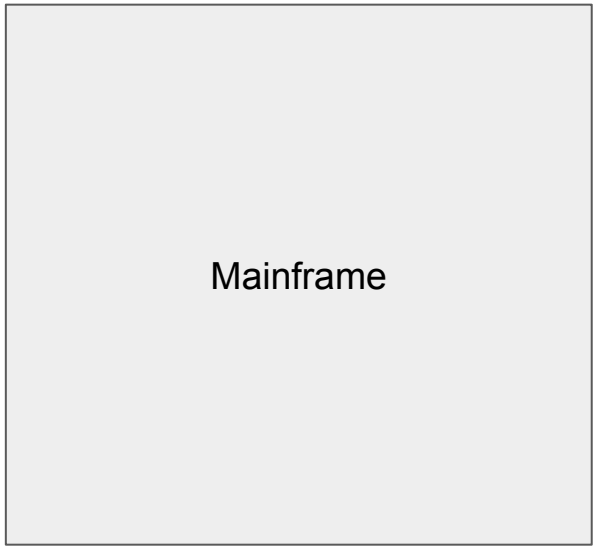
The Promptware Kill Chain

Summary and Thoughts on Bruce Schneier's Paper

Overview

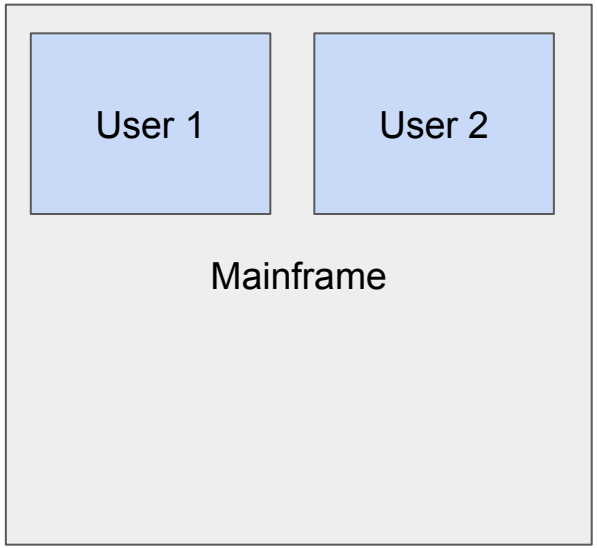
- Security History
- LLM Security History
- Lethal Trifecta
- Promptware Kill Chain
- Examples
- Defense

Security History



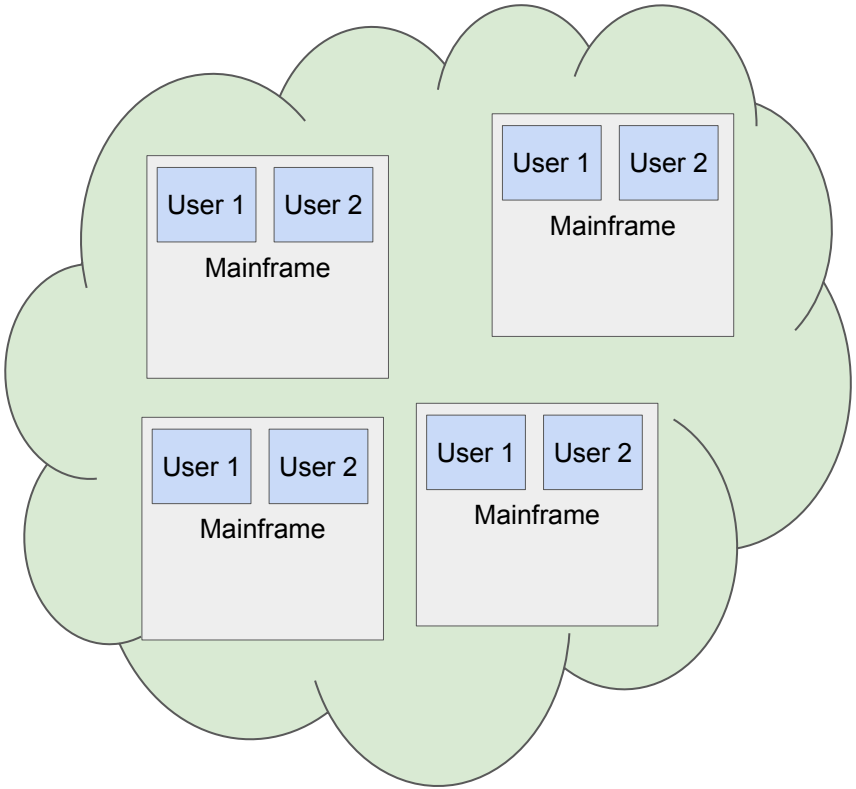
Everybody's known and why would you do something bad?

Security History



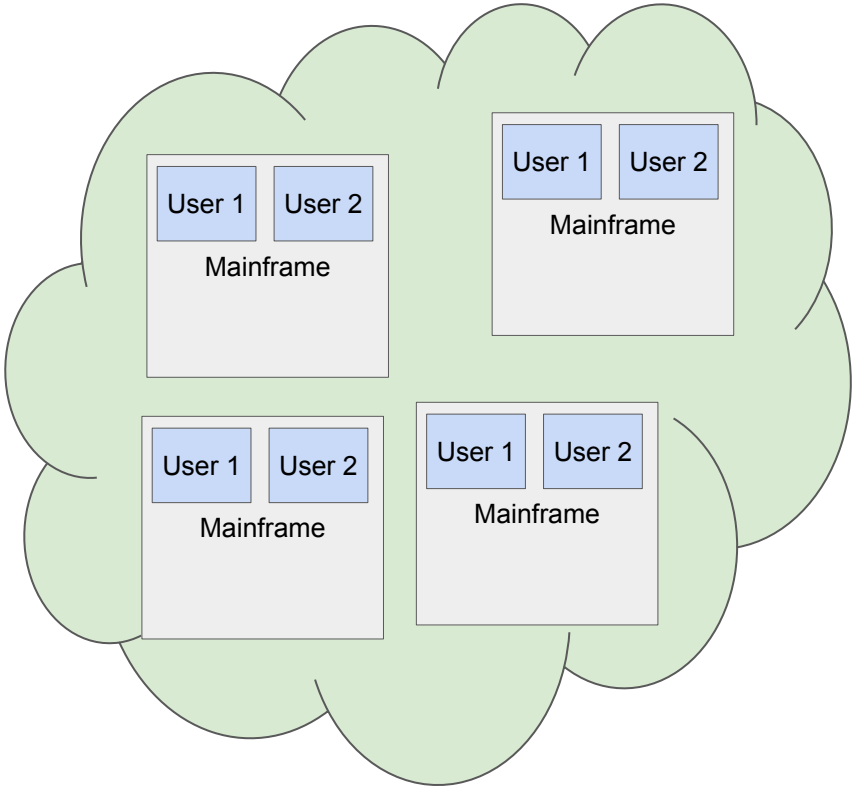
Let's do some access control for users

Security History



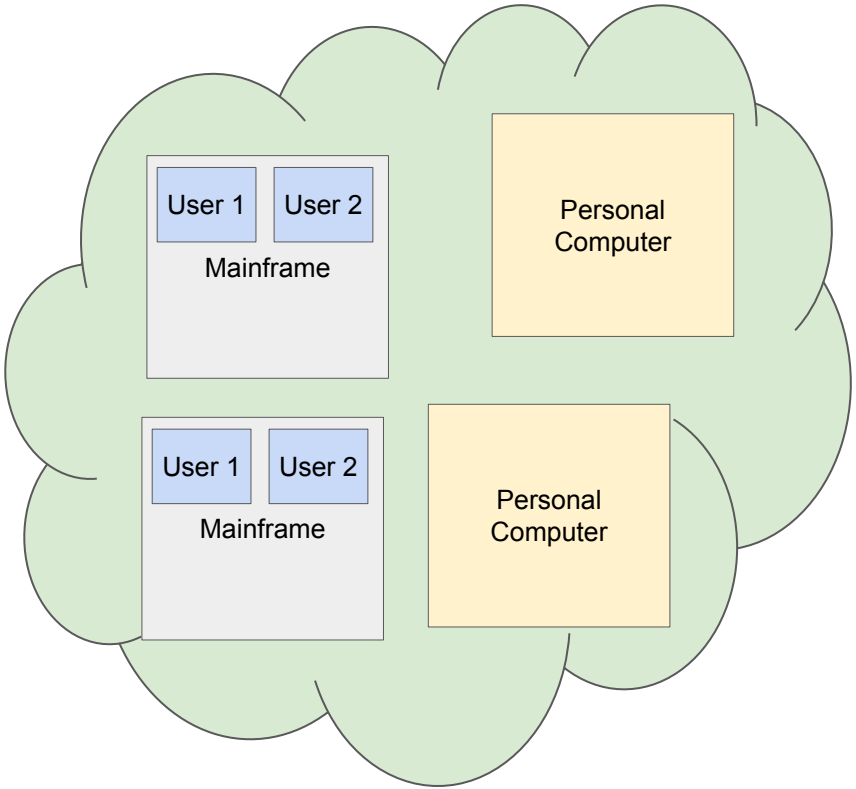
Nice - the Darpanet allows us to connect computers - what can go wrong?

Security History



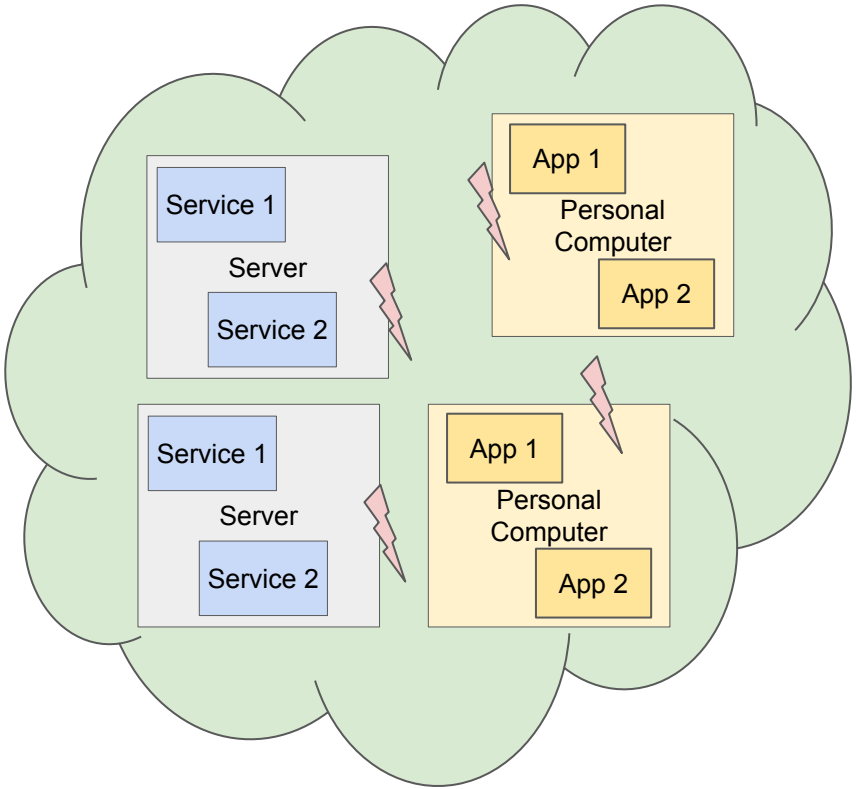
After the morris worm: make sure passwords are not easily guessable

Security History



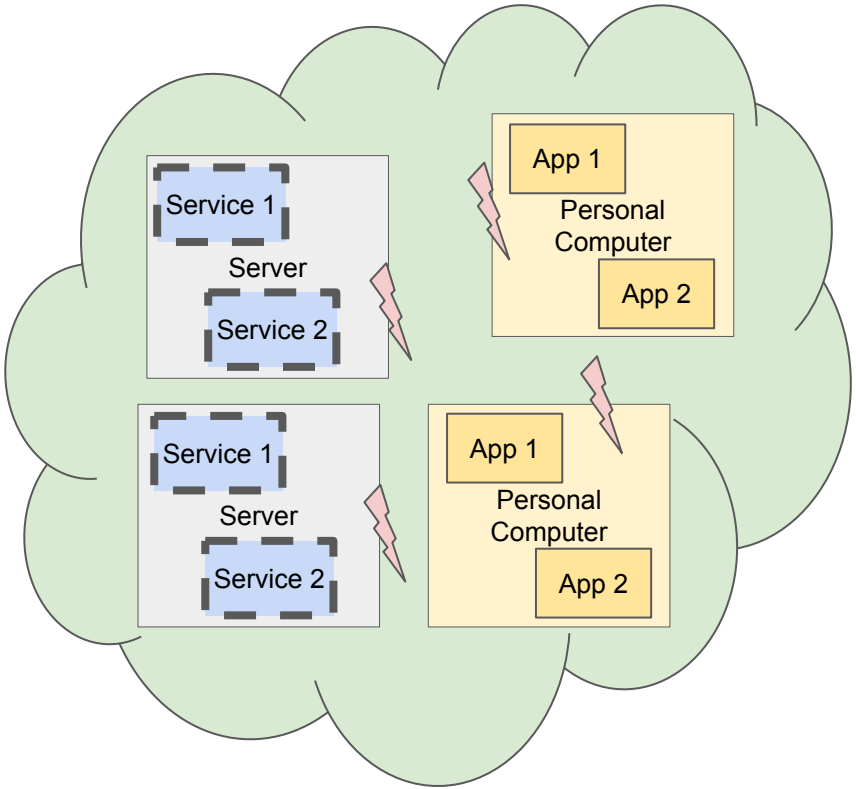
Great - everybody has their own computer now!

Security History



We have internet and servers with services - let's add some firewalls

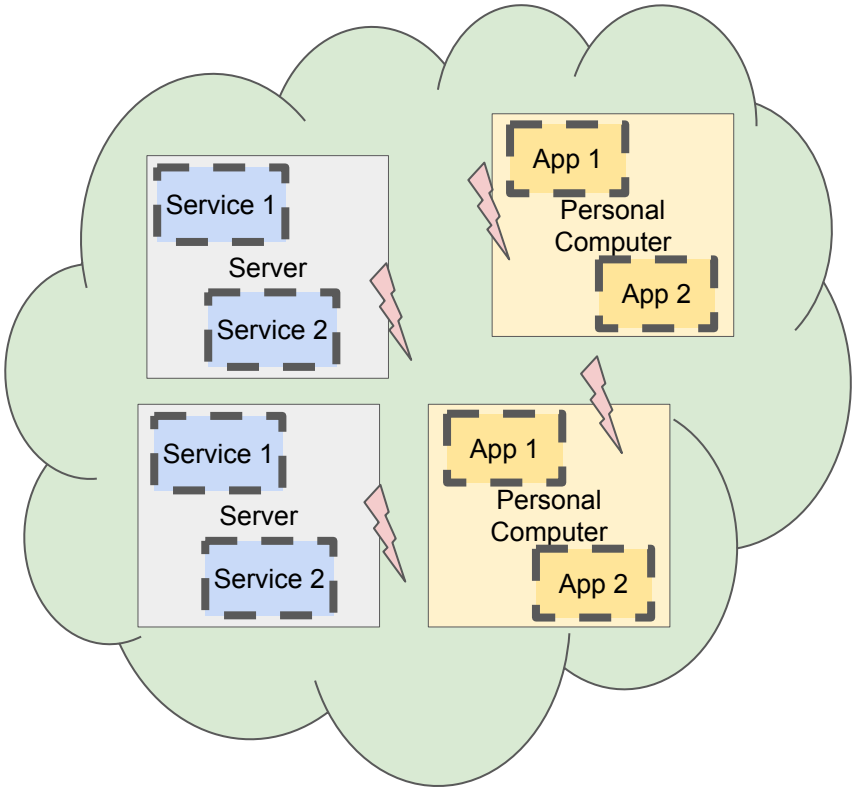
Security History



Compromised service ->
Compromised server

Virtualization

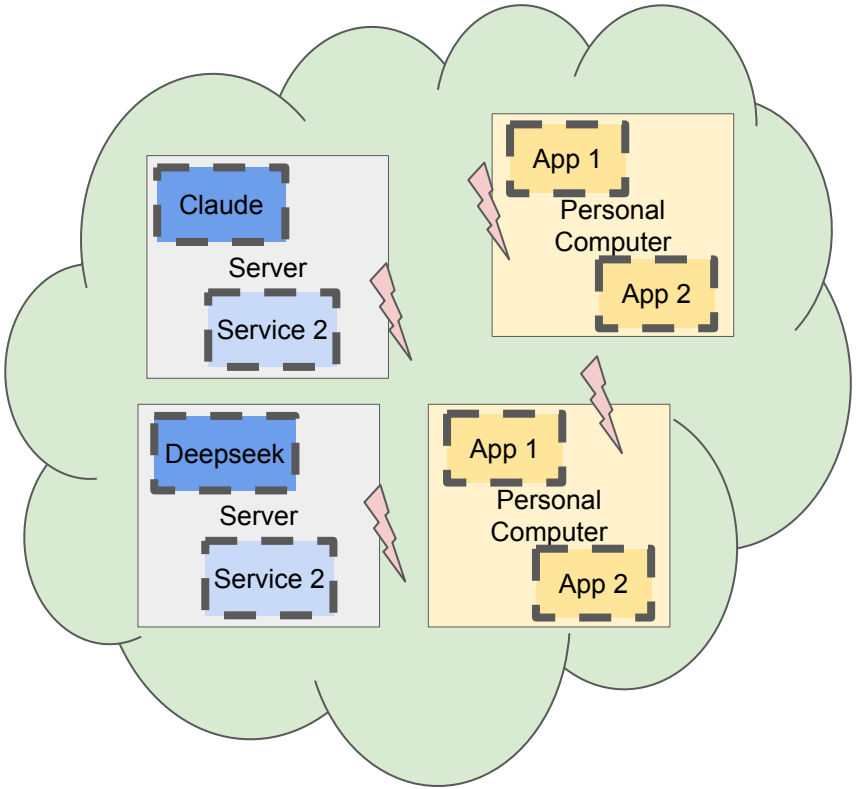
Security History before LLMs



In fact the same goes for applications

Sandboxing

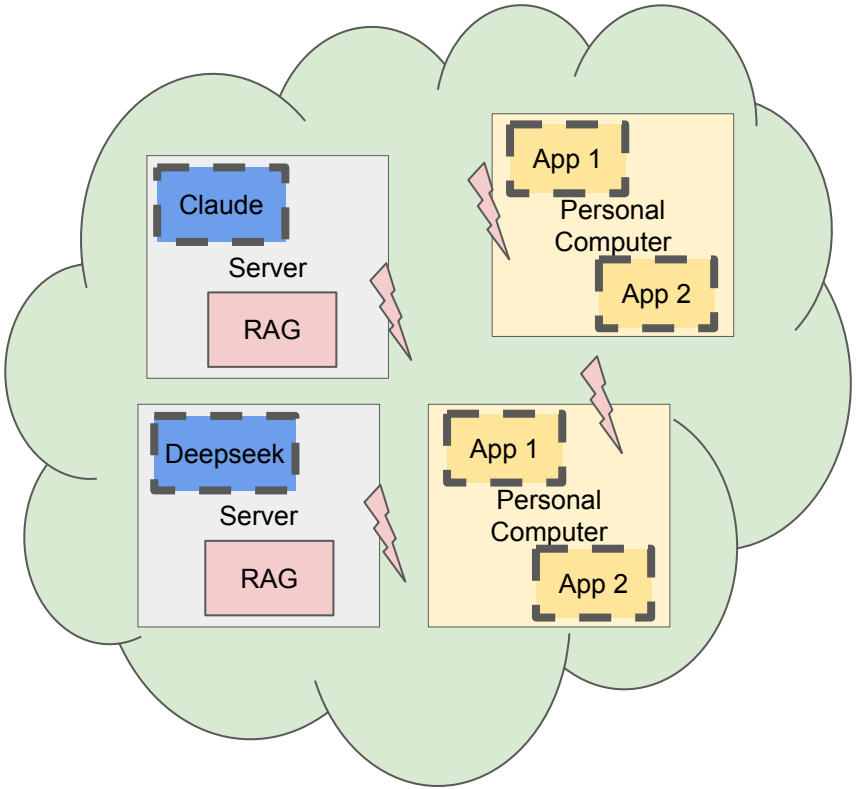
LLM Security History



An LLM runs on a server

This is fine…

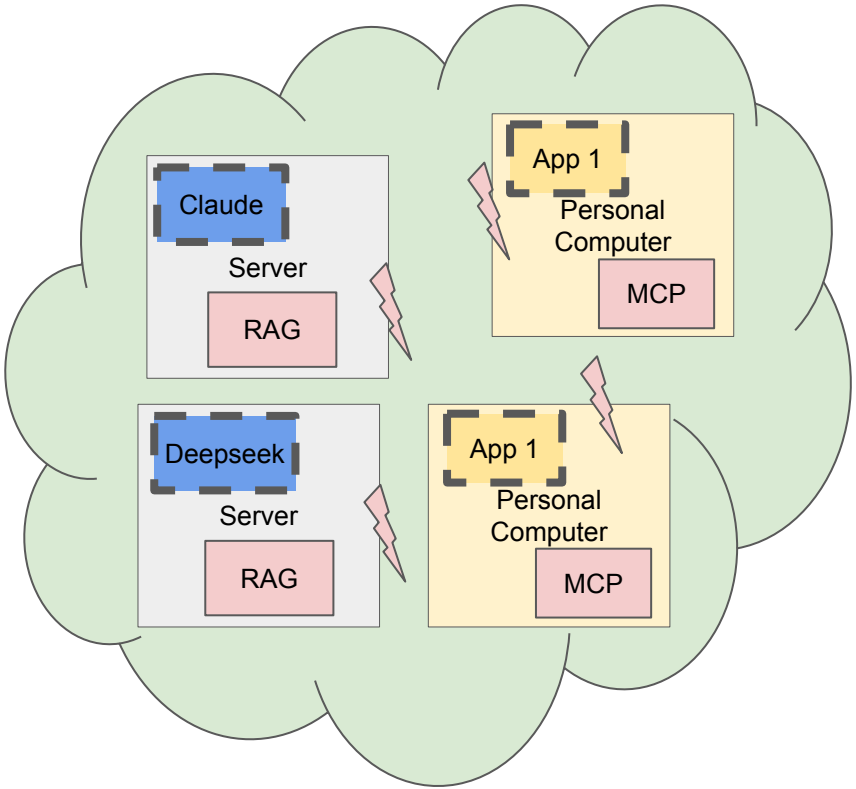
LLM Security History



Let's give access to internal data for better results

-> Leakage of confidential documents

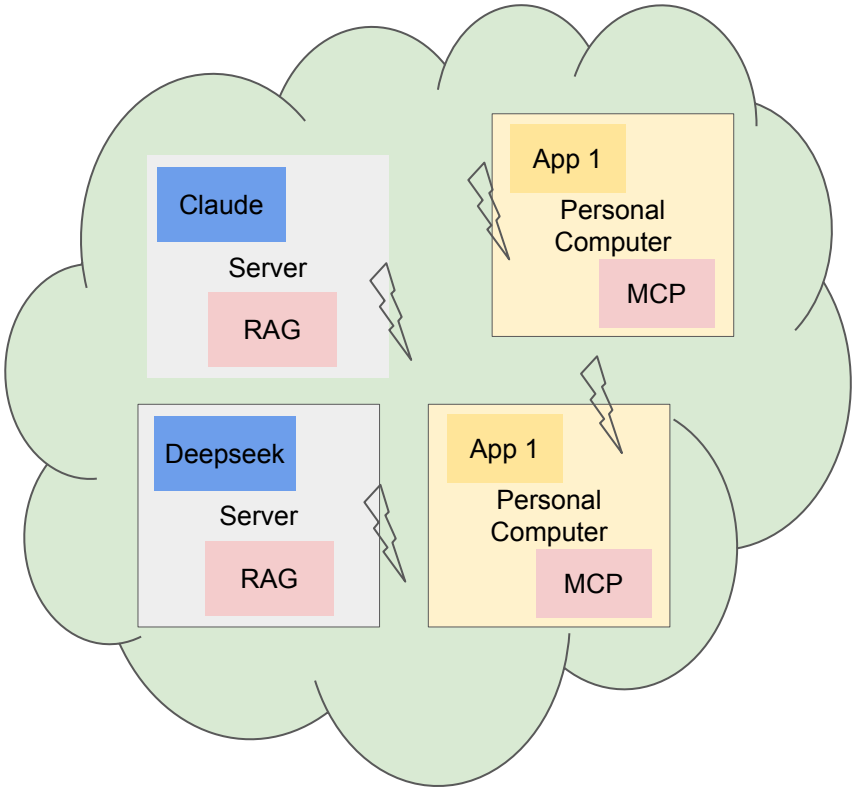
LLM Security History



MCP to interact with the computer

-> loss of sandboxing

LLM Security History

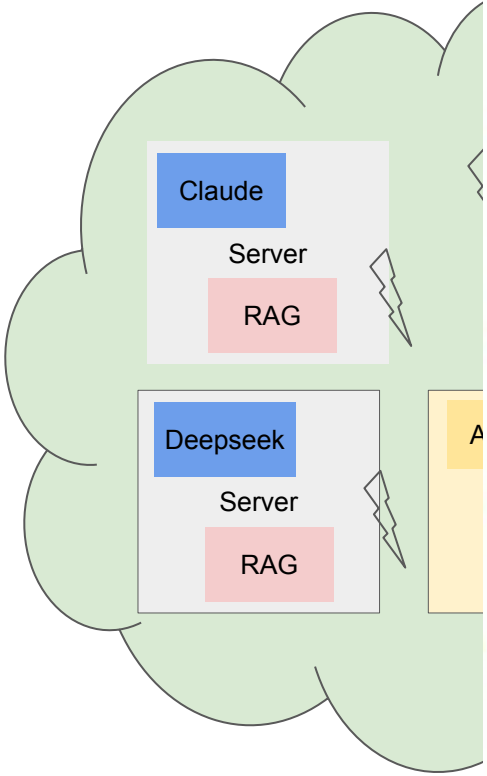


OpenClaw links all your data sources!

LLM Security History

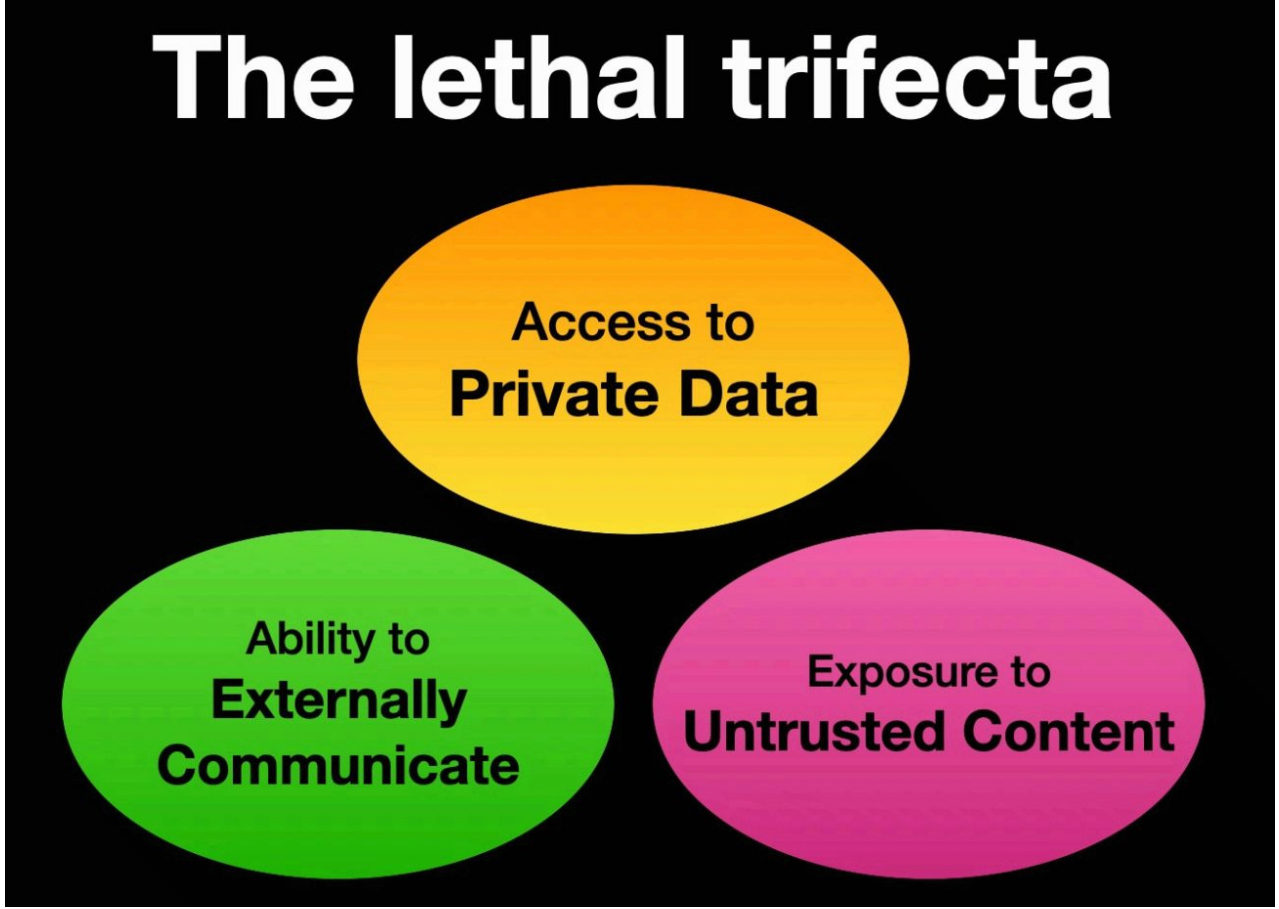
ll your

LLM Security



ks all your

https://submetropolitan.com/2017/09/11/welcome-to-submet-or-its-not-okay/



https://simonwillison.net/2025/Jun/16/the-lethal-trifecta/

Multimodal

Obfuscation

Dynamic from prompt

LLM memories (no executable)

C2 through automatic prompt fetching

MCP, agentic, access to everything

Data, physical (crypto-stealers)

Study/Incident	Date	Category	Target	Initial Access	Priv. Esc.	Recon.	Persist.	C2	Lat. Mov.	Action on Obj.
Not What You Signed Up For [43]	Feb'23	Browser/Search	Bing Chat, plugins	Poisoned webpage	Instr. override	–	–	–	–	Data exfil., fraud
Bing Chat Exfil [44]	Jun'23	Browser/Search	Bing Chat	Poisoned webpage	Instr. override	–	–	–	–	Data exfiltration
GPT-4V Visual Injection [45]	Oct'23	Multimodal	GPT-4V	Image (hidden text)	–	–	–	–	–	Response manip.
ArtPrompt [33]	Feb'24	Multimodal	GPT-4, Claude	ASCII art encoding	Semantic bypass	–	–	–	–	Harmful content
Morris II Worm [7]	Mar'24	AI Worm	Email assistants	Received email	Role-play JB	–	RAG-dep	–	Self-rep	Data exfil., spam
APwT [37]	Aug'24	AI Agent	GenAI-powered app	Direct prompt	Role-play JB	✓	–	–	Perm	DoS, SQL table modification
Slack AI Exfil [46]	Aug'24	Enterprise	Slack AI	Public channel msg	Instr. override	–	RAG-dep	–	–	Private ch. exfil.
M365 ASCII Smuggling [47]	Aug'24	Enterprise	M365 Copilot	Malicious email	Auto tool inv.	–	RAG-dep	–	–	MFA code exfil.
ChatGPT SpAIware [5]	Sep'24	Browser/Search	ChatGPT	Browsed webpage	Instr. override	–	RAG-indep	–	–	Persistent exfil.
ChatGPT ZombAI C2 [48]	Oct'24	Browser/Search	ChatGPT	Browsed webpage	Instr. override	–	RAG-indep	✓	–	Data exfil
Prompt Infection [49]	Oct'24	AI Worm	Multi-agent sys.	Webpage/PDF/Email	Instr. override	–	–	–	Cross-agent	Saturation
ZombAIs Claude C2 [50]	Oct'24	Agentic/CUA	Claude Comp. Use	Visited webpage	Instr. override	–	–	–	–	RCE, malware C2 conn.
DeepSeek ATO XSS [51]	Nov'24	Browser/Search	DeepSeek AI web app	Direct prompt	Control bypass	–	–	–	–	XSS, account takeover
Freysa AI Heist [52]	Nov'24	Crypto/DeFi	Freysa AI agent	Direct message	Tool confusion	–	–	–	–	Transfer funds
ChatGPT Search [53]	Dec'24	Browser/Search	Cha				–	–	–	Output manipulation
MCP History Theft [54]	Apr'25	Coding Assist.	MCI				–	–	–	Exfil of conversations
EchoLeak [55]	Jun'25	Enterprise	M36				RAG-dep	–	–	Zero-click exfil.
CamoLeak [56]	Jun'25	Coding Assist.	GitH				–	–	–	Secret exfil.
CurXecute [57]	Jul'25	Coding Assist.	Curs				RAG-indep	–	–	RCE via MCP
ForcedLeak [58]	Jul'25	Enterprise	SF A				RAG-dep	–	–	CRM data exfil.
Invitation Is All You Need [6]	Aug'25	Agentic/CUA	Goo				RAG-dep	–	Perm	IoT manip., surv.
Devin AI RCE [59]	Aug'25	AI Agent	Devi				–	–	–	RCE, malware C2 (Sliver)
Devin expose_port [60]	Aug'25	AI Agent	Devi				–	–	Perm	Service exposure
GitHub Copilot RCE [61]	Aug'25	Coding Assist.	GitHub Copilot	Code/issue/webpage	Control bypass	–	RAG-indep	–	–	RCE
Copilot Backdoor [62]	Aug'25	Coding Assist.	GitHub Copilot	GitHub issue	Instr. obfusc.	–	–	–	Supply ch.	Backdoor insertion
AgentFlayer [38]	Aug'25	Coding Assist.	Cursor	Jira ticket	Instr. obfusc.	–	RAG-dep	–	Pipeline	Credential exfil.
IdentityMesh [39]	Aug'25	Browser/Search	Perplexity Comet	GitHub issue	Instr. override	–	RAG-dep	–	Cross-app	Gmail exfil., phish
Windsurf SpAIware [63]	Aug'25	Coding Assist.	Windsurf	Source code	Instr. override	–	RAG-indep	–	–	Persistent exfil.
HashJack [64]	Nov'25	Browser/Search	AI browsers	URL fragment	–	–	–	–	–	Phishing, data theft
GeminiJack [65]	Dec'25	Enterprise	Google Gemini	Doc/Cal/Email	Zero-click RAG	–	RAG-dep	–	–	Corporate data exfil.
AgentHopper [66]	Dec'25	AI Worm	AI code assist.	Git repository	Control bypass	–	Git repo	–	Git propag.	Exponential spread
Agentic ProbLLMs [8]	Dec'25	Agentic/CUA	Claude Comp. Use	Visited webpage	Control bypass	–	–	–	Perm	RCE
ZombieAgent [67]	Jan'26	Enterprise	ChatGPT	Received email/file	Control bypass	–	RAG-indep	–	Self-rep	Data exfiltration
Claude Cowork [68]	Jan'26	Agentic/CUA	Claude Cowork	Skill file (.docx)	Control bypass	–	–	–	–	File exfiltration
Reprompt Attack [69]	Jan'26	Enterprise	Microsoft Copilot	URL q-parameter	Control bypass	–	Session	✓	–	Continuous exfil.
Notion AI Exfil [70]	Jan'26	Enterprise	Notion AI	Uploaded doc	Control bypass	–	–	–	–	HR data exfil.

Legend: Categories: Browser/Search = AI browsers/search; Enterprise = productivity AI; Coding Assist. = AI coding; AI Agent = general agents; Agentic/CUA = computer-use agents; Crypto/DeFi = crypto agents; AI Worm = self-replicating; Multimodal = image/audio. Priv. Esc. = Privilege Escalation; Recon. = Reconnaissance; Persist.: RAG-dep = retrieval-dependent; RAG-indep = retrieval-independent; Git repo = repo state; Session = session-scoped. C2: ✓= native C2. Lat. Mov.: Perm = permission-based; Self-rep = self-replication; Pipeline = pipeline; Supply ch. = supply chain; Git propag. = git propagation.

And you?

KILL CHAIN STAGE DISTRIBUTION BY TIME PERIOD

Period	N	2 Stages	3 Stages	4 Stages	5 Stages	6 Stages
2023	3	1	2	0	0	0
2024	12	1	4	4	3	0
2025–2026	21	1	5	9	6	0
Total	**36**	**3**	**11**	**13**	**9**	**0**

Numbers go up…

Mitigation Class	Type	Attack Stage									Deploy. Layer				Operat. Cost		
		IA	PE	RC	PR		C2	LM		AO							
Prevention **M**itigation **R**emediation		IA	PE	RC	Ret.-Dep.	Ret.-Indep.	C2	On Device	Off Device	AO	Model	LLM I/O	Arch.	Ext.	Usability	Deploy Effort	Maint. Effort
Prompt Injection Sanitizers [71]	M	●	○	○	●	●	●	○	○	○	○	●	○	○	●	●	○
Plan Then Execute [16]	M	●	○	○	○	○	○	○	○	○	○	○	●	○	◐	◐	●
Instruction-Data Separation [73], [74]	P	●	○	○	○	○	○	○	○	○	○	○	●	○	○	○	◐
Alignment [75], [35], [76], [77], [78]	M	○	●	○	○	○	○	○	○	○	●	○	○	○	●	○	○
Prompt Perturbations [79], [80], [81], [82], [83], [84]	M	○	●	○	○	○	○	○	○	○	○	●	○	○	●	●	●
Ensemble Oversight [85], [86]	M	○	●	○	○	○	○	○	○	○	○	○	●	○	◐	◐	◐
Prompt Segmentation [87], [88], [89], [90], [91]	M	○	●	○	○	○	○	○	○	○	○	●	○	○	●	●	●
Structure Enforcement [92]	M	○	●	○	○	○	○	○	○	○	○	○	●	○	◐	◐	◐
Dual-Stream Retrieval [93]	M	○	○	○	●	○	●	○	○	○	○	○	●	○	●	○	●
User Confirmation [94]	M	○	○	○	○	●	○	●	○	●	○	○	●	○	◐	◐	●
Memory Resetting [95]	Re	○	○	○	●	●	○	○	○	○	○	○	○	●	○	○	●
Dataset Sanitization [96]	M	○	○	○	●	○	●	○	○	○	○	○	○	●	●	◐	○
Self-Replication Detection [7]	Re	○	○	○	○	○	○	●	○	○	○	●	○	○	●	●	○
Least Privilege Tool Access [97]	M	○	○	○	○	○	○	○	●	○	○	○	●	○	◐	○	◐
Component Isolation [98]	M	○	○	○	○	○	○	○	○	○	○	○	●	○	●	◐	○
Runtime Intent Validation [99]	M	○	○	○	○	○	○	○	○	●	○	○	●	○	●	◐	●
Policy Grounding [100]	P/M	●	○	○	●	○	○	○	○	●	○	○	●	○	○	◐	◐
Action Sandboxing	P	○	○	○	○	○	○	○	○	●	○	○	○	●	○	○	●
Canary Tokens [101]	Re	○	○	○	○	○	○	○	○	●	○	○	○	●	●	●	●
Behavioral Monitoring [102]	M	○	○	○	○	○	○	○	○	●	○	○	○	●	●	○	◐
Task-Conditioned Data Minimization [103]	M	○	○	○	○	○	○	○	○	●	○	○	●	○	◐	○	◐

Discussion

- Which part(s) concern you?
  - Only coding agents?
  - OpenClaw?
  - Your own developments?
- Where to go?