



Simplifying Security with Local AI Agents

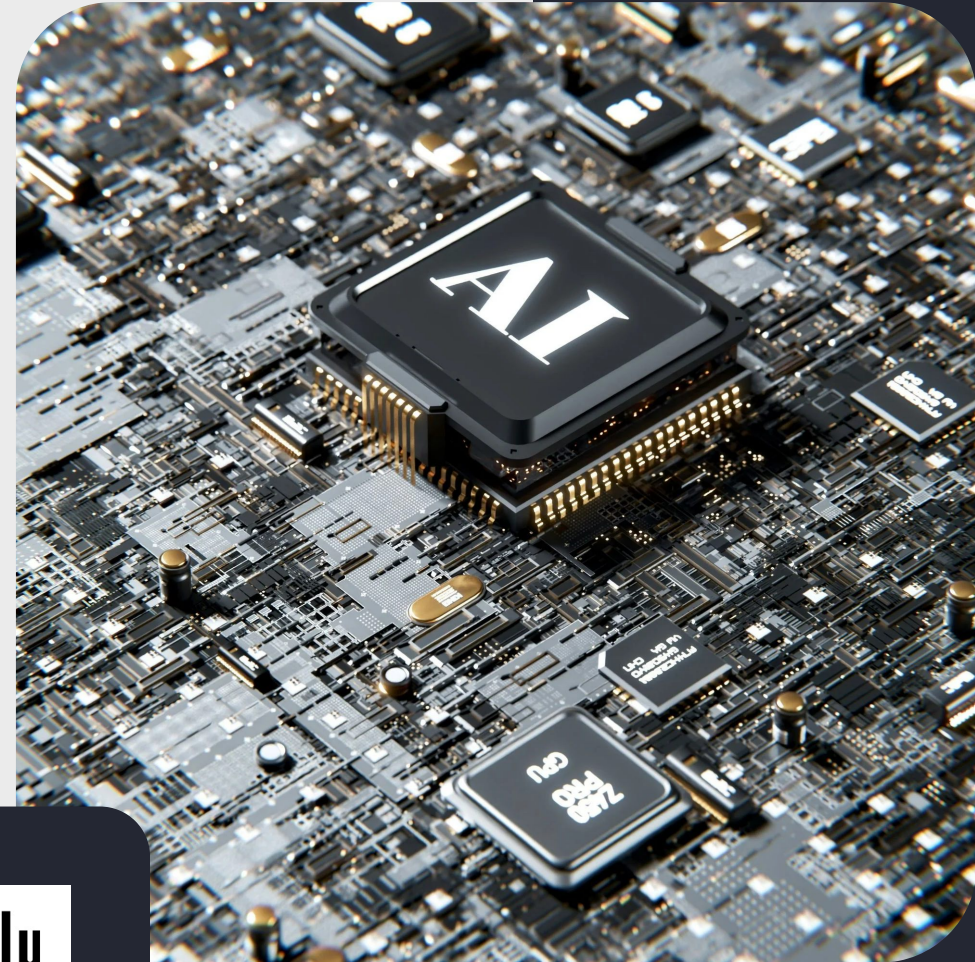
By Abhinandan Khurana



Agenda

TL;DR

- Help you understand and streamline your own research
- When, Why and How?
- Problems & possible Solutions



n/u

Disclaimer

What not to expect?

- Code Discussion
- "Perfect", no flaw PoC Demo
- this session is enough!





Audience POV?

Problems?

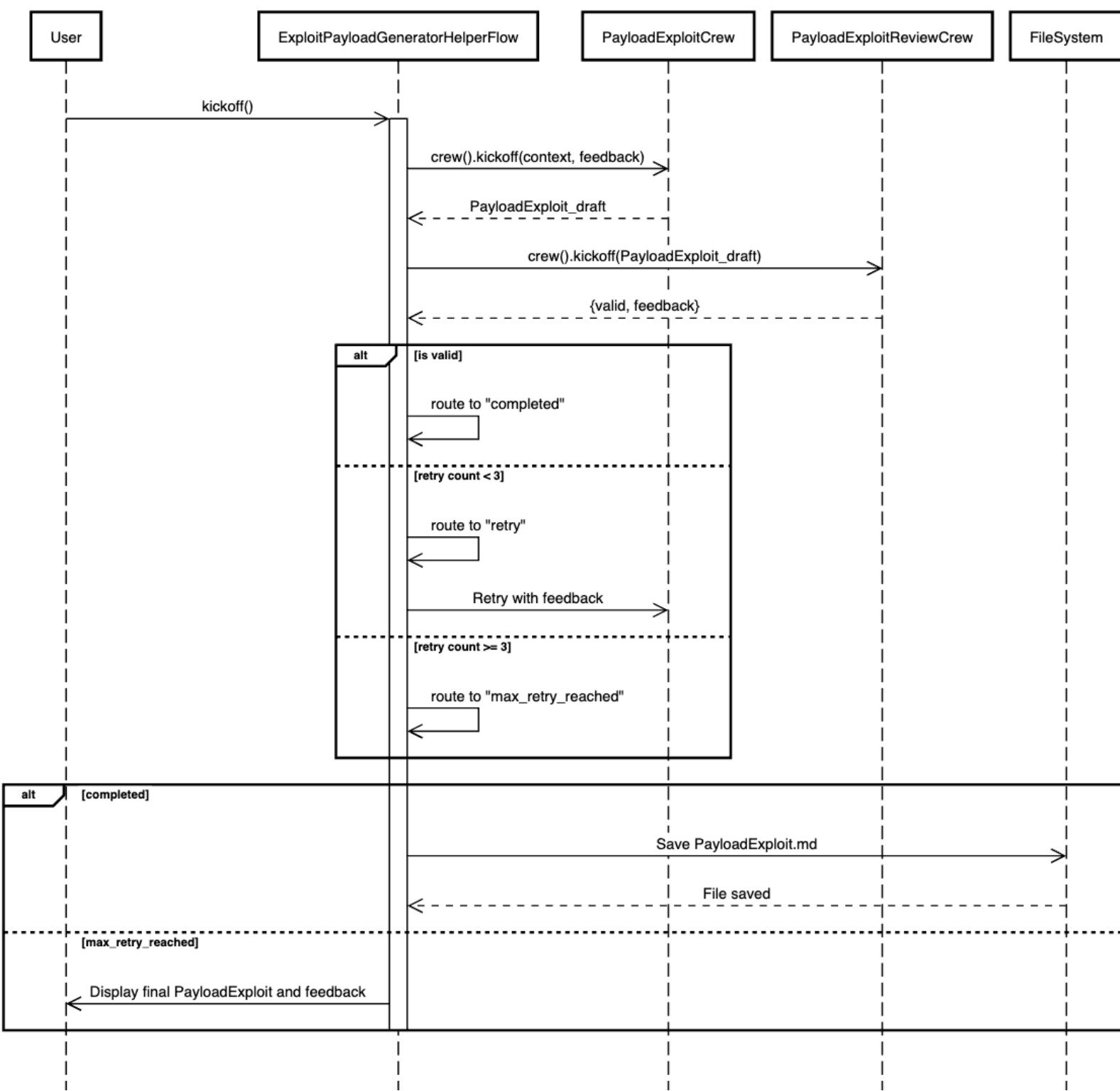
Challenges?

Ideas?

Demo 1

Exploit Payload
Generator Agent





How does it work?



Introduction to AI Agents



1

Prompt Chaining

2

Routing

3

Parallelization

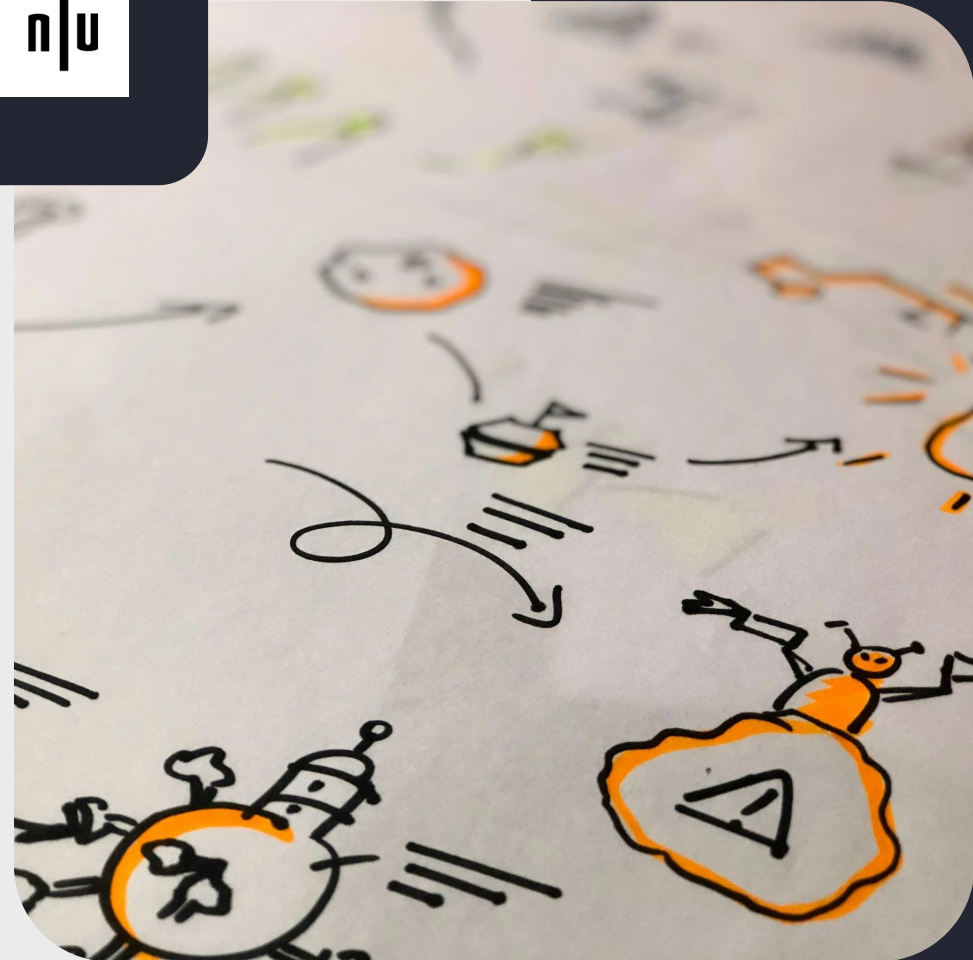
4

Orchestrator-Workers

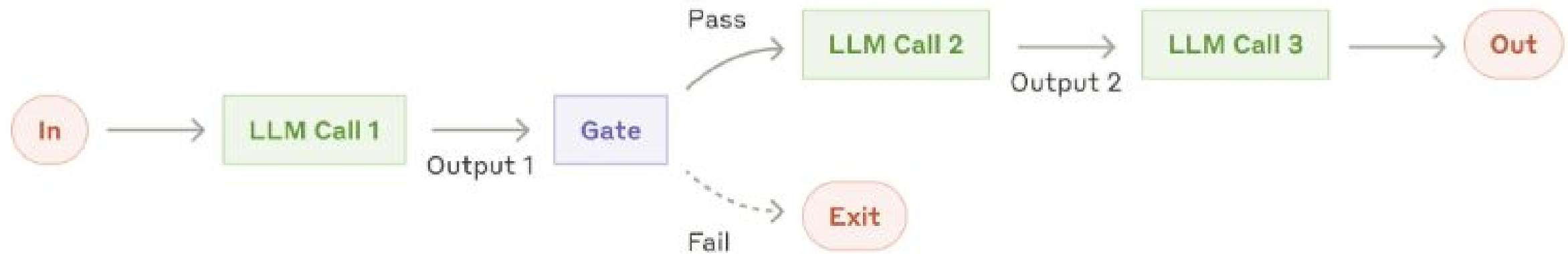
5

Evaluator-Optimizer

n/u



Prompt Chaining

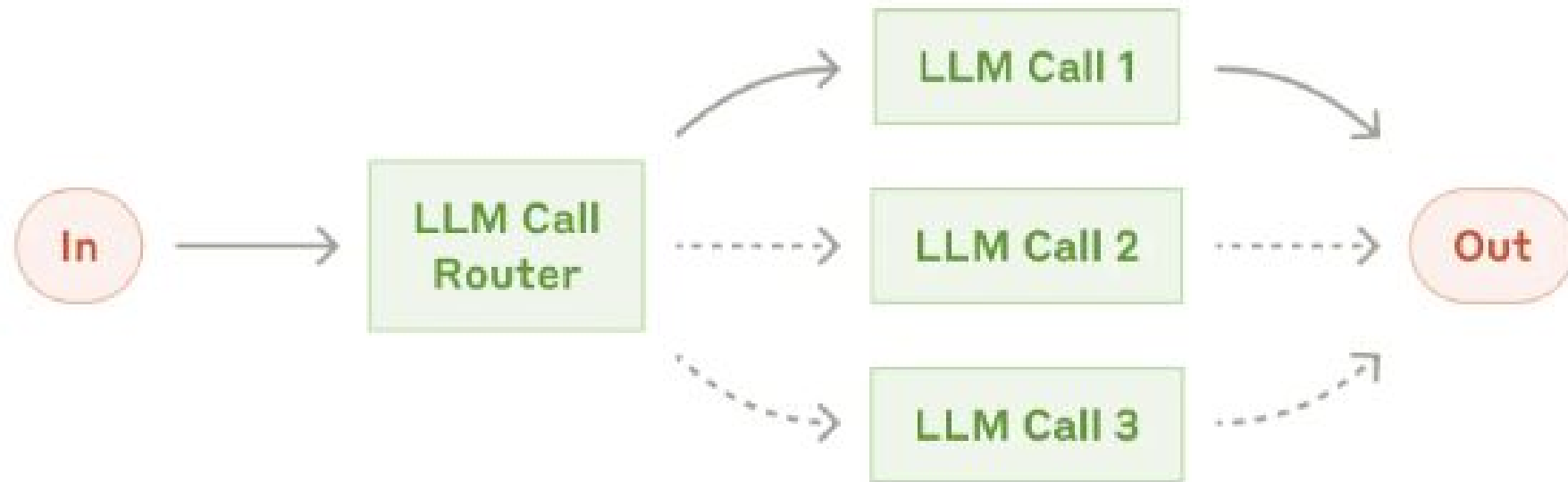


Workflow: Chaining

accuracy over
latency

Example: document
optimization

Routing



Workflow: Routing

complex to specific

Example: divide to
efficient llm tasks

Parallelization



Workflow: parallel

Sectioning & Voting

Example: adversary
network intrusion

Orchestrator-Workers

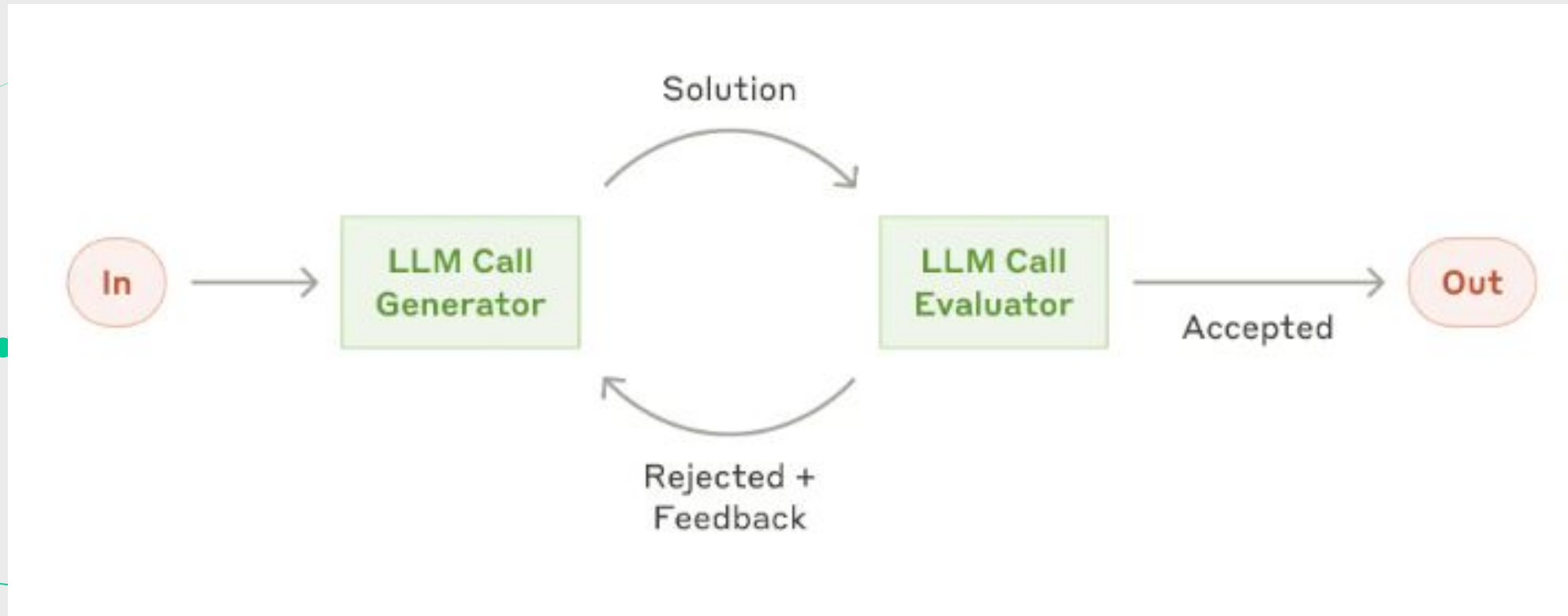


Workflow: master-slave

dynamic process

Example:
investigate and
patch SCA issues

Evaluator-Optimizer



FeedBack Loop

iteration

Example: Demo 1

What local LLMs to use for cyber security usecase?



Hermes3



whiterabbitneo



deepseek



dolphin mixtral



qwen2.5



additional

What open-source LLMs or OLMs are you in search of? #2000 in total.

Search

Best Uncensored LLMs. Top-Ranked Uncensored Large Language Models. *

Was this list helpful?

Model Size

OK

999B

Model VRAM

0

768

Columns

Quick Filters

Reset Filters & Sorting

Model Name	Maintainer	Size	Score	VRAM (GB)	Quantized	License
Oxy 1 Small	oxyapi	14B	0.43	29.7	•	apache-2.0
EVA Qwen2.5 72B V0.2	EVA-UNIT-01	72B	0.41	146	•	other
Dobby Mini Unhinged Llama 3.1 8B	SentientAGI	8B	0.41	16.1	•	llama3.1
Hermes 3 Llama 3.2 3B	NousResearch	3B	0.39	6.5	•	llama3
Hermes 3 Llama 3.1 8B	NousResearch	8B	0.39	16.1	•	llama3
Violet Twilight V0.2	Epiculous	12B	0.38	24.5	•	apache-2.0
Wayfarer 12B	LatitudeGames	12B	0.38	24.5	•	apache-2.0
Dobby Mini Leashed Llama 3.1 8B	SentientAGI	8B	0.38	16.1	•	llama3.1
Hermes 3 Llama 3.1 70B	NousResearch	70B	0.37	141.9	•	llama3

What tech stack to use for building agents?



crewAI



Langchain



llamaindex



smolagents



langflow (no
code)

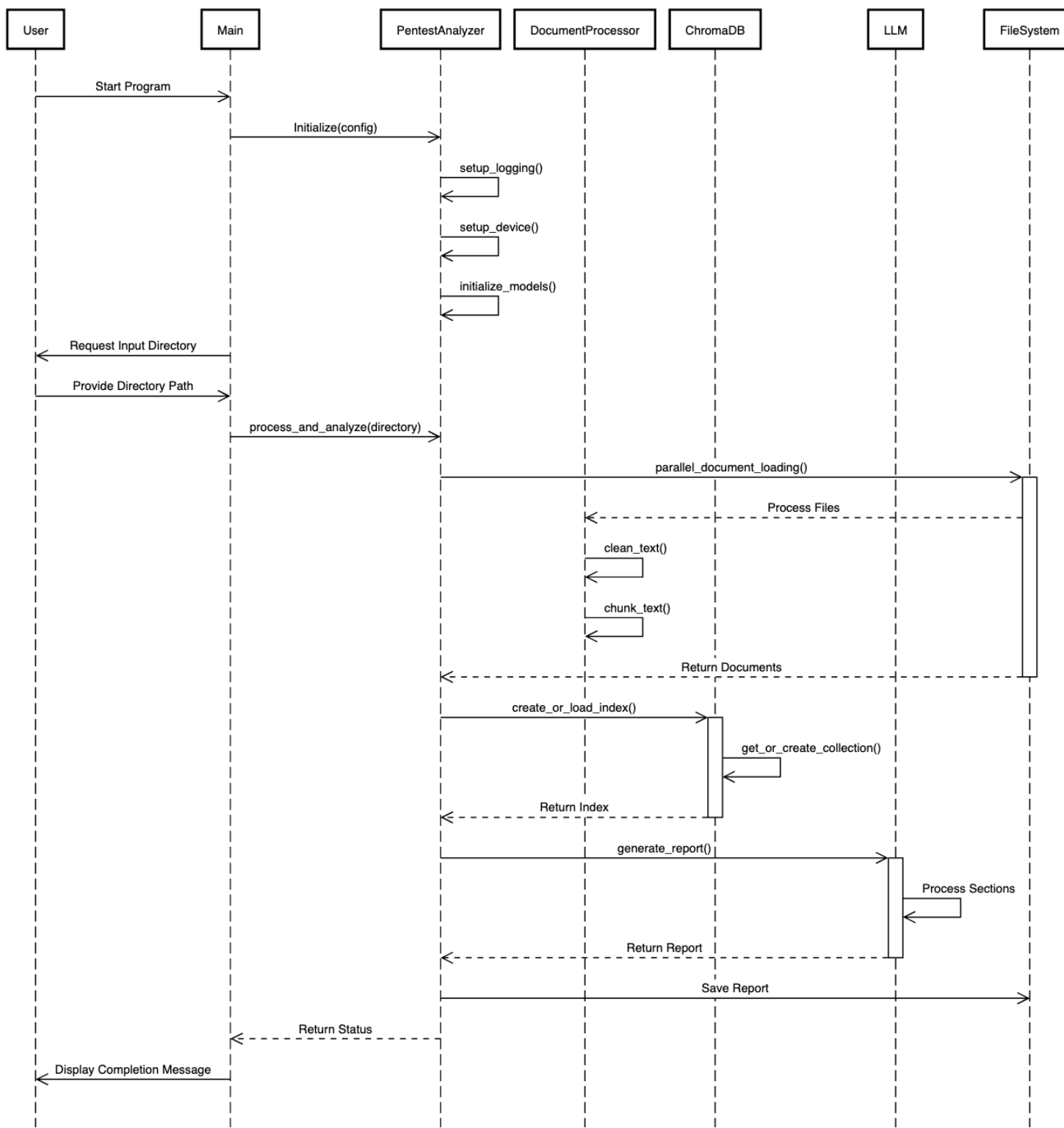


n8n (no code)

Demo 2

RAG based Pentest
report Generator
Agent





How does it work?



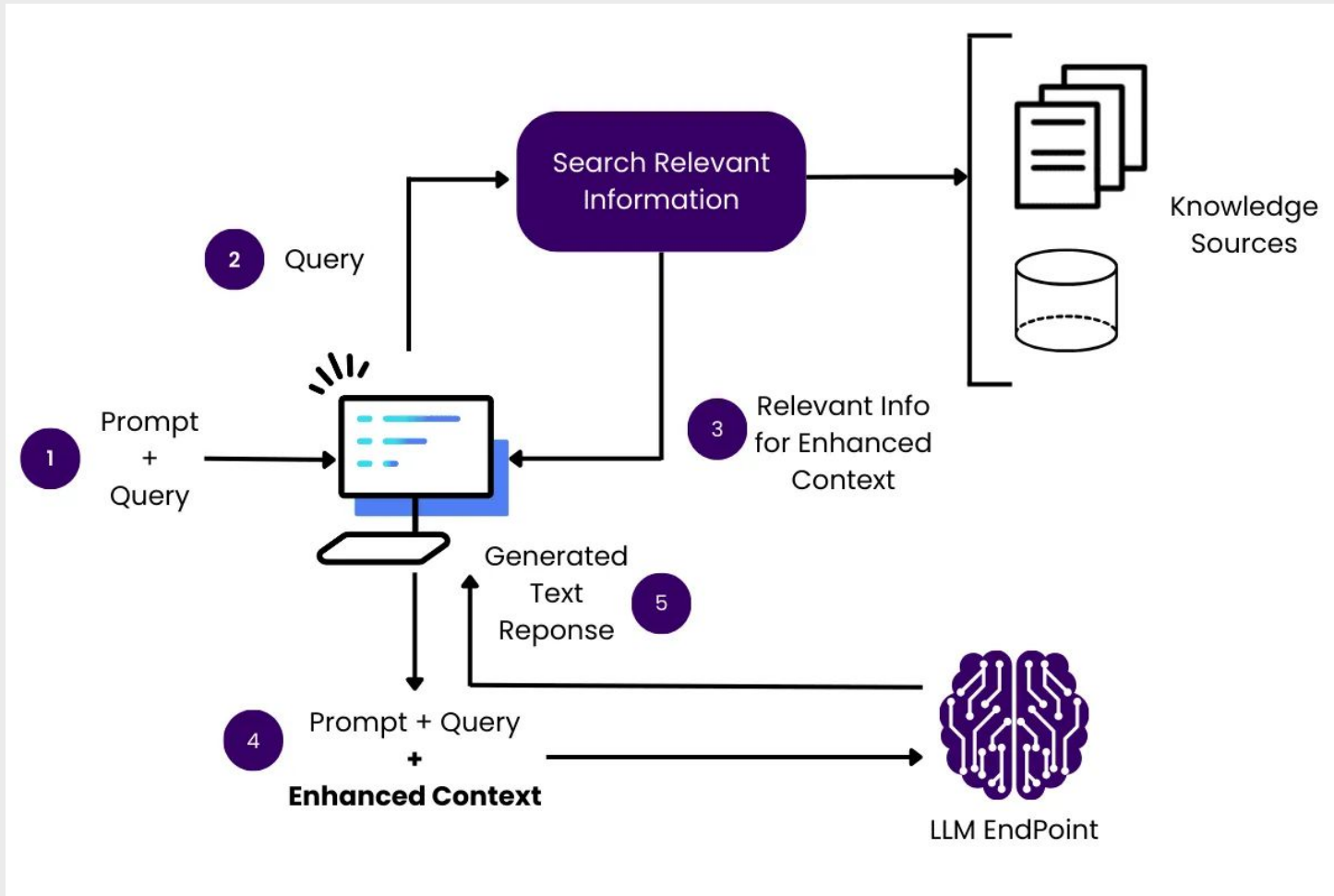
The background of the slide features a dark, semi-transparent image of three scientists in white lab coats working in a laboratory. They are gathered around a piece of equipment, possibly a microscope or a computer monitor. Overlaid on this image are several teal-colored geometric graphics, including lines connecting dots to form a network or graph structure. These graphics are positioned in the corners and along the sides of the slide.

What is Retrieval Augmented Generation(RAG)?



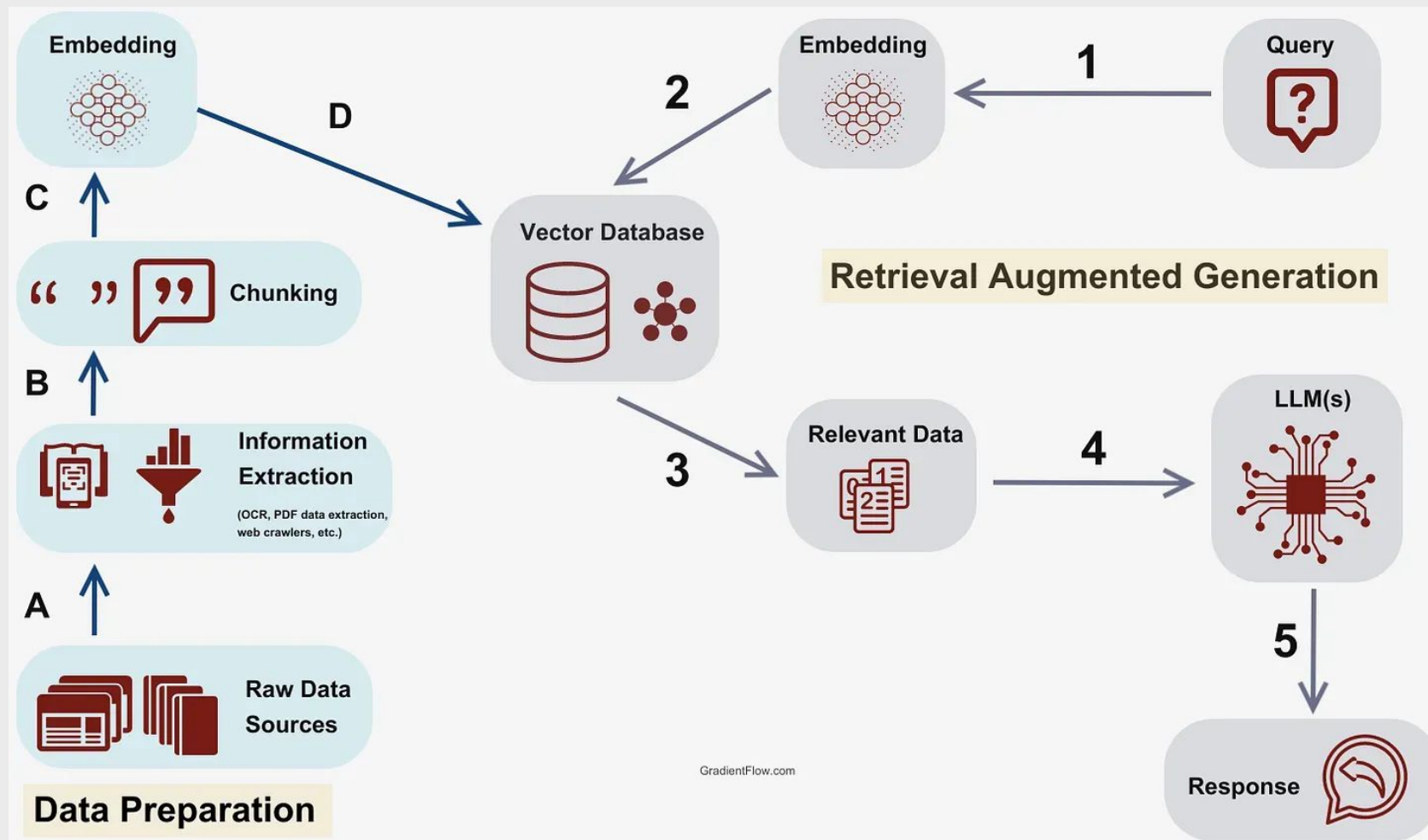
Retrieval Augmented Generation

What is RAG?



How does it work?

- Text chunking
- Embedding
- Creating a vector Database



The background features a dark, blurred image of a laboratory flask containing a white substance. Overlaid on this is a network of teal lines and dots. The dots are located at various points: one near the top right, one near the top center, one near the bottom left, one near the bottom center, and one near the bottom right. The lines connect these dots in a web-like pattern.

Any Questions?





Thank you!

GitHub: <https://github.com/Abhinandan-Khurana>

Topmate: https://topmate.io/abhinandan_khurana