SecurityBoat Community

Connect, Learn & Grow!!

LLMs from Scratch

For Security Engineers

SecurityBoat Community
Connect, Learn & Grow!!

About Me

Cloud Security Researcher & Trainer

Securing Cloud & Cloud Native Infrastructure

Building **https://CloudSecurity.Club**

Blogs at **https://badshah.io**

SecurityBoat Community
Connect, Learn & Grow!!

Agenda

Cut the AI hype

Use LLMs if it’s the solution to your usecase

SecurityBoat Community
Connect, Learn & Grow!!

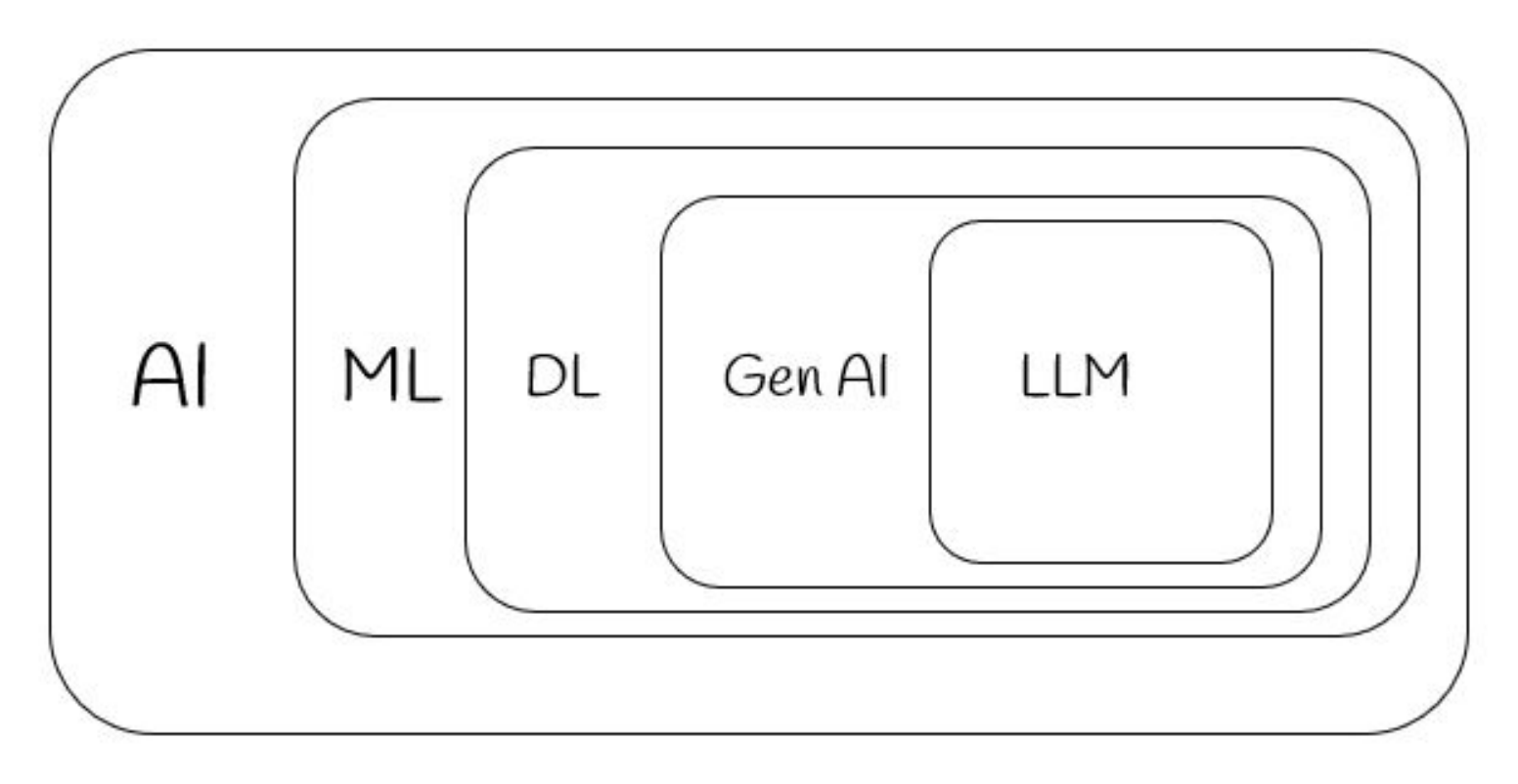
What's Artificial Intelligence?

Top Down Approach - Simulating human intelligence by machines

Bottom Up Approach - 🤷

SecurityBoat Community
Connect, Learn & Grow!!

Large Language Models (LLMs)



Source: https://www.storagefreak.net/2023/08/generative-ai-vs-large-language-models

SecurityBoat Community
Connect, Learn & Grow!!

Pro Tip #1

Ask vendors what do they mean by “AI-Powered”

SecurityBoat Community
Connect, Learn & Grow!!

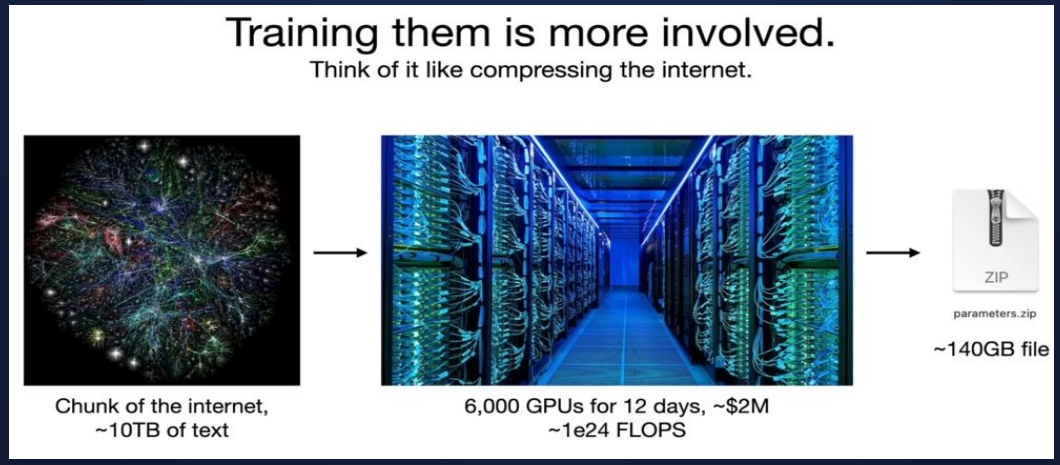
Introduction to Large Language Models

Large Language Models (LLMs)

Pattern prediction systems

General purpose

Trained on large amounts of text (books, blogs, Reddit, etc)



Source: https://www.youtube.com/watch?v=zjkBMFhNj_g

Important Components in LLMs

parameters (model training) - numerical values that encode complex patterns

inference engine (model inference) - uses parameters to predict the next word

LLM AI Model	Parameters	Year
BERT	340 million	2018
GPT-2	1.5 billion	2019
Meena	2.6 billion	2020
GPT-3	175 billion	2020
LaMDA	137 billion	2022
BLOOM	176 billion	2022

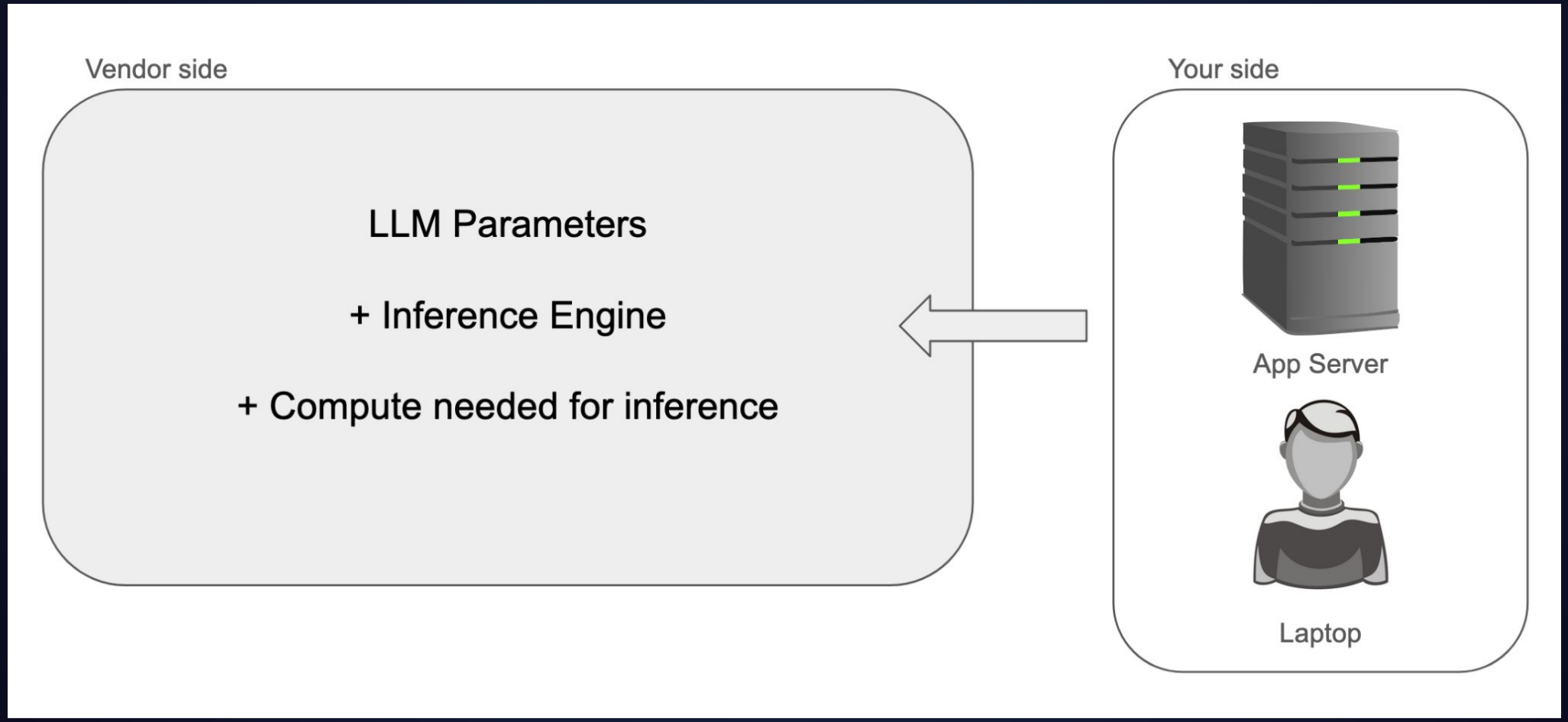
Source: https://cdn.labellerr.com/Comparing%20LLMs/comp-llm-1.webp

SecurityBoat Community
Connect, Learn & Grow!!

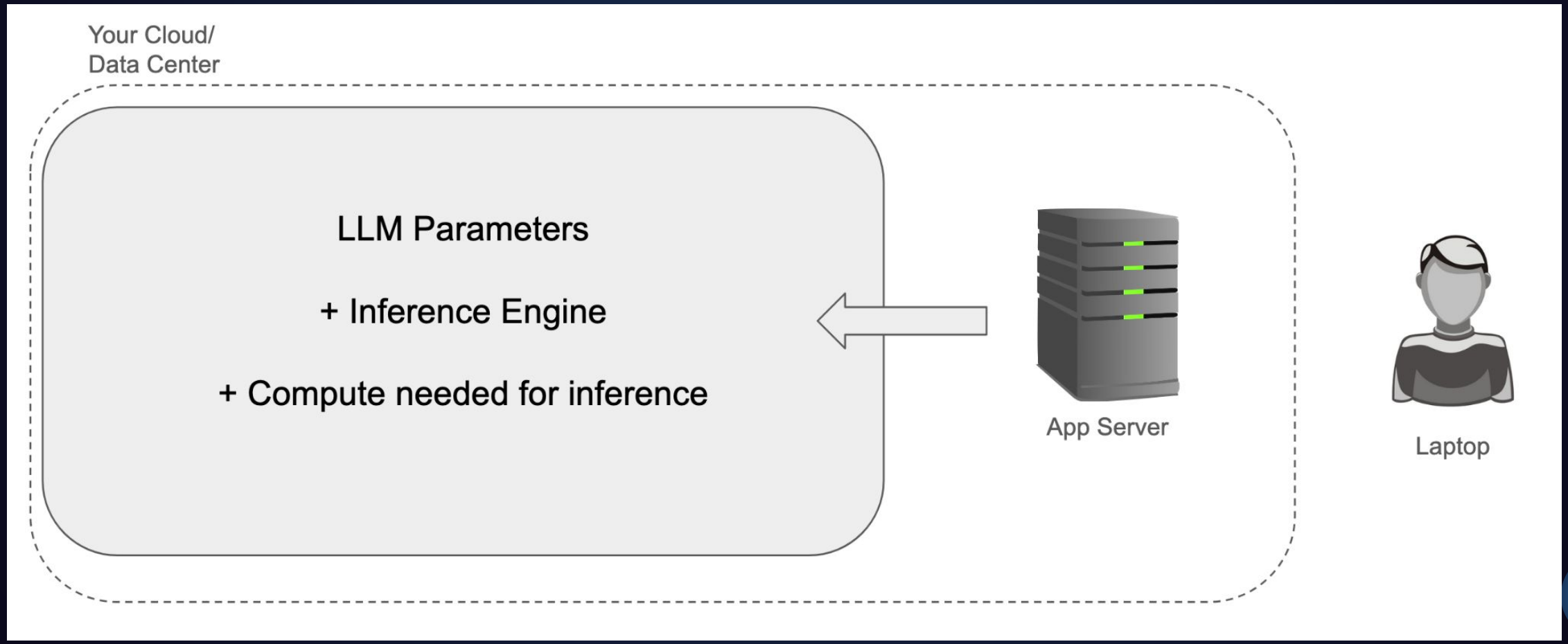
Demo #1

LLM Output Predictions

Proprietary LLMs (OpenAI, Claude, etc)



Open Weight / Open Source LLMs



SecurityBoat Community
Connect, Learn & Grow!!

Which LLM to use?

Depends on what the model is trained on (language, opinions, fake data, etc)

Bigger parameters (70B, 90B) means:

- not necessarily best one for your task
- more capable of handling variety of tasks
- more compute for output generation

Llama 3.1	July 23, 2024	• 8B • 70.6B • 405B
Llama 3.2	September 25, 2024	• 1B • 3B • 11B • 90B[40][41]
Llama 3.3	December 7, 2024	• 70B

Source: https://en.wikipedia.org/wiki/Llama_(language_model)

SecurityBoat Community
Connect, Learn & Grow!!

Pro Tip #2

Find an LLM that works best for your use case

Check the Model Cards

Instruction tuned models

Category	Benchmark	# Shots	Metric	Llama 3 8B Instruct	Llama 3.1 8B Instruct	Llama 3 70B Instruct	Llama 3.1 70B Instruct	Llama 3.1 405B Instruct
General	MMLU	5	macro_avg/acc	68.5	69.4	82.0	83.6	87.3
	MMLU (CoT)	0	macro_avg/acc	65.3	73.0	80.9	86.0	88.6
	MMLU-Pro (CoT)	5	macro_avg/acc	45.5	48.3	63.4	66.4	73.3
	IFEval			76.8	80.4	82.9	87.5	88.6
Reasoning	ARC-C	0	acc	82.4	83.4	94.4	94.8	96.9
	GPQA	0	em	34.6	30.4	39.5	46.7	50.7
Code	HumanEval	0	pass@1	60.4	72.6	81.7	80.5	89.0
	MBPP ++ base version	0	pass@1	70.6	72.8	82.5	86.0	88.6
	Multipl-E HumanEval	0	pass@1	-	50.8	-	65.5	75.2
	Multipl-E MBPP	0	pass@1	-	52.4	-	62.0	65.7

Source: https://github.com/meta-llama/llama-models/blob/main/models/llama3_1/MODEL_CARD.md

SecurityBoat Community
Connect, Learn & Grow!!

Getting ~~*best*~~ better output from Large Language Models

SecurityBoat Community
Connect, Learn & Grow!!

Pro Tip #3

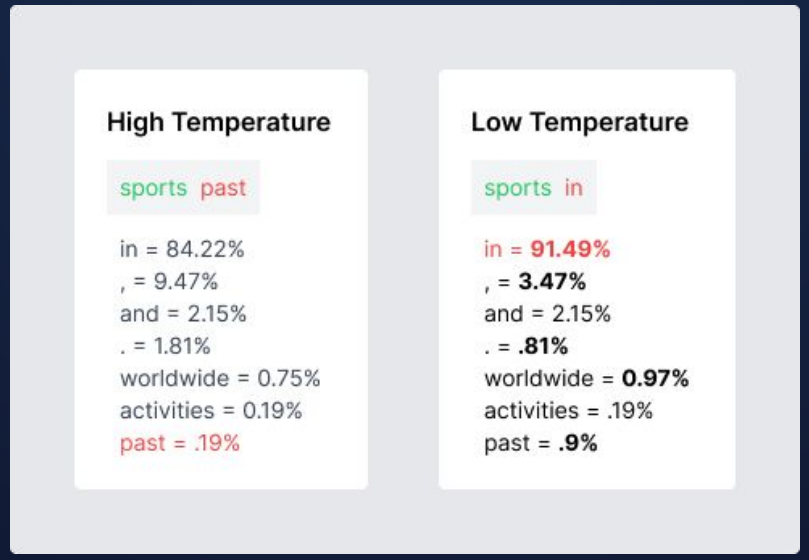
For better outputs from the right model:

1. Update configurations
2. Update your system and user prompts

Configuration Fine-tuning - **Temperature**

Temperature (between 0-2) defines the randomness of generated text

- 0 = Less random and more deterministic
- 2 = More random (and sometimes weirder output)



Source: https://www.prompthub.us/blog/understanding-openai-parameters-how-to-optimize-your-prompts-for-better-outputs

SecurityBoat Community
Connect, Learn & Grow!!

Configuration Fine-tuning - **Top P**

Top P (between 0-1) defines how many possible words to consider

- 0.1 = Lesser choice of words
- 1.0 = More choice of words

Source: https://www.prompthub.us/blog/understanding-openai-parameters-how-to-optimize-your-prompts-for-better-outputs

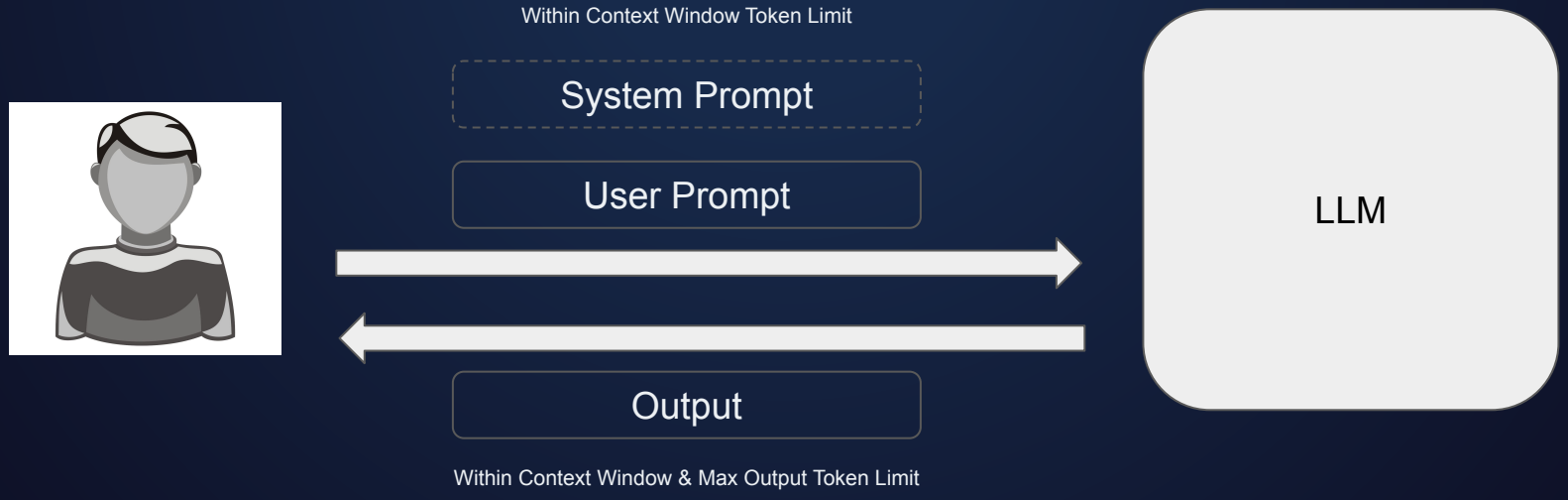
SecurityBoat Community
Connect, Learn & Grow!!

Prompt Fine-tuning

LLMs work with **tokens** - sequence of characters

System Prompt - Instructions provided to set context, tone and responses

User Prompt - Input by the end-user during their interaction



SecurityBoat Community
Connect, Learn & Grow!!

Demo #2

AI-Powered Hello World Program!

SecurityBoat Community
Connect, Learn & Grow!!

Prompt Fine-tuning - **One-Shot / Few-Shot Prompts**

LLMs are good at performing **zero-shot** tasks

Ex: “Make this paragraph better”

One-shot / Few-Shot Prompts - Providing examples in prompt

(c) Zero-shot

Q: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?
A: The answer (arabic numerals) is

(Output) 8 ✗

(b) Few-shot-CoT

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?
A: Roger started with 5 balls. 2 cans of 3 tennis balls each is 6 tennis balls. 5 + 6 = 11. The answer is 11.

Q: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?
A:

(Output) The juggler can juggle 16 balls. Half of the balls are golf balls. So there are 16 / 2 = 8 golf balls. Half of the golf balls are blue. So there are 8 / 2 = 4 blue golf balls. ***The answer is 4.*** ✓

Source: https://www.promptingguide.ai/techniques/cot

SecurityBoat Community
Connect, Learn & Grow!!

Prompt Fine-tuning - **Chain of Thought**

Chain of Thought - Asking the LLM to “think step-by-step” or giving detailed examples

(d) Zero-shot-CoT (Ours)

Q: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?
A: ***Let's think step by step.***

(Output) There are 16 balls in total. Half of the balls are golf balls. That means that there are 8 golf balls. Half of the golf balls are blue. That means that there are 4 blue golf balls. ✓

(b) Few-shot-CoT

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?
A: Roger started with 5 balls. 2 cans of 3 tennis balls each is 6 tennis balls. 5 + 6 = 11. The answer is 11.

Q: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?
A:

(Output) The juggler can juggle 16 balls. Half of the balls are golf balls. So there are 16 / 2 = 8 golf balls. Half of the golf balls are blue. So there are 8 / 2 = 4 blue golf balls. **The answer is 4.** ✓

Source: https://www.promptingguide.ai/techniques/cot

SecurityBoat Community
Connect, Learn & Grow!!

Going Beyond Large Language Models

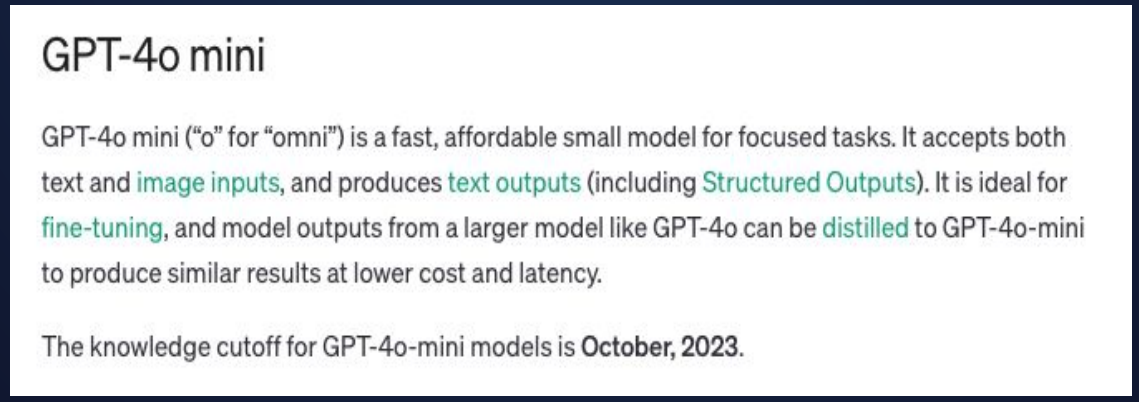
SecurityBoat Community
Connect, Learn & Grow!!

Problems with LLMs

LLMs are trained of data that was public and maybe old now

- News articles
- Documentation
- Didn't train on your company's internal data

Text generation isn't sufficient. Action is more valuable.



GPT-4o mini

GPT-4o mini ("o" for "omni") is a fast, affordable small model for focused tasks. It accepts both text and image inputs, and produces text outputs (including Structured Outputs). It is ideal for fine-tuning, and model outputs from a larger model like GPT-4o can be distilled to GPT-4o-mini to produce similar results at lower cost and latency.

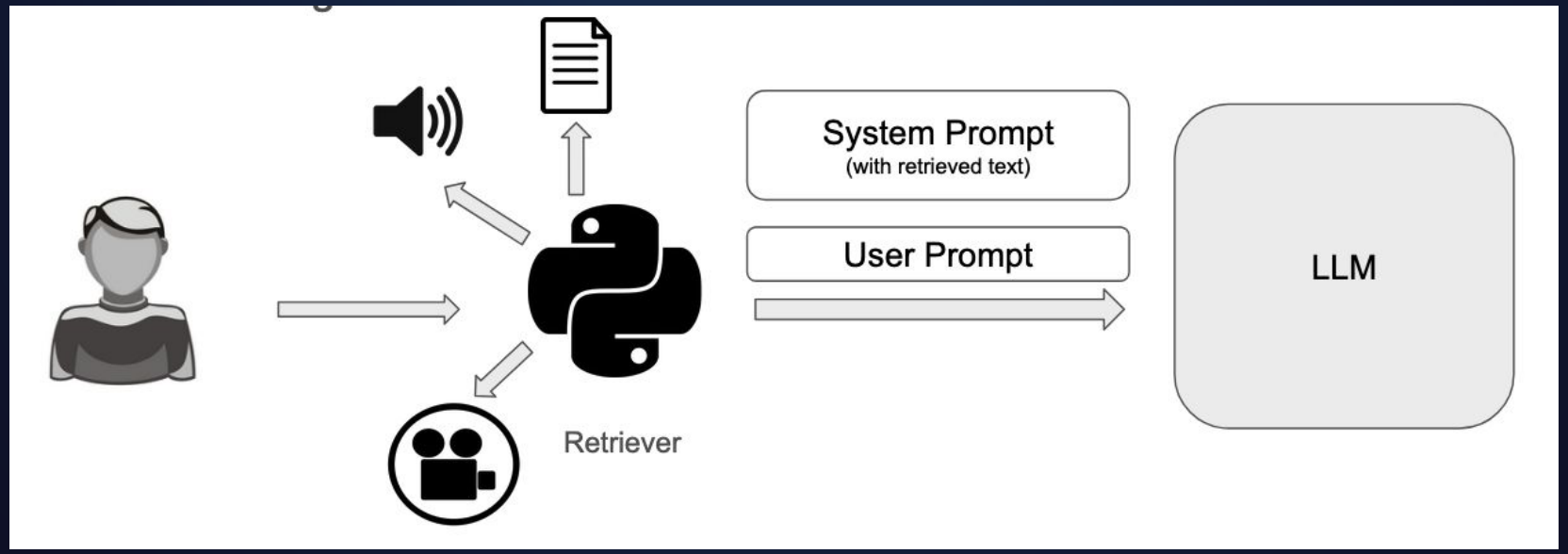
The knowledge cutoff for GPT-4o-mini models is **October, 2023**.

Source: https://platform.openai.com/docs/models

Retrieval Augmented Generation (RAG)

You fetch the relevant data and pass on the prompt to LLM

Multimodal RAG fetches relevant data from different data types - text, audio, video & images



AI Agents

Allows taking actions (writing code, executing external tools, etc)

Semi-autonomous or fully-autonomous

Can be used for open-ended problems (fix the code, find the best XYZ, etc)

> • Agents, on the other hand, are systems where LLMs dynamically direct their own processes and tool usage, maintaining control over how they accomplish tasks.

Source: https://www.anthropic.com/research/building-effective-agents

SecurityBoat Community
Connect, Learn & Grow!!

LLMs in Security

Practical Examples

- Retrieve data from personal notes
  - Redteam/Pentest/Freelance engagements
  - Previous unreported low hanging bug
- Offensive tasks involving natural language
  - Phishing / Vishing
  - Social Engineering
- Code/Policy/Query/CICD pipeline generation
  - Malware/Red team tools
- Threat modeling/Design Reviews

SecurityBoat Community
Connect, Learn & Grow!!

Demo #3

Using LLM for Cloud Security Research

SecurityBoat Community
Connect, Learn & Grow!!

My Thoughts

LLMs are useful when you have **non-recurring unstructured random inputs** (possibly in human language) that **can't be automated** and the **produced output is not detrimental**

Good examples:

- Idea to Code/Query/Image/Text
- Contextual actions
- Threat modeling / Design Reviews*

Bad examples:

- Replacing well performing custom built ML models
- Replacing dashboards
- Using AI as security tools

SecurityBoat Community

Connect, Learn & Grow!!

THANK YOU