
LLMS
PROMPTING, AGENTS, ASSISTANTS, OH MY

JOSH RICKARD

ABOUT

- Blue Team | DFIR | SOAR | Threat Detection
- Product Management | Software Engineering

- GitHub.com/msadministrator
- letsautomate.it

THE ANALOGY

- We’re in Kansas (kinda) but MIZ!
- Dorothy is us, humans in this scenario
- We all have to travel the yellow brick road of progression and learn as we go

- GPT-3 - 2020
- CHATGPT - 2022
- GPT-4 - 2023
- CLAUDE, GEMINI, LLAMA - 2024
- DEEPSEEK R1 - 2025
- OPENCLAW - 2026

TORNADO

PLAYERS

- OpenAI / ChatGPT
- Anthropic / Claude
- Google / Gemini
- Meta / LLaMA
- xAI / Grok
- OpenClaw & Cursor
- The AI startups

WHAT THE F*** IS GOING ON

- LLM - Large Language Model
- Prompt - Your Instruction to the model - Garbage In -> Garbage Out
- Context Window - How much text the LLM can “see” at once
- RAG - Retrieval Augmented Generation - Inject data into the context of an LLM
- Agent - An LLM that calls tools, takes actions, loops until done.
- MCP - Model Context Protocol - Standardized API calls so models can access them
- Vibe Coding - Prompting your way to code without understanding it

INVESTORS & EVANGELIST

- Magical end to this road
- Some see progress
  - No more working! Enjoy life!
  - Intergalactic space travel
- Some see despair
  - Being replaced
  - Overlords

OUR CONSCIENCE (OR JUST ME)

- Environmental impact
- Mental health impact
- Scraping & hosting costs
- The other costs

AND GET WALKING

WAKE UP

SCARECROW

- Simple prompting
- Vibe coding
- Small to Medium complexity

CREATE A PYTHON SCRIPT TO PARSE THE MITRE ATT&CK JSON FILE

PROMPTING

- Use natural language to describe your goal, question, problem, etc.
- Provide context (resources, constraints, examples, output format, etc)
- Keep track and iterate on prompts for the best results (use other models as well)
- Use speech-to-text if that is easier (I prefer writing)

MID PROMPT

- As a security analyst and phishing threat detection engineer.
- As a software engineer experienced with Python, Flask and Docker.
- Create a simple Flask (bootstrap) application to collect DNS, WHOIS/RDAP, HTML and other open-source threat intelligence for a given phishing url.

MORE ON PROMPTING

- Vibe Coding
  - No true thought on design / architecture
  - Great for prototyping / ideation
- More advanced prompting
  - Add requirements, constraints, tools, what not to do, etc.

PERSONAS -> IDENTITIES -> PERSONALITIES

- Set the stage by using your existing experience
  - As a senior threat detection engineer (stack them)
  - Specify the technologies, frameworks, principles, etc. to focus on
  - Specify the repository structure
  - Provide perspectives & goals (use Critical Systems Thinking)
- When done, tell the LLM to output this personality and lessons learned for future projects

TOOLS

- Specify the tools that the LLM has access to
  - Local docker setup
  - MCP(s)
- Provide boundaries / guidelines for tool use
  - Use **uv** over *poetry*
  - Only invoke the **enrichment_function** when given an IPv4 address
  - Only perform hunt after all data extracted from sample

DOCUMENTATION (RESOURCES)

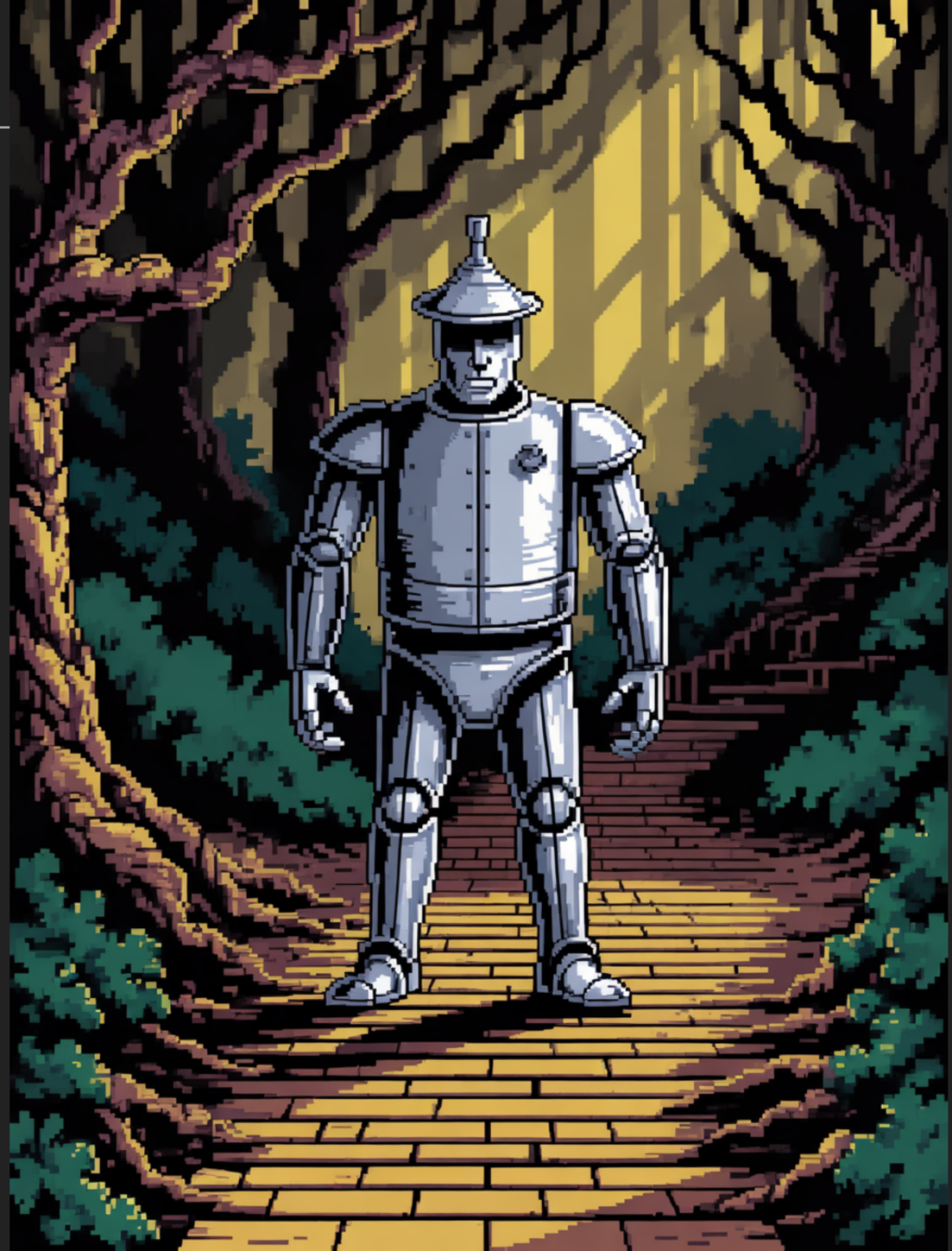
- Refer to API specifications / JSON Schemas / repositories
- Provide a list of terms that are exclusively used internally (jargon)
- Provide examples of past / previous problems
- Specify what **NOT** to do
  - Document when not to use a tool
  - Define what to do if a situation occurs

SECURITY RISKS

- Prompt injection
- Data leakage
- Hallucination / Misinformation
- Abuse / automation for offensive use cases
- Authorization abuse / mis-use

TIN MAN

- APIs (REST)
- MCP (Model Context Protocol)
- Tools (plugins)
- Skills (SKILL.md) / RAG / Memory
- Assistants / Copilots



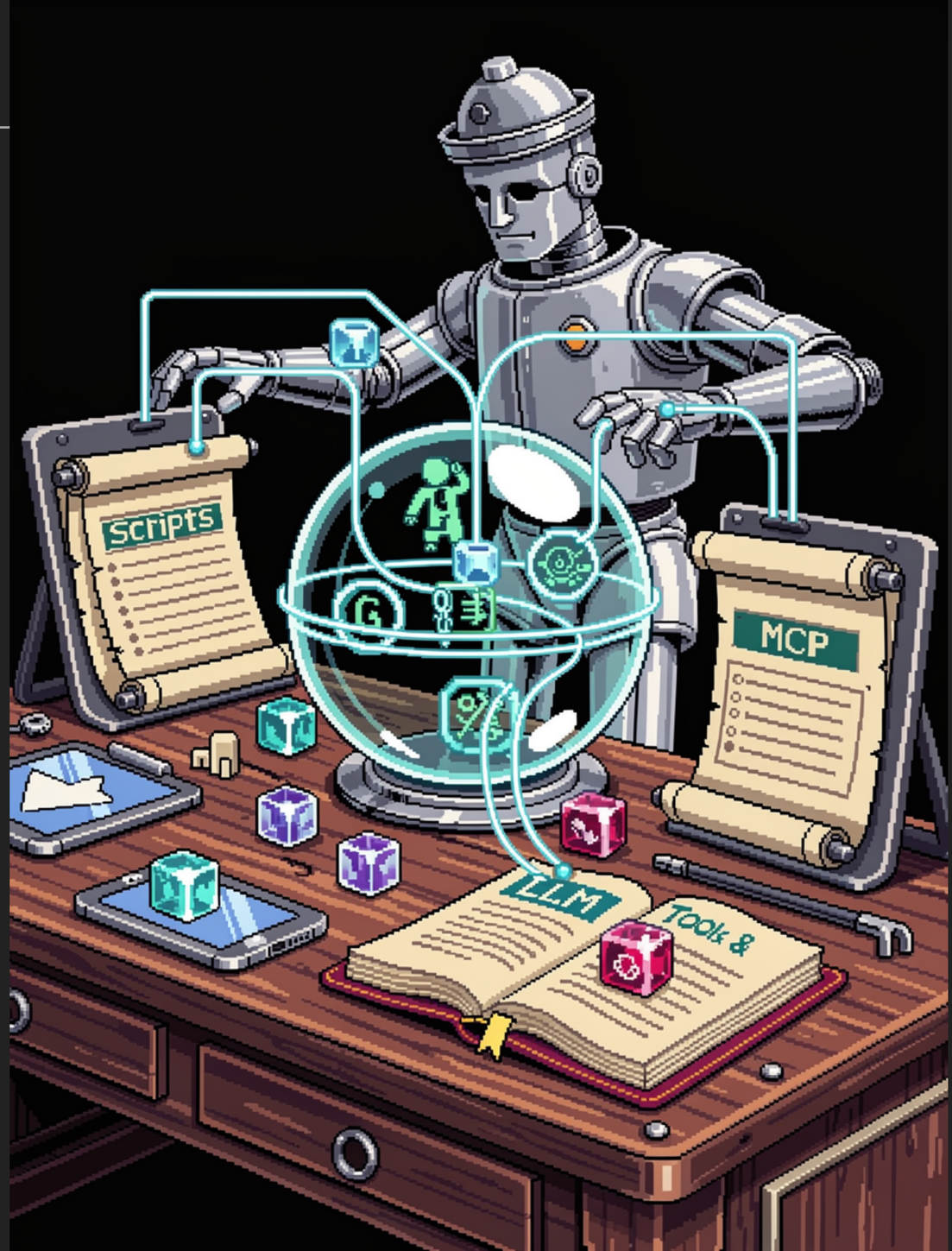
MODEL CONTEXT PROTOCOL

- Standardized way for LLMs to communicate to APIs (external services)
- You can abstract business logic and standardize response structure that an LLM can understand (words)
- A standard framework to make development simple
- I still think the industry should have standardized on gRPC & Protobuf (HTTP2/3) (and other security reasons)

modelcontextprotocol.io

TOOLS

- Define **tools** using different sources like scripts, MCPs, APIs themselves, local files, and more.
- Define how and when to use these **tools**
- Provide context for understanding the results of any resource calls
- Results flow back into context

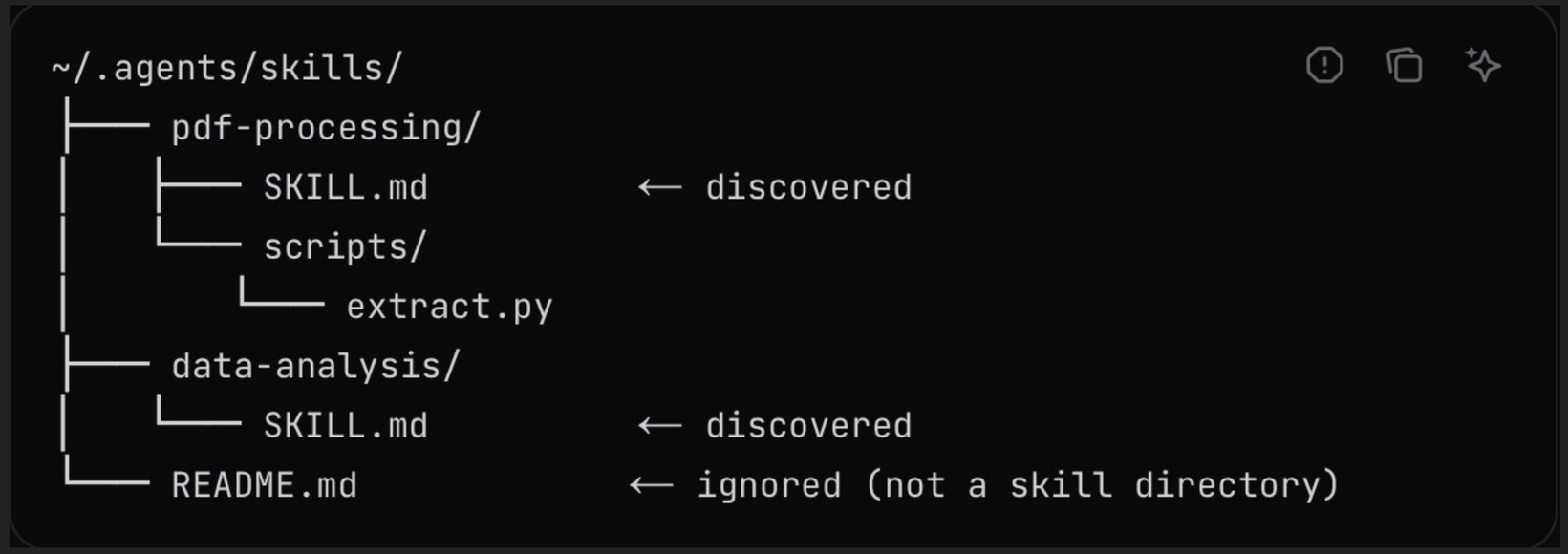


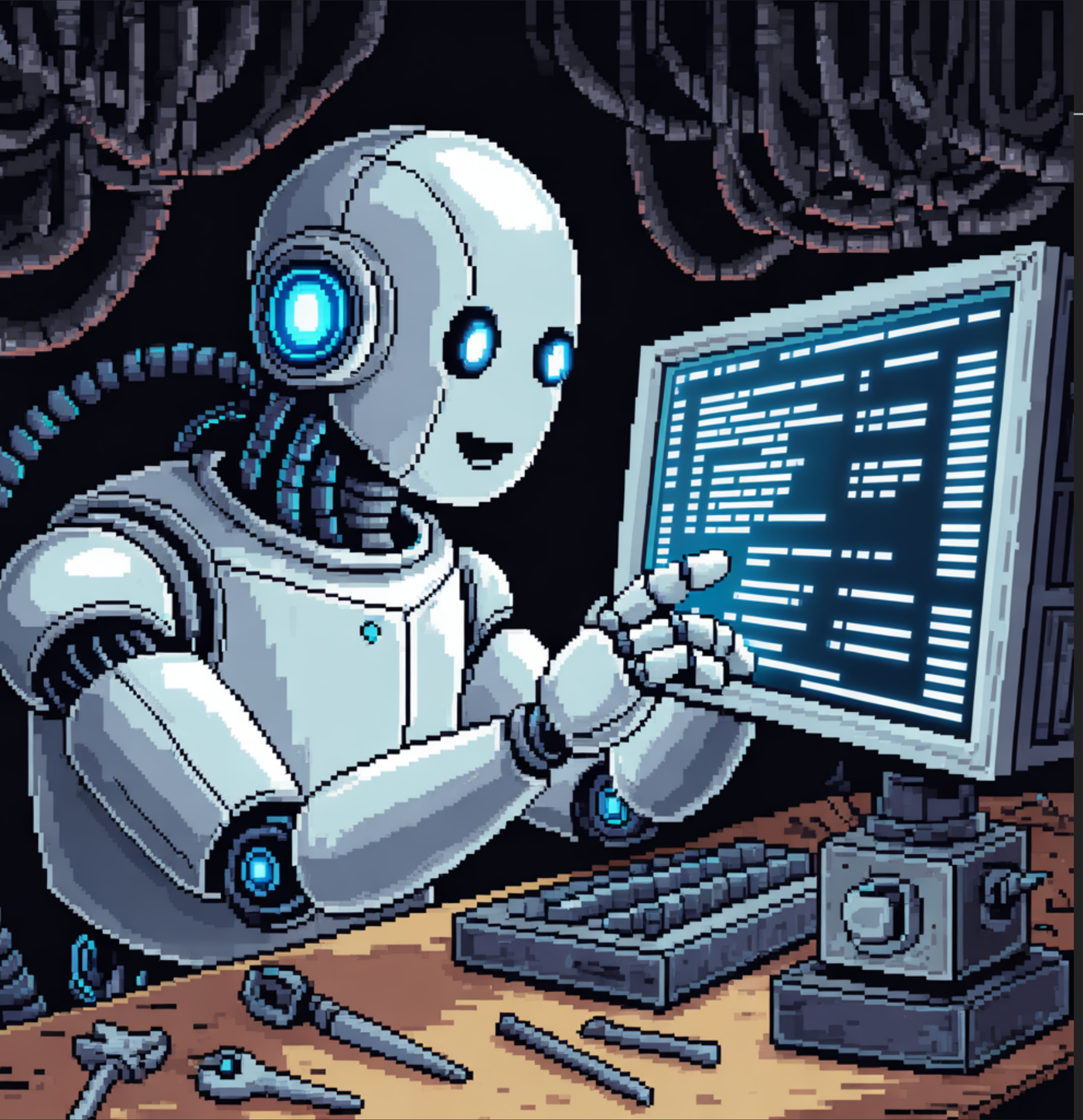
SKILLS

- A folder of instructions, scripts, and resources that agents use to do things more accurately and efficiently
- A SKILL.md (markdown file) provides the LLM who, what, when, why and how, to understand the data itself

SKILL.MD

- A markdown file which references
  - Other skills
  - Documentation
  - Scripts
  - Etc.





~~ASSISTANTS~~ / COPILOTS

- GitHub Copilot / PAI / OpenClaw / Cursor
- Agentic development environments
- State-machines used for orchestration of work via a workflow (hooks, LangChain, etc.)

LION

- ChatGPT / Claude / Gemini
  - Newer reasoning models
  - Adds UI, persistent memory, tools, etc.
- Multi-Agent Frameworks / Orchestration
  - Finite State Machines
  - Workflow architecture

CURSOR & CLAUDE & DEEPSEEK R1

- Reasoning models
- Reinforcement Learning-based Reasoning (rewards)
  - Outcome-based
  - Chain-of-Thought
  - Process-based
  - Structured reasoning
  - Program-of-Thought
  - Tool-augmented reasoning

AUTONOMOUS AGENTS (ASSISTANTS)

- Advanced & custom workflows (traditional state machines)
- Defined & continually updated preferences based on interactions with an LLM
- Keep & utilize a memory system
- OpenClaw https://github.com/openclaw/openclaw
- PAI - https://github.com/danielmiessler/Personal_AI_Infrastructure

YOUR SECURITY KNOWLEDGE

- Create a CLAUDE.md / AGENTS.md
- List out your:
  - Roles (or ask LLM to do that based on blog, LinkedIn, etc.)
  - What tools/skills you are strong in or favor
  - Goals (career or otherwise)
  - Tools available to you (MCP, EDR, SOAR, SaaS, etc.)
- Refer to your Incident Response Process Documentation (right!?)

NO ONE REALLY F***ING KNOWS

EMERALD CITY

FUTURE & BEYOND

- Agent-to-Agent Protocol
  - Negotiations, e-commerce, possibly your interface to the new internet
- Micro data-centers (hosted Agents down the street)
  - More energy efficient GPUs
  - Cheaper to run models with low latency
- AGI is still a long ways away
  - Unless we can get a model to build AGI for us :)

THE WIZARD


YOU


END

REFERENCES

- How I use LLMs for Security Work: Part 1 (https://substack.com/@l0phius/p-185980735)
- How I use LLMs for Security Work: Part 2 (https://substack.com/@l0phius/p-189795560)
- Agentic Web: A New Internet Built for Agents, Not Browsers (https://letsautomate.it/article/agentic-web/)