LOCAL LARGE LANGUAGE MODELS

PATRIK BLOMMASKOG

CADEC 2025.01.23 & 2025.01.29 | CALLISTAENTERPRISE.SE

CALLISTA

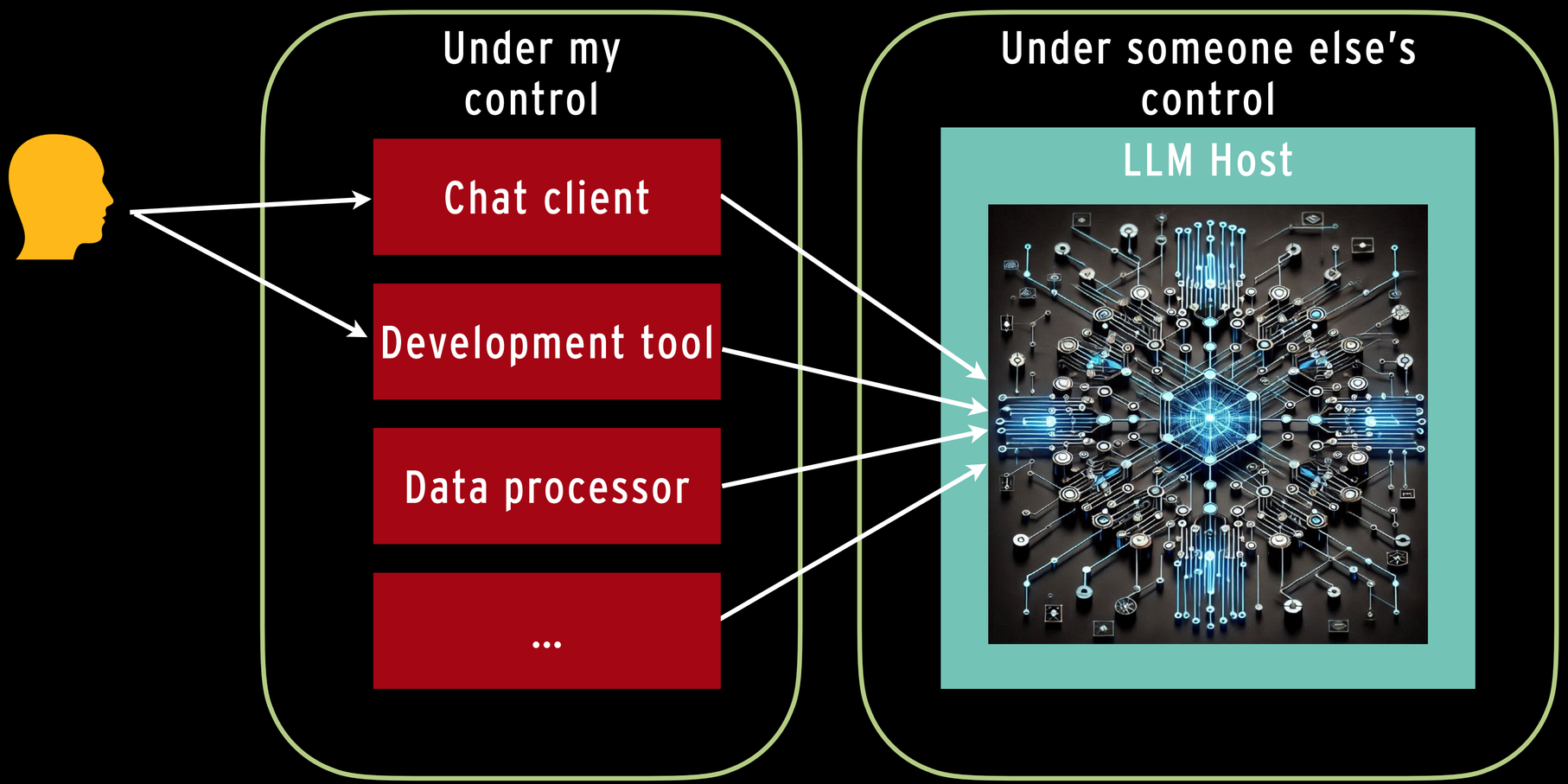
STIG VILDMARK - THE WILDERNESS GUIDE

- Knows all about the wilderness
- Prefers to be offline
- Has a great personality
- Makes use of what you are carrying

AGENDA

- Run an LLM locally - what and why
- Technical aspects
- Use in practice 🤞

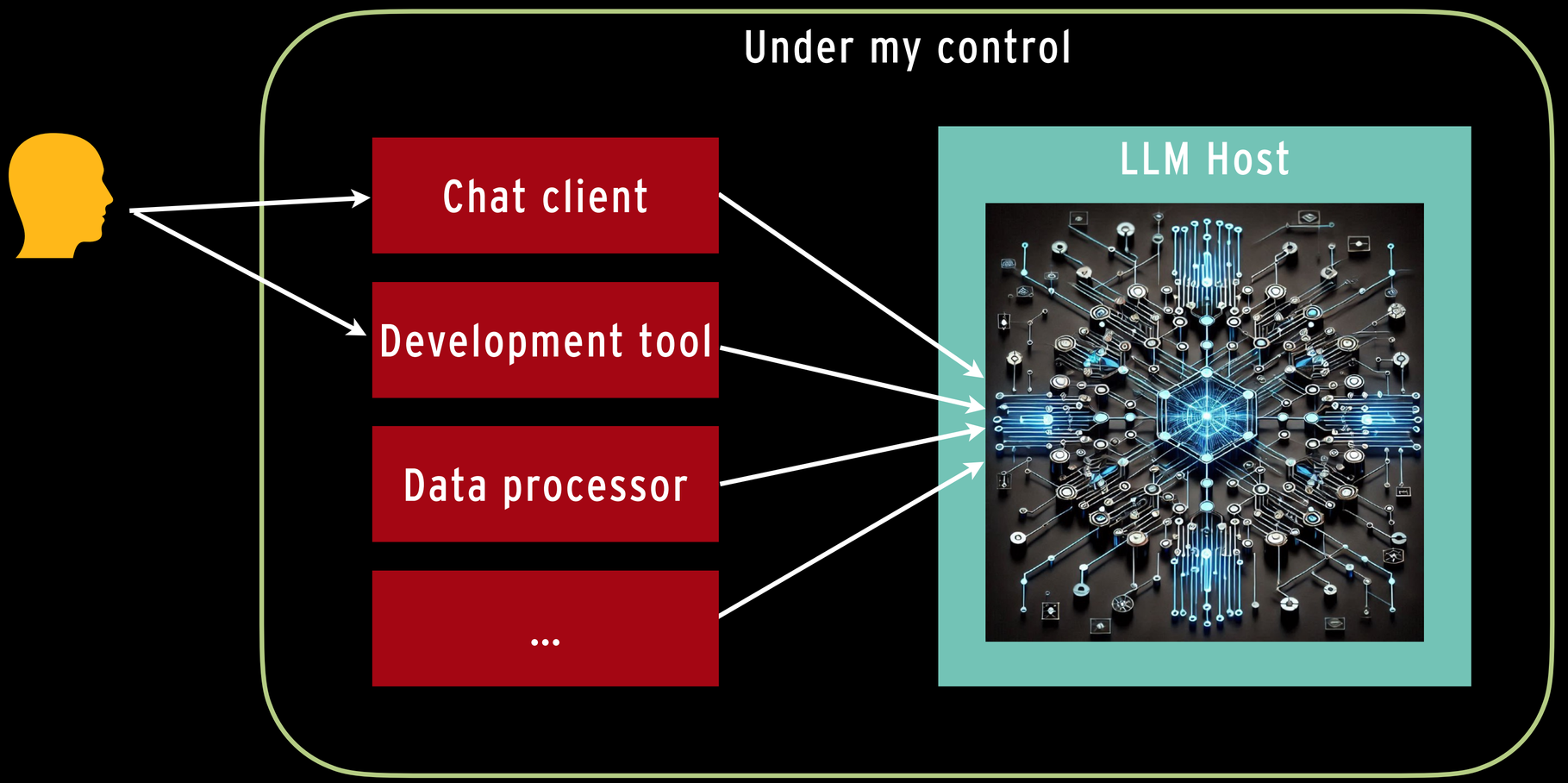
Using a cloud provided LLM



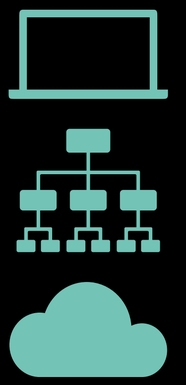
GPT-4
Claude
Gemini
…

DEFINITION

Run an LLM locally



Llama
Qwen
Mistral
Deep Seek
...



TRAINING VS INFERENCE

- Training is very expensive and slow
- Fine tuning is less so
- Inference is cheap(ish) and fast(ish)

DEMO: OLLAMA

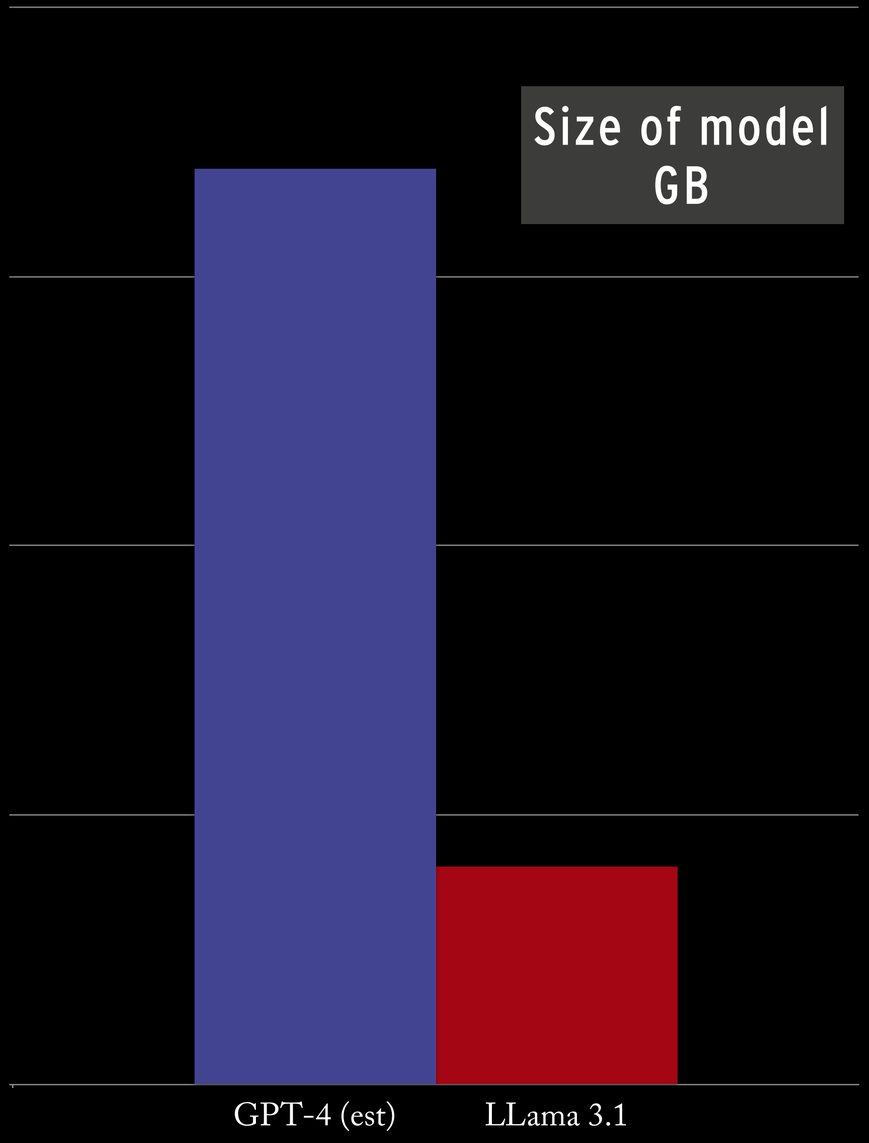
”Never work with animals or children”

– W.C. Fields, ~1930s

”Never demo with stochastic tools”

– Me, soon

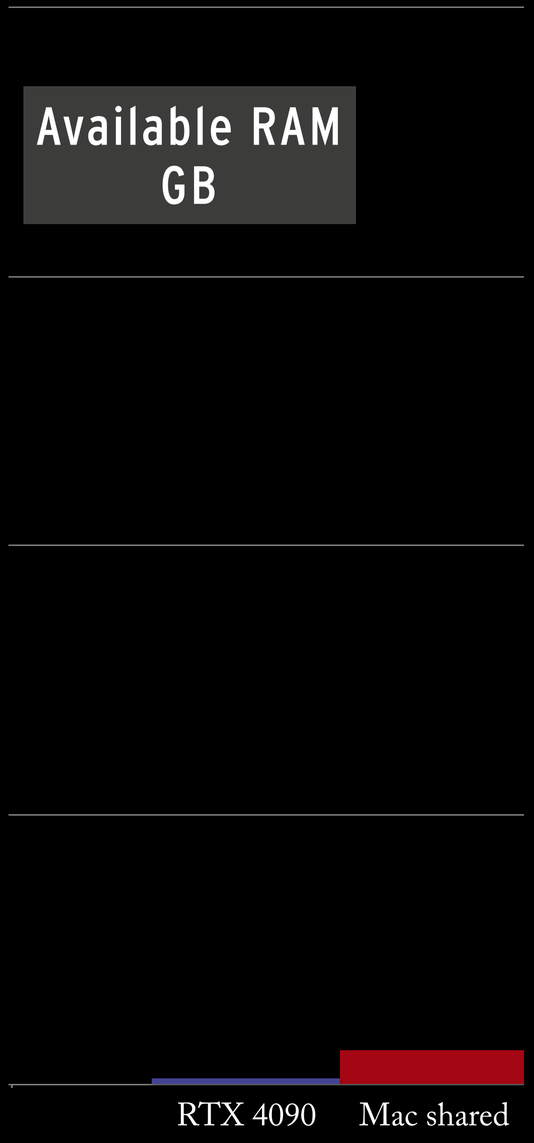
RESOURCES



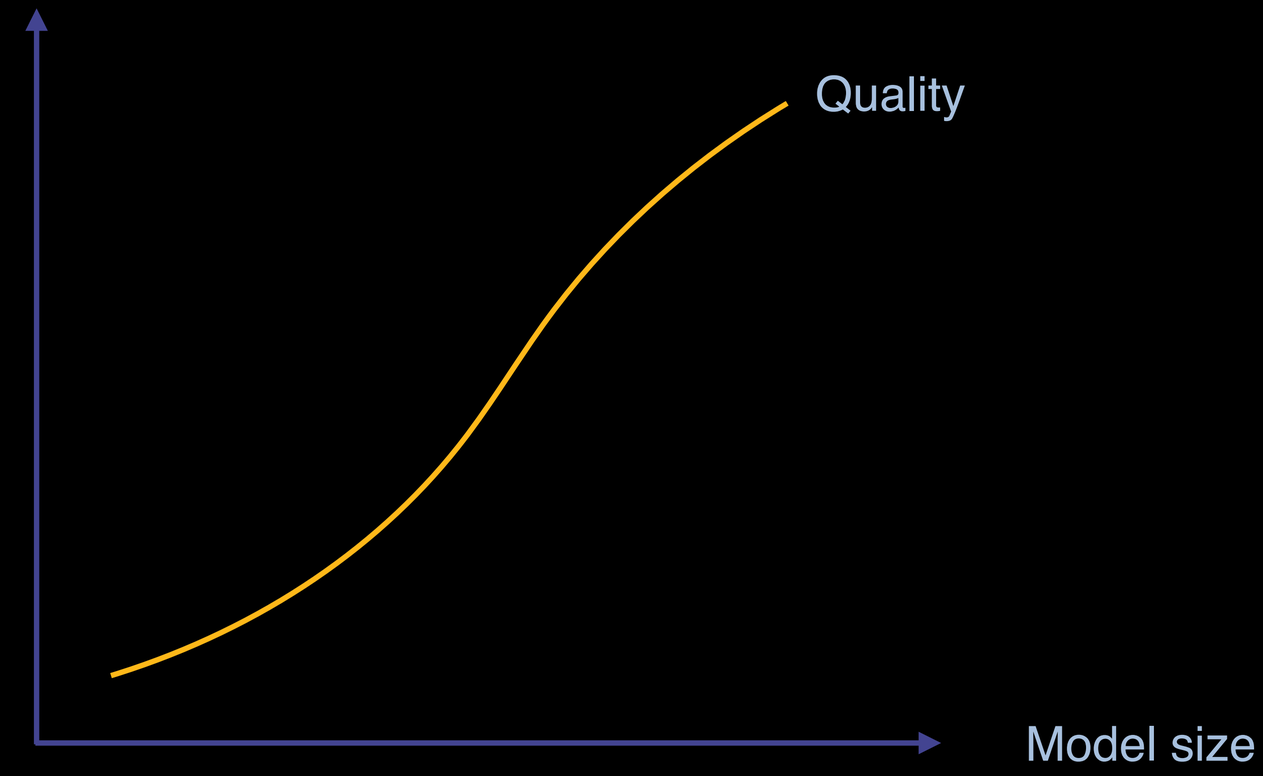
CPU

GPU

NPU



SIZE VS QUALITY



SIZE VS QUALITY AND SPEED

LOCAL LLM SALESMAN OF THE YEAR

- A local model will likely be slower, dumber, or both

WHICH ONE IS BETTER?

IT DEPENDS

Images: Porsche Sprint Challenge, Fergusson, Wikipedia

LOCALLY PRODUCED VS CLOUD

- Advantages
  - Privacy
  - Offline
  - Cost
  - Immutability
  - Adaptability
  - Censorship
  - Licensing
- Disadvantages
  - Not usable from all tools
  - Speed
  - Smartness
  - Not following state of the art
  - Up-front cost
  - Maintenance

WAYS TO FIT THE MODEL

- Smaller representation
- Specialisation, e.g. for coding

SMALLER REPRESENTATION

Quantising

Original

Fewer bits

Lower resolution

Fewer parameters

End result

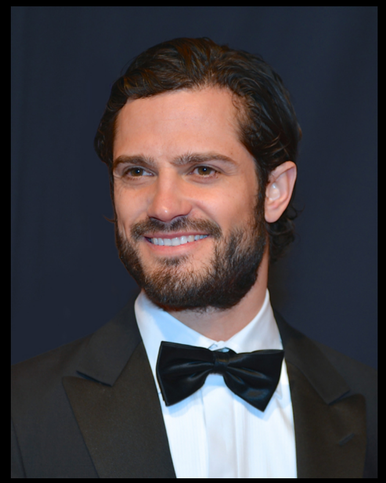
Image: Wikipedia

KING SIZE

Silvia and Carl XVI Gustaf

Victoria

Madeleine



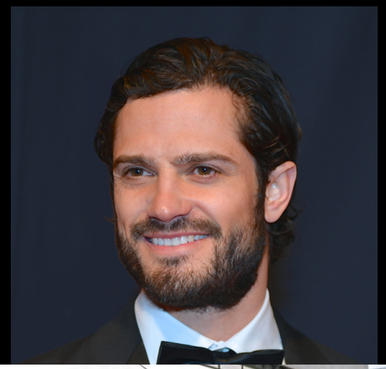
Carl Philip

CALLISTA

Images: Wikipedia

KING SIZE

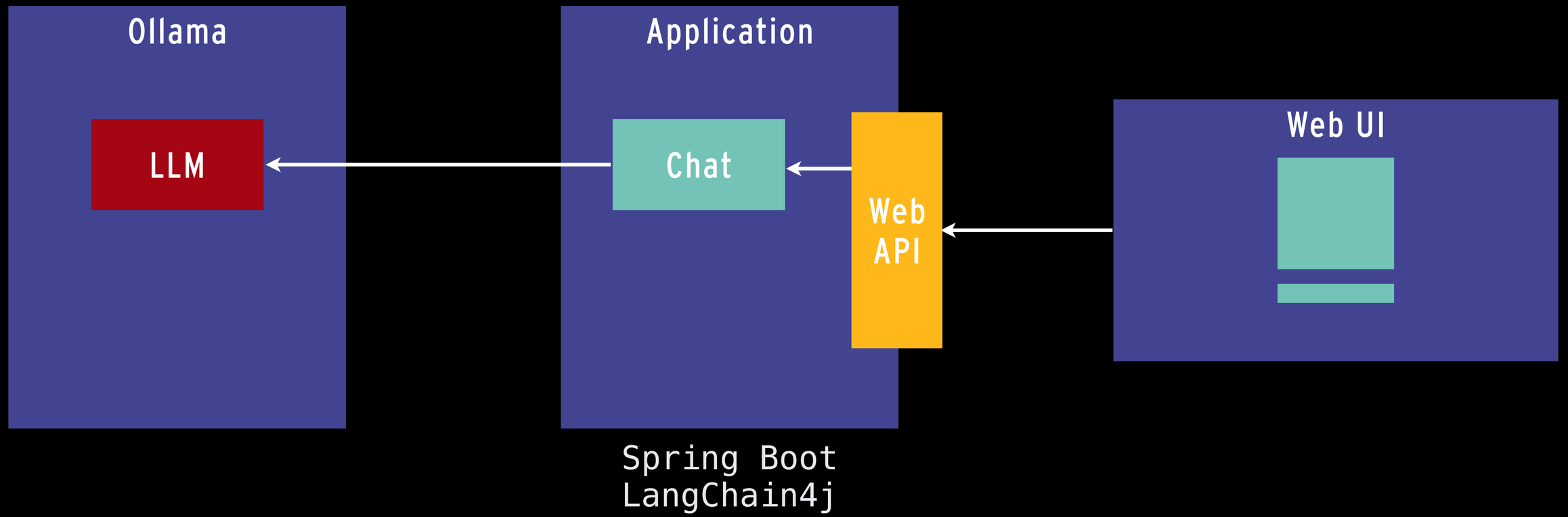
Silvia and Carl XVI Gustaf



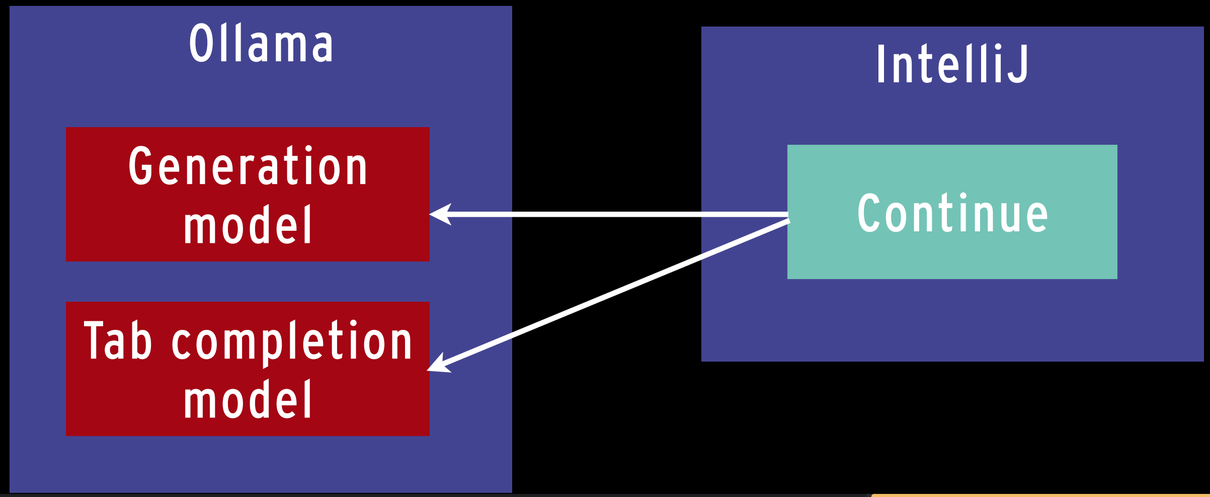
STIG VILDMARK

- Knows all about the wilderness
- Prefers to be offline
- Has a great personality
- Makes use of what you are carrying

A CHAT APPLICATION



LOCAL LLM FOR DEVELOPMENT TOOLS



ChatController.java

```
17
18    @PostMapping("/chat")
19    public ResponseEntity<String> chat(@RequestBody ChatRequest chatRequest) {
20        String response = chat.chat(chatRequest.message());
21
22        HttpHeaders headers = new HttpHeaders();
23        headers.setContentType(MediaType.valueOf("text/plain; charset=UTF-8"));
24        return new ResponseEntity<>(response, headers, HttpStatus.OK);
25    }
26 }
27
```

Add a test case

qwen2.5-coder:32b ▾ Enter

Continue

What on earth does this code do?

qwen2.5-coder:32b @ ⌘⏎ @codebase ⏎ Enter

5 context items

This collection of files appears to be part of a Java Spring Boot application that integrates with a language model service via the langchain4j library. Let's break down what each file does:

build.gradle (0-37)

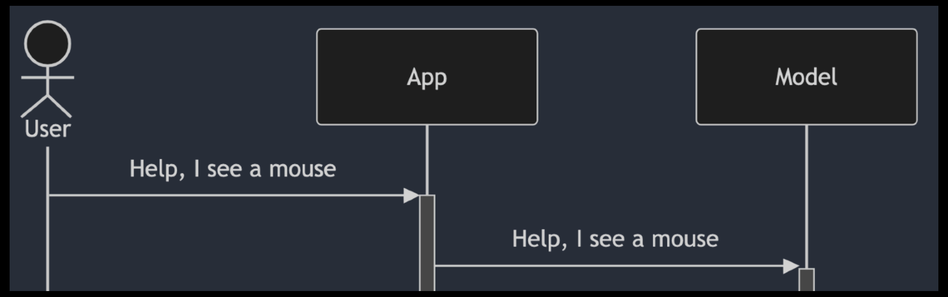
Purpose: This is the Gradle build script for the project.

- **Plugins applied**:
  - ChatController.java : Provides Java compilation,

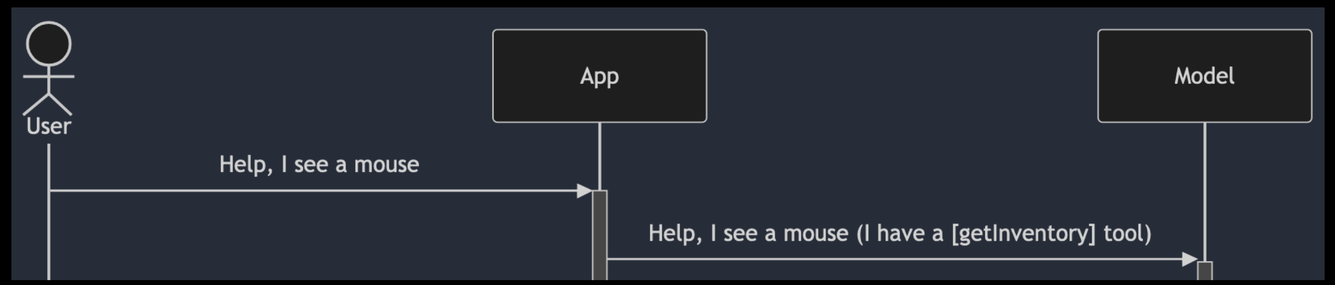
STIG VILDMARK

- Knows all about the wilderness
- Prefers to be offline
- Has a great personality
- Makes use of what you are carrying

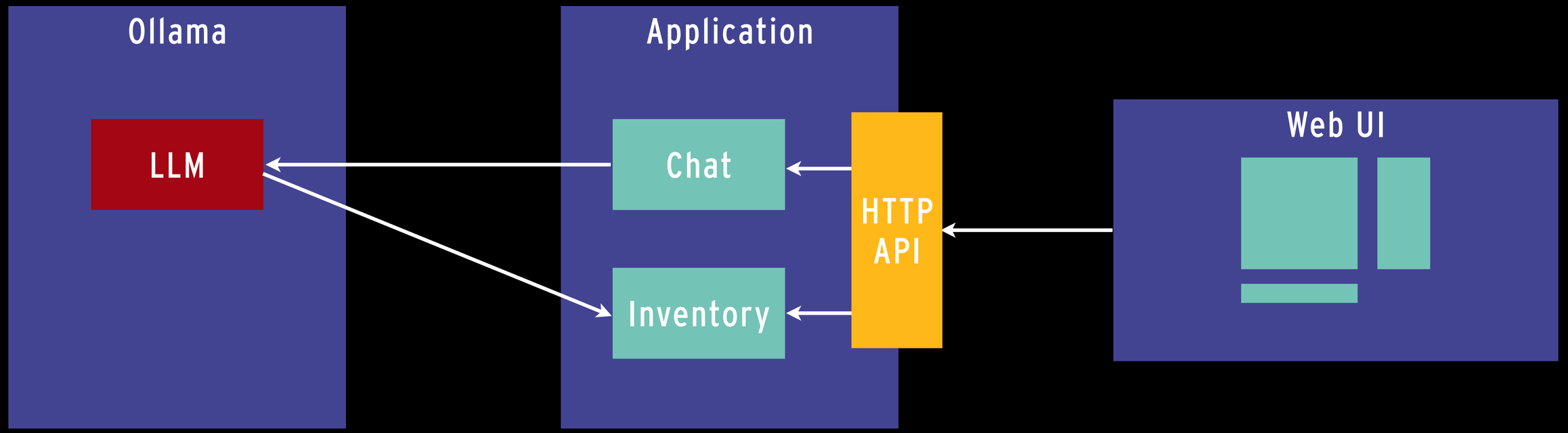
THE MODEL IS BLIND



TOOLS

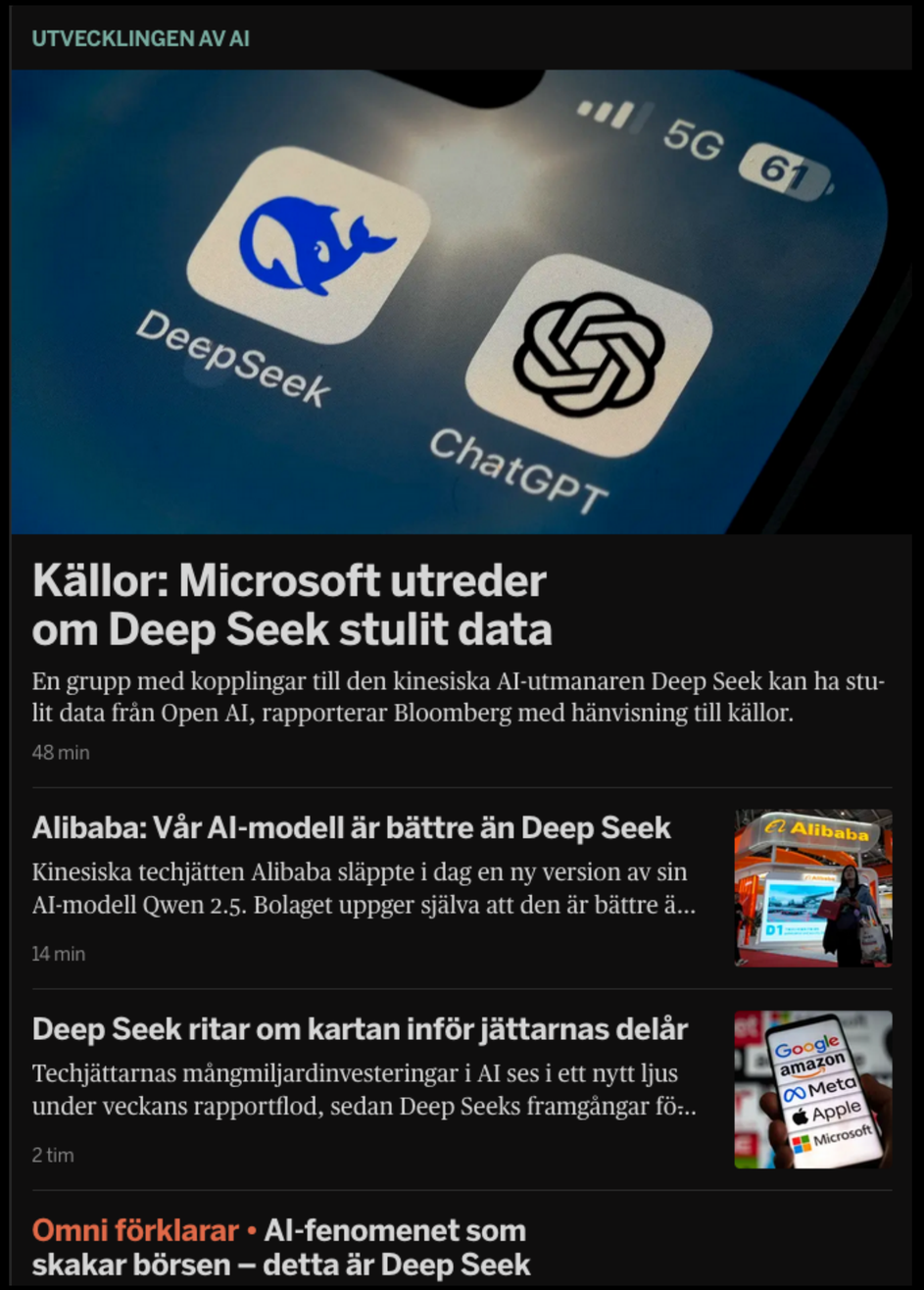

User
App
Model
Help, I see a mouse
Help, I see a mouse (I have a [getInventory] tool)

THE APPLICATION



SUMMARY

- Very easy to get started
- Unpredictability can be a challenge
- Disappointing as a development tool
- Size does matter
- You can’t plow with a race car
- Fast moving field



SOFTWARE AND SITES

Demonstrated

- *Ollama:* https://ollama.com
- *LangChain4j:* https://github.com/langchain4j
- *Continue:* https://www.continue.dev

Also useful

- *LM Studio:* https://lmstudio.ai
- *Hugging Face:* https://huggingface.co