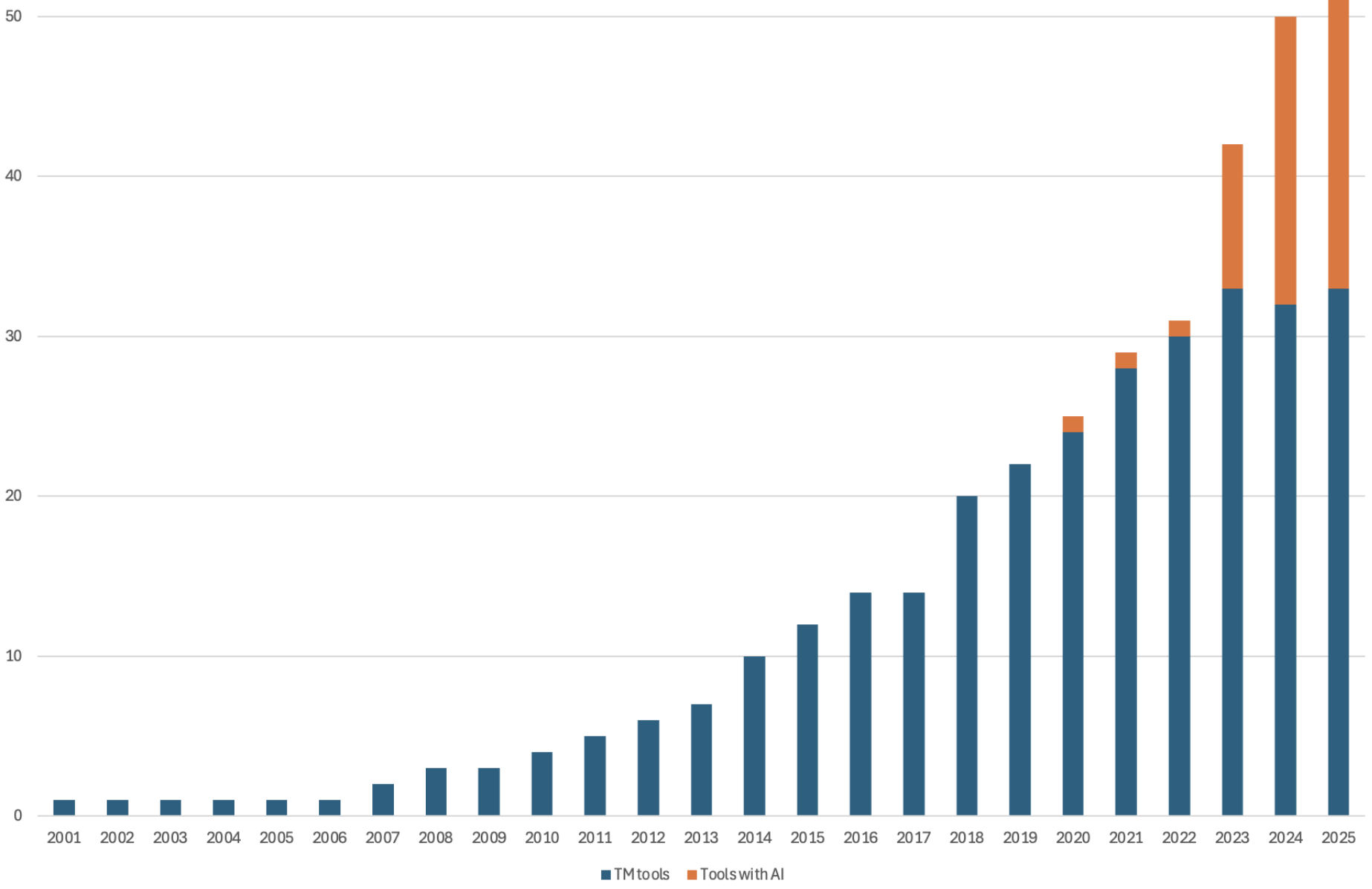
A Definition of AGI, arXiv:2510.18212v2 [cs.AI] 23 Oct 2025

TMGI ?

Threat Modeling General Intelligence

A system capable of replicating and surpassing the capabilities of the most expert human threat modelers across all TM domains.

The rise of (genAI) TM tooling



Threat Modeling Tool Directory

- ADTool
- Adversarial Robustness Toolbox (ART)
- Agent Wiz
- AI Security Analyzer
- Aribot
- Arrows
- AttackTree
- Attack Tree GPT
- AttackTree
- AT-AT
- CAIRIS
- CyberSage
- Deciduous
- Devici
- Dragon-GPT
- Ent
- ForkTM
- Gram
- Iriusrisk
- itemis SECURE
- Threat Modeling GPTs
- Microsoft Threat Modeling Tool
- Open Weakness & Vulnerability Modeler
- OWASP Threat Dragon
- PILLAR
- Prime
- PyTM
- Raindance
- RiskTree
- SAP Threat Modeling Tool
- SD Elements
- SeaMonster
- SeaSponge
- securiCAD
- SecurITree
- SecurityReview
- Seezo
- SPARTA
- StartLeft
- STRIDE GPT
- TaaC-AI
- td-ai-modeler
- Threagile
- Threat Composer
- Threat Designer
- ThreatCanvas
- Threatcl
- ThreatModeler
- ThreatPad
- ThreatPlaybook
- Threats Manager Studio
- Threatspec
- Threatware
- TicTaaC
- Tutamen Threat Model Automator

What is “threat modeling”, anyway?

The DICE framework

What are we working on?

What can go wrong?

What are we going to do about it?

Did we do a good enough job?

(*) Or, when using LLMs : Describe

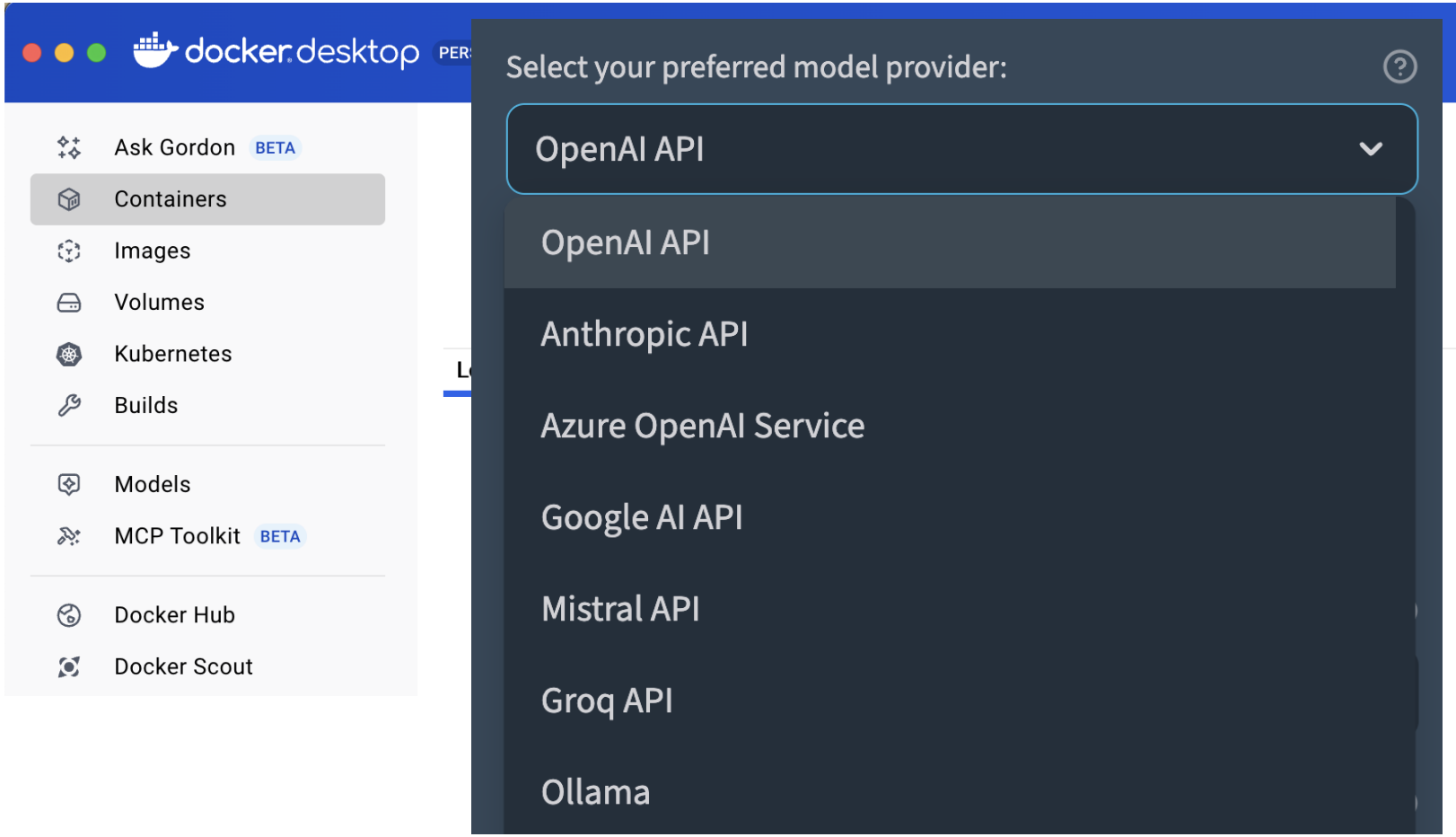
Diagram automation

https://www.eraser.io/ai/system-architecture-diagram-generator

STRIDE-GPT

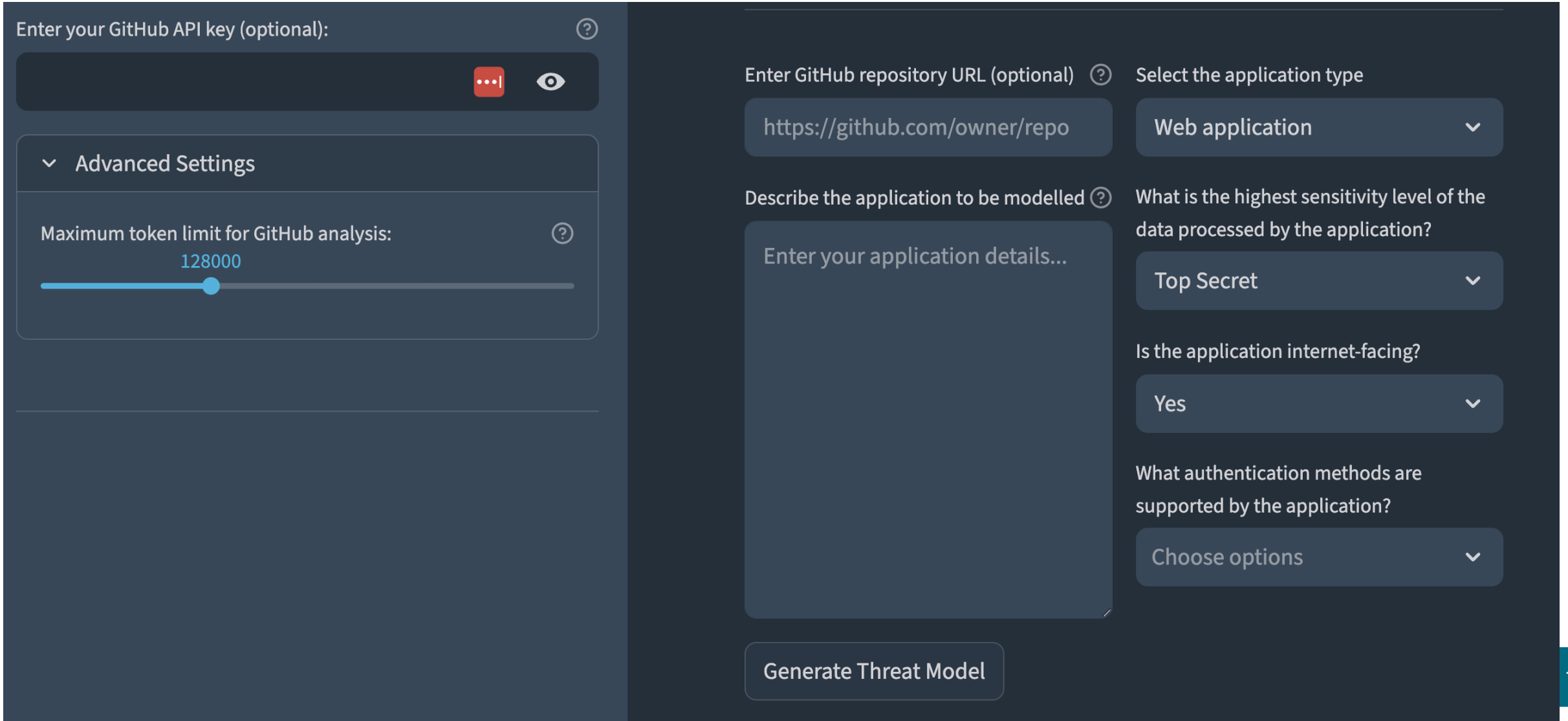
Local tool by Matt Adams to empower non-security experts in threat modeling

Can use online or offline LLM



STRIDE-GPT

Supports GitHub repo ingestion



STRIDE-GPT

Threat Model - generated

Improvement Suggestions

- Provide an architecture diagram with trust boundaries showing React frontend, Node.js API, real-time collaboration service (e.g., WebSocket/Socket.IO), MongoDB, and any reverse proxies or load balancers.
- Detail the OAuth2/OpenID Connect flow: libraries used, use of PKCE, state/nonce handling, redirect URI configuration, token validation (iss/aud/signature), and token lifetimes/storage.
- Clarify how MFA integrates with social login: which factors are supported (TOTP, WebAuthn, push), enrollment and recovery flows, when MFA is prompted, and how enforcement is applied across sessions and devices.
- Describe the real-time collaboration protocol: transport (WebSocket/SSE), authentication method for socket connections, per-message authorization checks, room/channel ID format, and how operations (OT/CRDT) are validated server-side.
- Explain the access control model: how ownership and sharing are represented in MongoDB (schema), roles (viewer/editor), and where authorization decisions occur in the backend.
- Specify whether encryption at rest is client-side or server-side; include key management details (KMS/HSM usage, key storage, rotation, wrapping, and access controls).
- Provide data flow details for sharing/invitations: how collaborators are added (by email, user ID, or links), invitation token format, expiry, binding to recipients, and acceptance workflow.
- List CORS and CSP configurations, including allowed origins, header exposure, and any third-party domains used by the frontend.
- Outline logging and auditing: which events are logged (auth, access, edits, sharing changes), log format, correlation IDs, retention, and tamper-resistance mechanisms.
- Describe input validation and database access patterns: use of Mongoose or other ODMs, parameterization, and protections against NoSQL injection.

e	Scenario	Suggested Mitigation(s)
	OAuth login CSRF due to missing/incorrect state and nonce validation. An attacker initiates an OAuth flow tied to their account and tricks a victim into completing it, causing the victim's browser session to be authenticated as the attacker.	- Use OAuth 2.1/OIDC Authorization Code Flow with PKCE; generate high-entropy state and nonce per login and bind both to the user session/server-side store - Verify returned state and nonce exactly; fail closed on any mismatch or reuse (maintain one-time-use cache of recent nonces) - Set redirect URIs to exact allowlist (no wildcards) and enforce response_mode=form_post to keep codes out of URL and Referer - Set auth cookies as HttpOnly, Secure, SameSite=Lax/Strict; add CSRF token for any login-init endpoint that mutates session - Display post-login interstitial showing the account to be linked/signed-in and require explicit user confirmation when the returned identity differs from existing session
	Improper verification of OAuth id_token (e.g., not	- Validate id_token per OIDC Core: verify issuer (iss) against discovery, audience (aud) equals your client_id, azp when multiple audiences, exp/nbf/iat with small clock skew, nonce when used - Fetch JWKS via OIDC discovery and cache keys;

queries).

insertion of attacker-controlled

TM-Bench

Benchmark for LLM-Based Threat Modeling using self-hostable models (also by Matt Adams)

Same prompts, same consumer grade hardware

Uses Claude3.7 Sonnet in the role of “LLM-as-a-judge” to score on

- STRIDE coverage 30%
- Threat completeness 30%
- Technical validity 30%
- JSON Compliance 10%

TM-Bench

Leaderboard

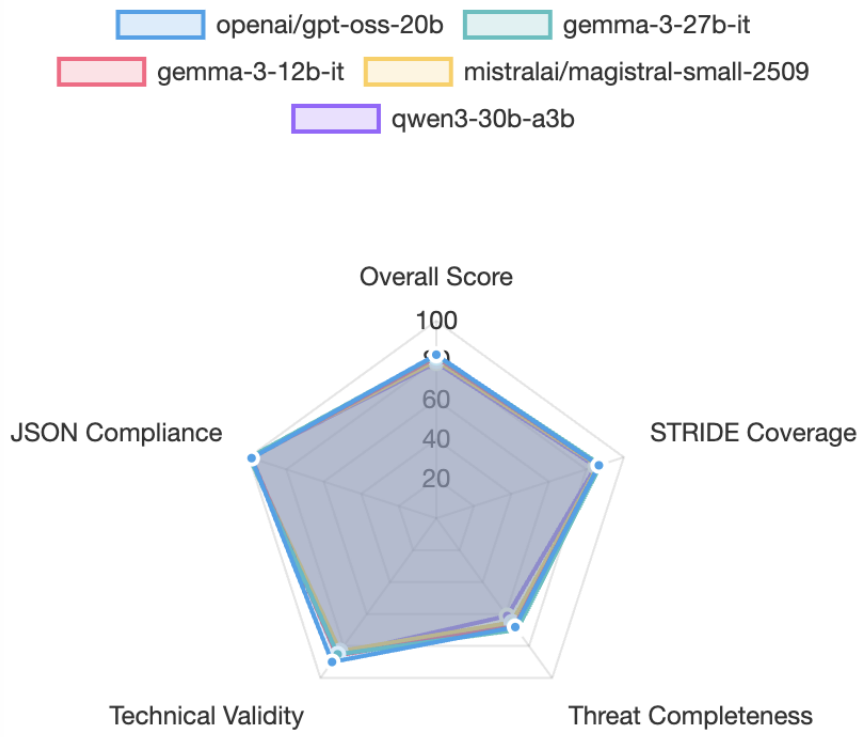
Search models or providers...

Overall ↓ | STRIDE | Complete | Valid | JSON

Rank	Model	Overall	STRIDE	Completeness	Validity	JSON
1	openai/gpt-oss-20b	82.8%	86.7%	68.1%	90.0%	98.6%
2	gemma-3-27b-it	82.5%	87.6%	69.9%	85.2%	99.7%
3	gemma-3-12b-it	81.6%	86.7%	67.6%	86.0%	98.6%
4	mistralai/magistral-small-2509	79.9%	86.2%	65.2%	83.2%	99.4%
5	qwen3-30b-a3b	78.8%	85.1%	61.2%	85.2%	99.6%
6	phi-4-reasoning-plus	78.6%	84.5%	61.0%	84.8%	99.5%
7	gemma-3-4b-it	77.4%	83.3%	62.2%	85.4%	88.5%
8	google/gemma-3n-e4b	77.1%	84.7%	61.0%	79.1%	99.0%
9	qwen3-8b	77.0%	83.8%	57.1%	85.3%	99.3%
10	qwen3-4b	75.6%	84.2%	55.5%	81.3%	98.1%

Performance Radar

Detailed metrics comparison for top models



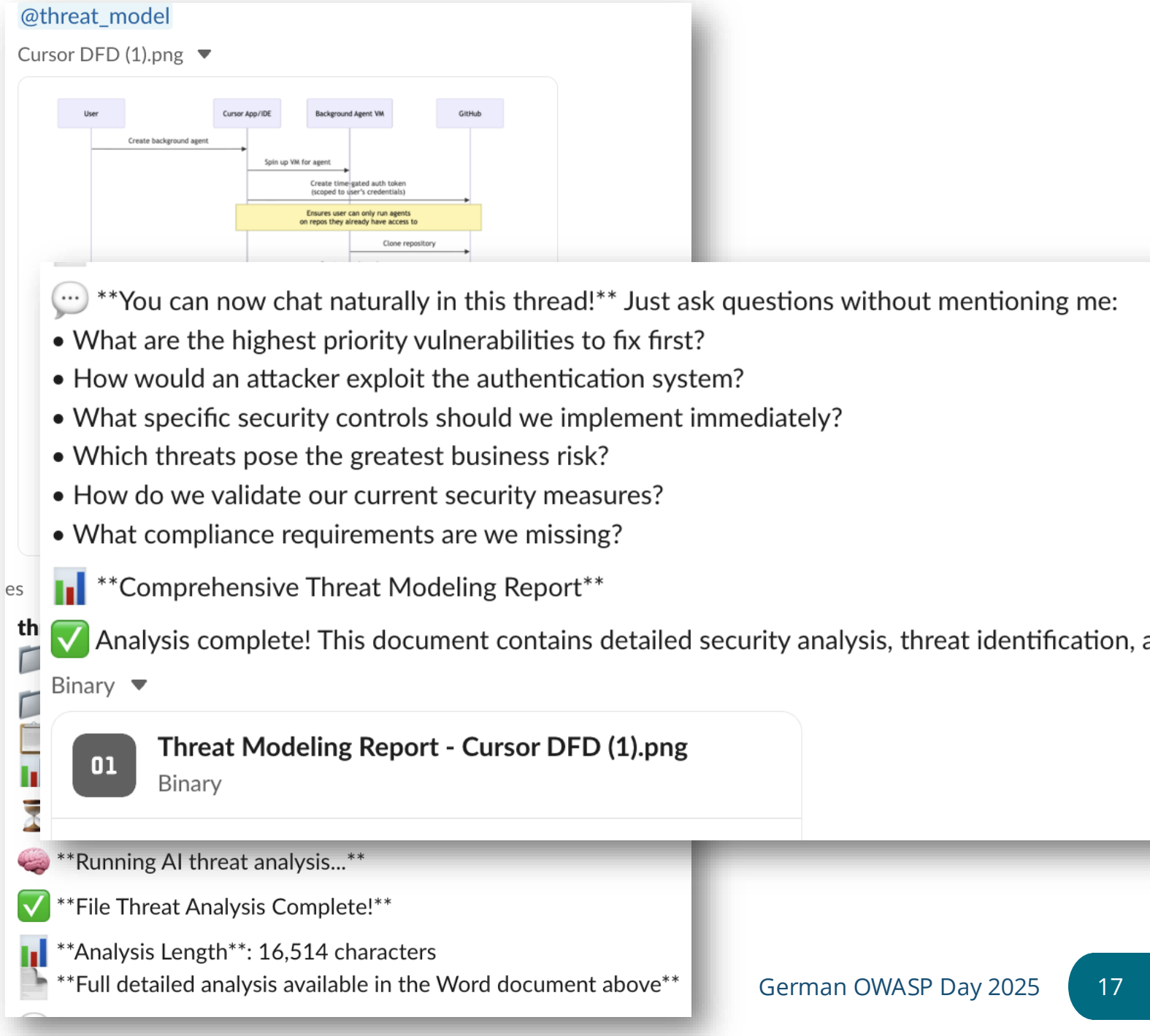
Integration in dev tools

Dipen Shah created a Slack-based tool to scale threat modeling

- 1st-level : bot in a dedicated Slack channel
- 2nd level : champions follow-up
- 3rd level : the central TM team validates

<u>Humans are the final approver</u>

Currently, as good as a “new trainee”
Next steps – provide it with org context, top 10 issues



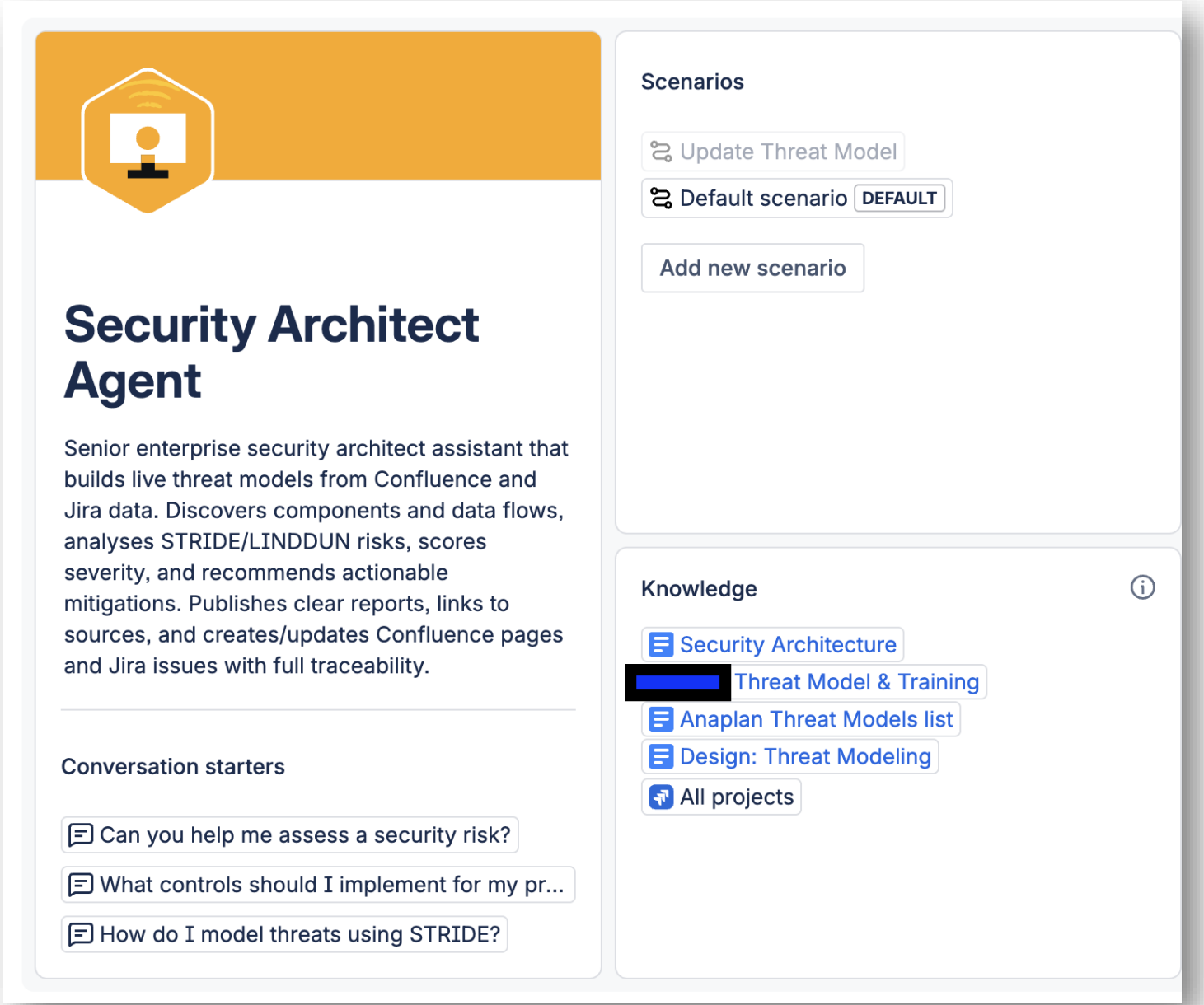
Integration in dev tools

Atlassian Rovo agent for a new project drafts (Lucian Corlan).

Based on the org methodology and examples

Scenarios within the agent package encapsulate the plan and tools, ensuring the same process runs consistently across reviews.

Disclaimer : Alpha state…



Industry example

JP Morgan Chase has developed/deployed AI to speed up threat modeling: "AI Threat Modeling Copilot" (AITMC)

AUSPEX: BUILDING THREAT MODELING TRADECRAFT INTO AN ARTIFICIAL INTELLIGENCE-BASED COPILOT

A PREPRINT

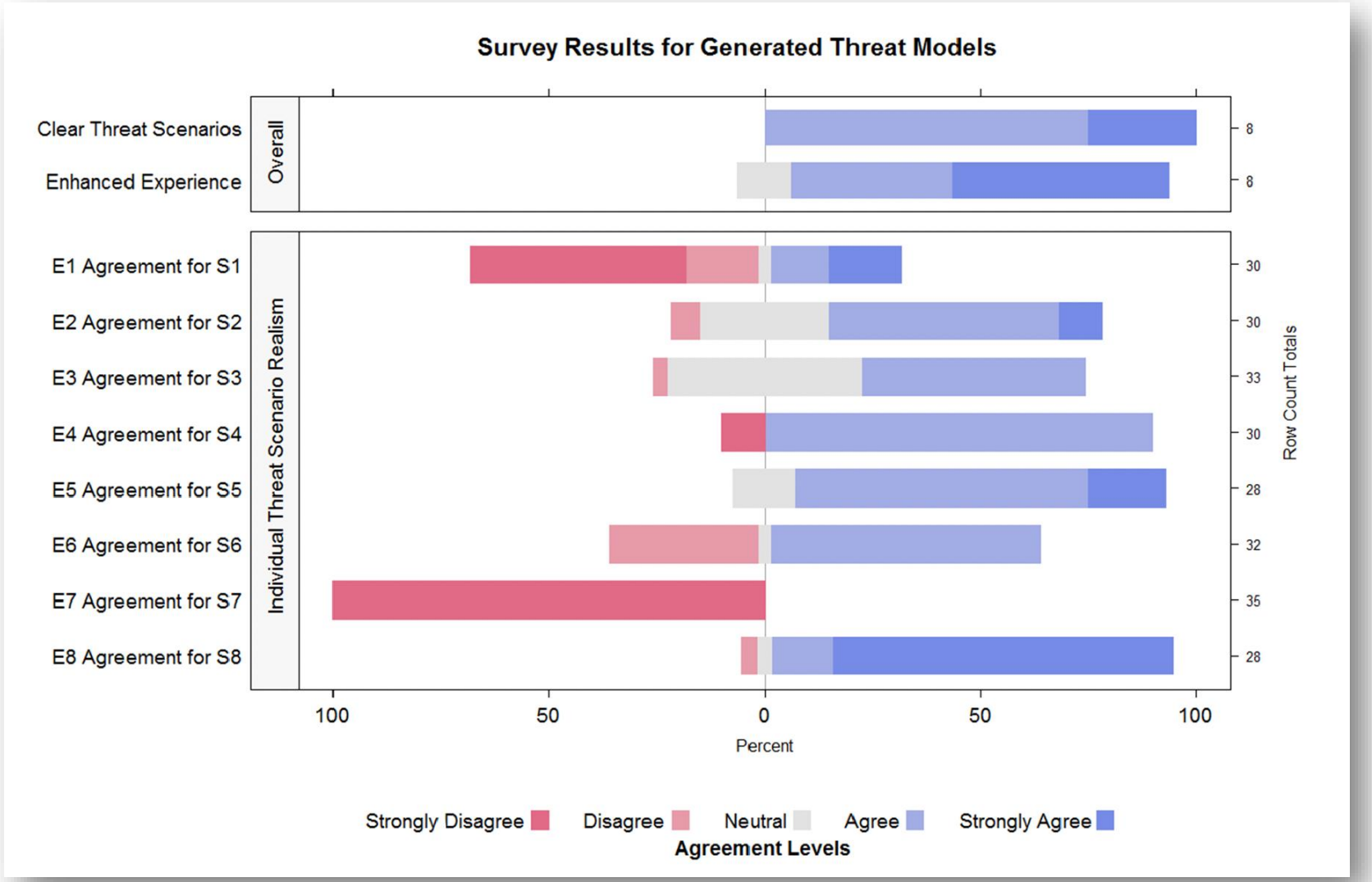
Andrew Crossman* JP Morgan Chase
Andrew R. Plummer JP Morgan Chase
Chandra Sekharudu JP Morgan Chase
Deepak Warrier JP Morgan Chase
Mohammad Yekrangian JP Morgan Chase

ABSTRACT

We present Auspex - a threat modeling system built using a specialized collection of generative artificial intelligence-based methods that capture threat modeling tradecraft. This new approach, called ***tradecraft prompting***, centers on encoding the on-the-ground knowledge of threat modelers within the prompts that drive a generative AI-based threat modeling system. Auspex employs tradecraft prompts in two processing stages. The first stage centers on ingesting and processing system architecture information using prompts that encode threat modeling tradecraft knowledge pertaining to system decomposition and description. The second stage centers on chaining the resulting system analysis through a collection of prompts that encode tradecraft knowledge on threat identification, classification, and mitigation. The two-stage process yields a threat matrix for a system that specifies threat scenarios, threat types, information security categorizations and potential mitigations. Auspex produces formalized threat model output in minutes, relative to the weeks or months a manual process takes. More broadly, the focus on bespoke tradecraft prompting, as opposed

https://arxiv.org/pdf/2503.09586

Industry example



https://www.jpmorganchase.com/about/technology/blog/aitmc

Industry example

"AITMC has driven 20% efficiency gain in our threat modeling process enabling faster models of new systems and broader scale

...

It helps engineers understand **WHY** controls are important not just **WHAT** they need to do to build secure systems",

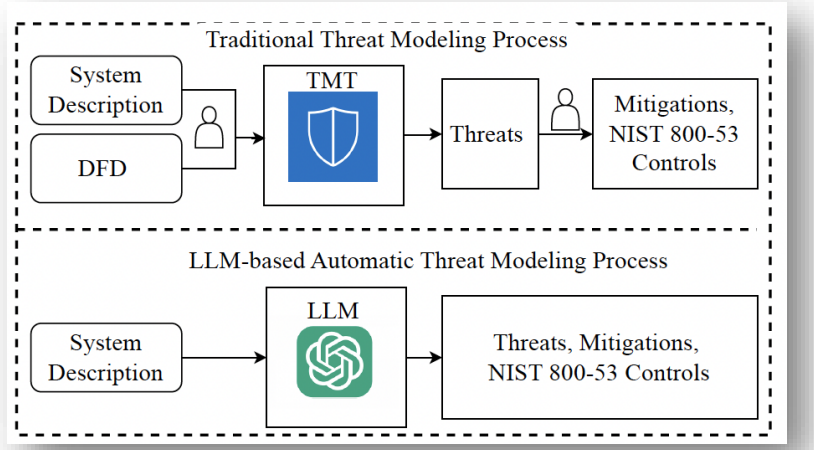
- Pat Opet, CISO JPMC

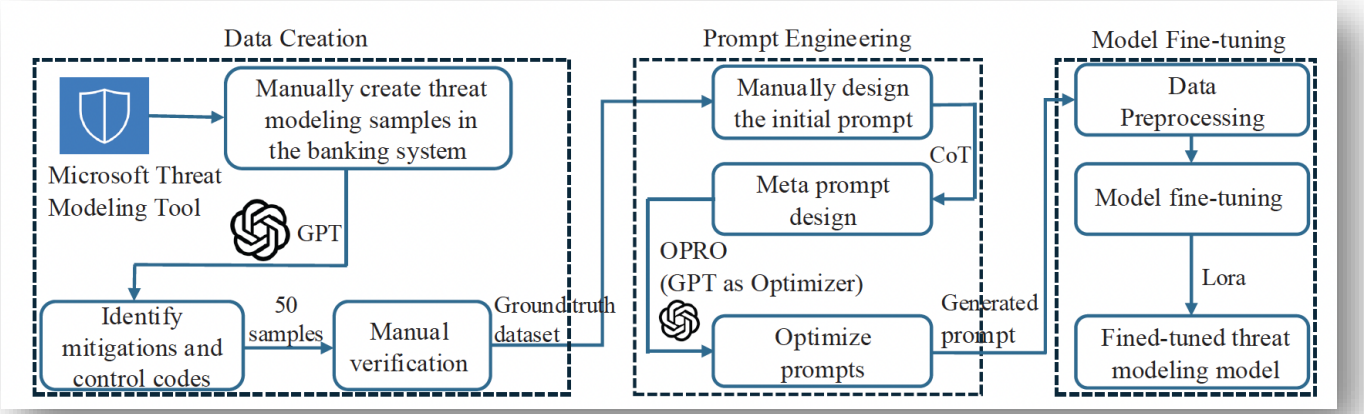
https://www.jpmorganchase.com/about/technology/blog/aitmc

More scientific research is ongoing

LLM-based threat modeling automates security from system descriptions to NIST controls, bypassing manual DFD and mitigation mapping.

The results illustrate that the combination of prompt engineering and fine-tuning techniques is highly effective for automated threat modeling.





ThreatModeling-LLM: Automating Threat Modeling using Large Language Models for Banking System,
arXiv:2411.17058v2 [cs.CR] 14 May 2025

AGI ?

An intelligence capable of performing *any* intellectual task a human can, across a wide variety of domains

Reality check …

Tooling is one (increasingly important) part

People, Process, and Technology

In a complex business context

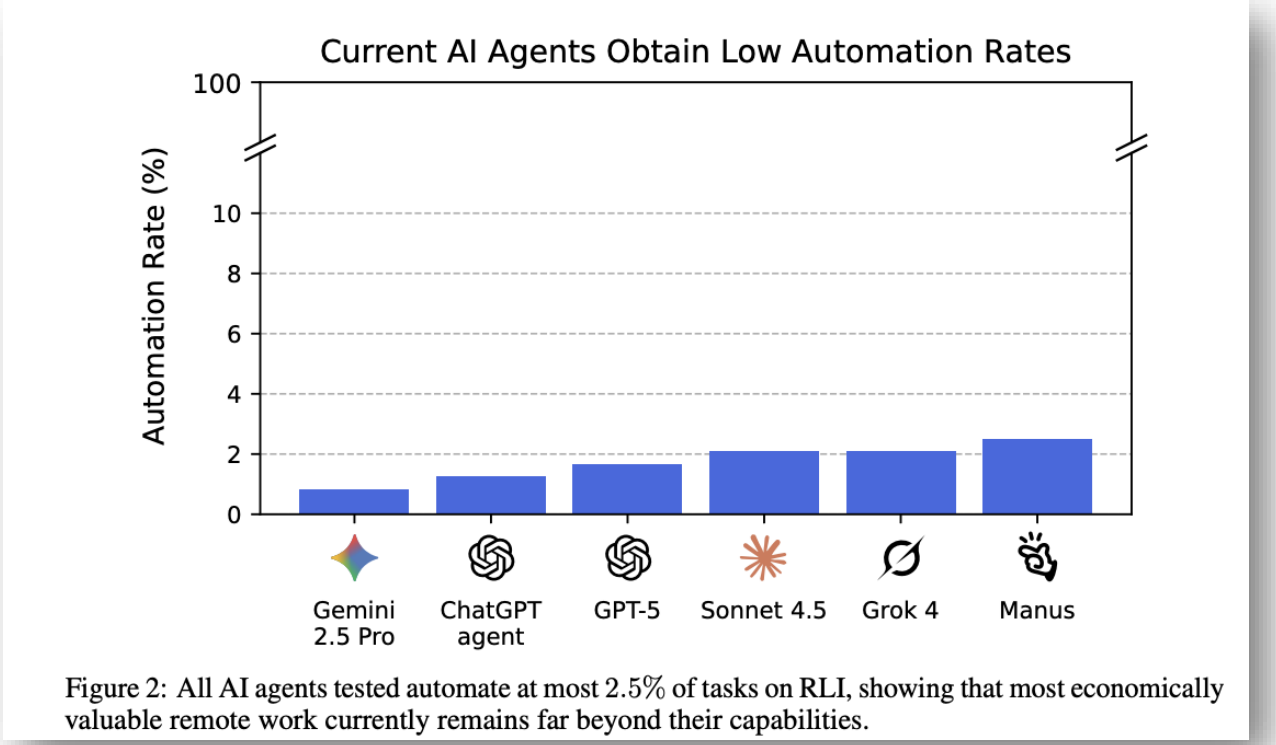


Figure 2: All AI agents tested automate at most 2.5% of tasks on RLI, showing that most economically valuable remote work currently remains far beyond their capabilities.

Remote Labor Index: Measuring AI Automation of Remote Work, arXiv:2510.26787v1 [cs.LG] 30 Oct 2025

Current state

Rapidly moving from TM-checklists to TM-Automation

We can now leverage AI as a mighty copilot

Humans can create with originality and vision, drawing on subjective feelings and cultural nuances.

Whereas AI is a powerful tool that lacks consciousness, emotions, and lived experiences.

The automation illusion

Automation supports, does not solve TM

Developers must own the process

The hard part is securing buy-in

Threat modeling is a soft skill

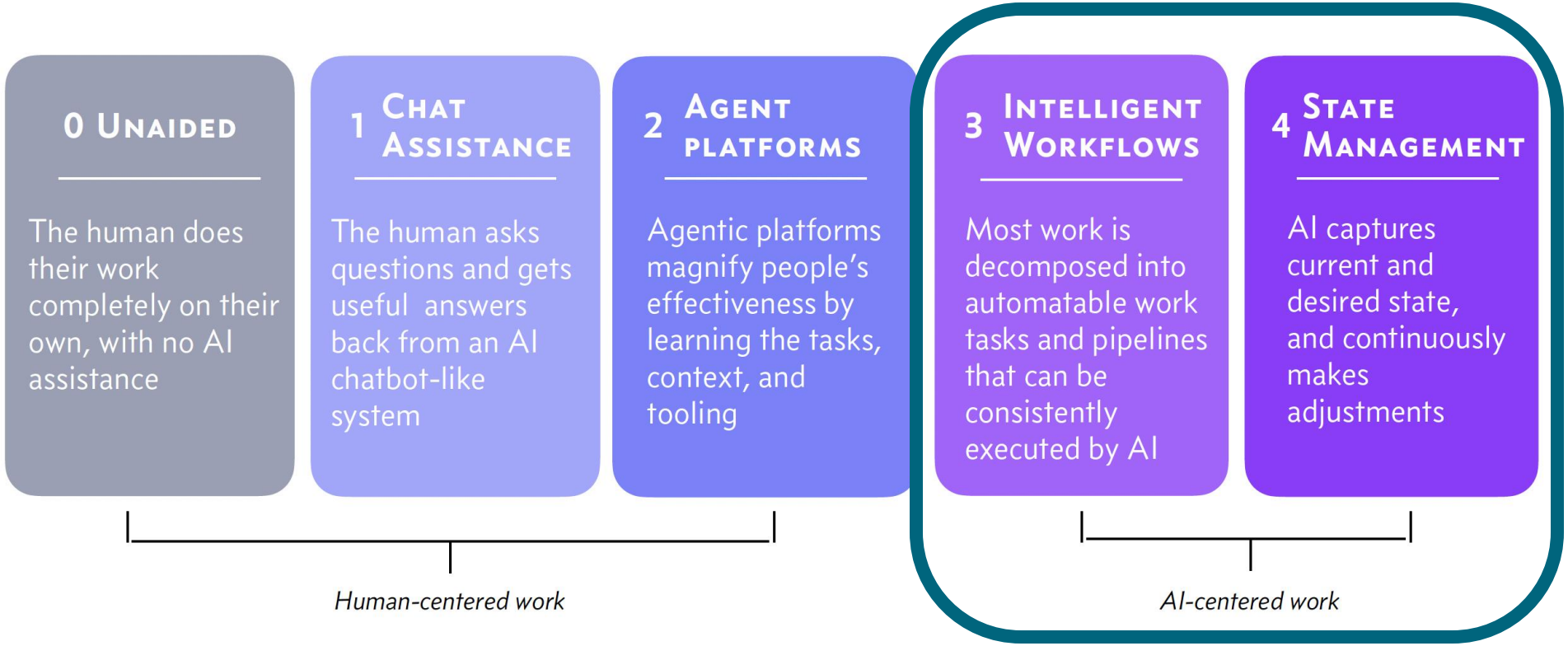
Process is key

Human centric threat modeling

AI augments rather than replaces human creativity

The future of TM is a collaboration between the two

How do we get here?



Daniel Miessler, OWASP DC25 Keynote

Threat Model like a surgeon

Increasingly essential, you:

- Make the final decisions
- Lead human / AI TM team
- Handle the complexity

Like surgeons, our expertise requires constant study, practice, and engagement with the body of knowledge on threat modeling.

Threat Modelers, Reinvented

AI Skeptic	• dabbling, small experiments, with low expectations
AI Explorer	• using AI for basic diagramming, STRIDE, and prompt iteration
AI Collaborator	• co-creating, multi-step prompting, agent orchestration
AI Strategist	• designing large agent workflows, focusing on delegation + verification

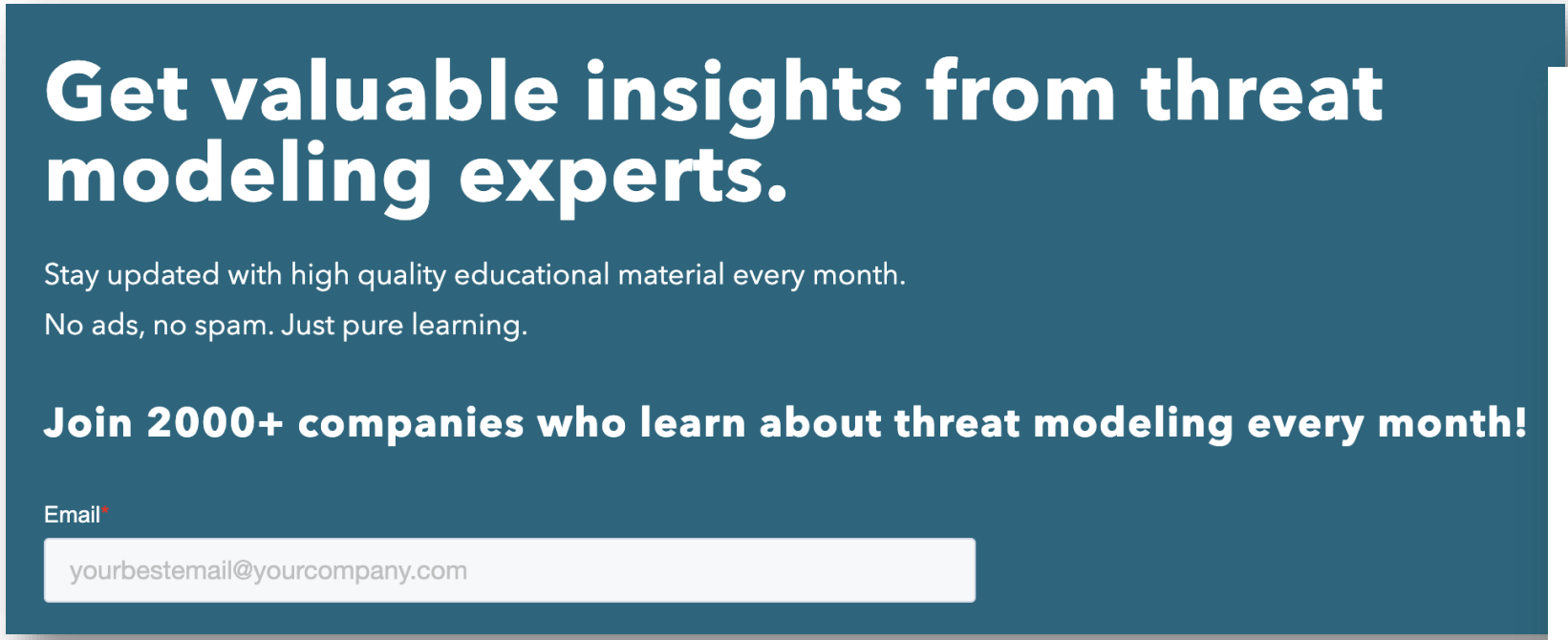
A Skill Shift, Not Elimination

Learn & leverage AI

Verify & own decisions

Increase your impact

Threat Modeling Insider - Newsletter - Register



Early access Threat Modeling Tool Directory

https://www.toreon.com/tmi-threat-modeling/

German OWASP Day 2025

Thank you