Demystifying the AI Jungle

AI India
Gurugram, India
April 12, 2026

Rainer Hahnekamp
Soverius AI
www.soverius.ai

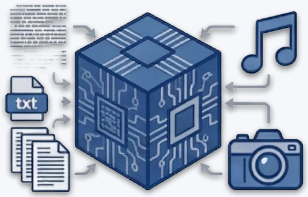
SOVERIUS
AI

Murat Sari

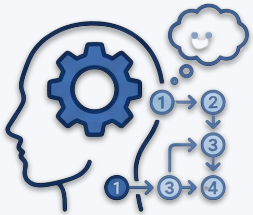
Rainer Hahnekamp

Agenda

Model



Pre-Training



Chain of Thought

Inference

Limitations

Agents/Tools

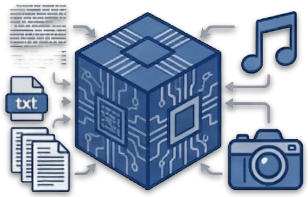
Context Size

Data Actuality

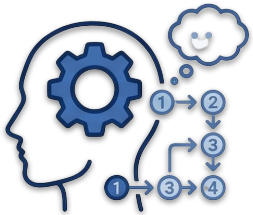
SOVERIUS AI

Agenda

Model



Pre-Training



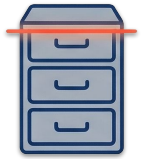
Chain of Thought

Inference

Harness

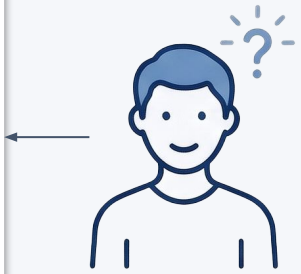
Limitations

Agents/Tools



Context Size

Data Actuality



SOVERIUS AI

Disclaimer

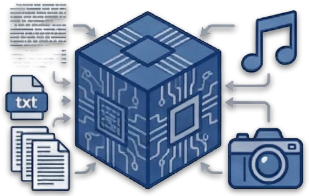
Simplified Examples

...but show the principles

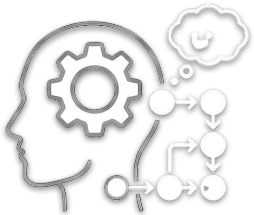
SOVERIUS AI

Pre-Training

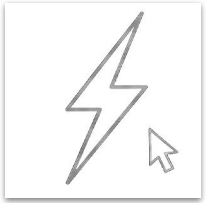
Model



Pre-Training



Chain of Thought



Inference

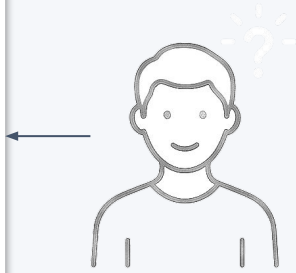
Harness

Limitations

Agents/Tools

Context Size

Data Actuality



SOVERIUS AI

Pre-Training

Text is fed in

```
Murat lives in Leoben and maintained OGRE.
Rainer is from Austria and likes Star Wars. He sees
programming as both hoby and profession.
...
```

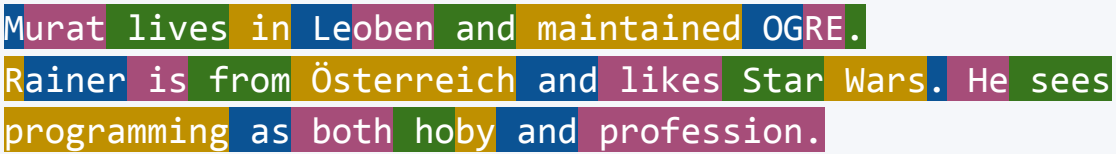
SOVERIUS AI

Pre-Training

Text is fed in

```
Murat lives in Leoben and maintained OGRE.
Rainer is from Austria and likes Star Wars. He sees
programming as both hoby and profession.
...
```

Text is split up into Tokens



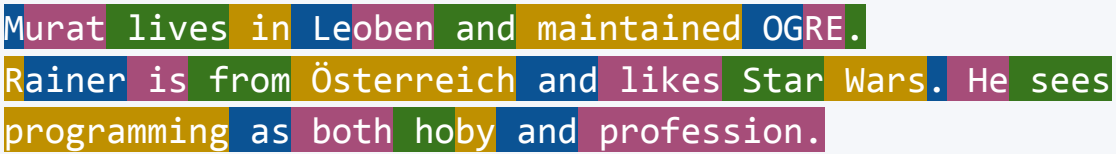
SOVERIUS AI

Pre-Training

Text is fed in

```
Murat lives in Leoben and maintained OGRE.
Rainer is from Austria and likes Star Wars. He sees
programming as both hoby and profession.
...
```

Text is split up into Tokens



Tokens become numbers

```
[
  44 58193 8354 306 2018 60006 326 26871 81993 1099 558,
  49 2573 382 591 70997 326 18861 11307 25778 13 1679 27432,
  23238 472 2973 312 30172 326 3872 13
]
```

SOVERIUS AI

Training starts...

Prediction 1/n

Murat

Prediction 1/n

```
Murat is

Murat Star Wars

Murat Rainer

Murat the

Murat lives
```

Prediction 2/n

Murat lives

Prediction 2/n

```
Murat lives the

Murat lives ho

Murat lives by

Murat lives lives

Murat lives in
```

SOVERIUS AI

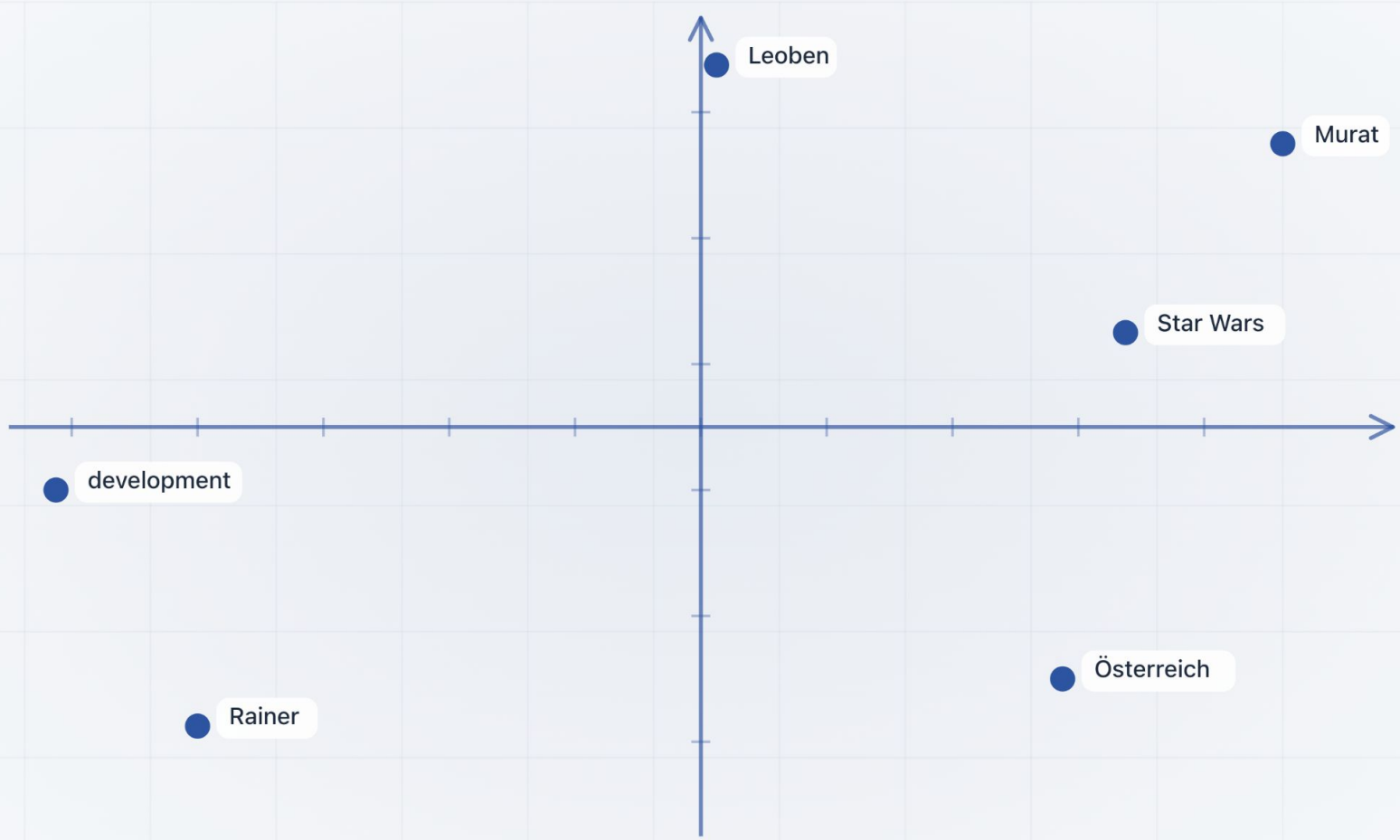
Prediction

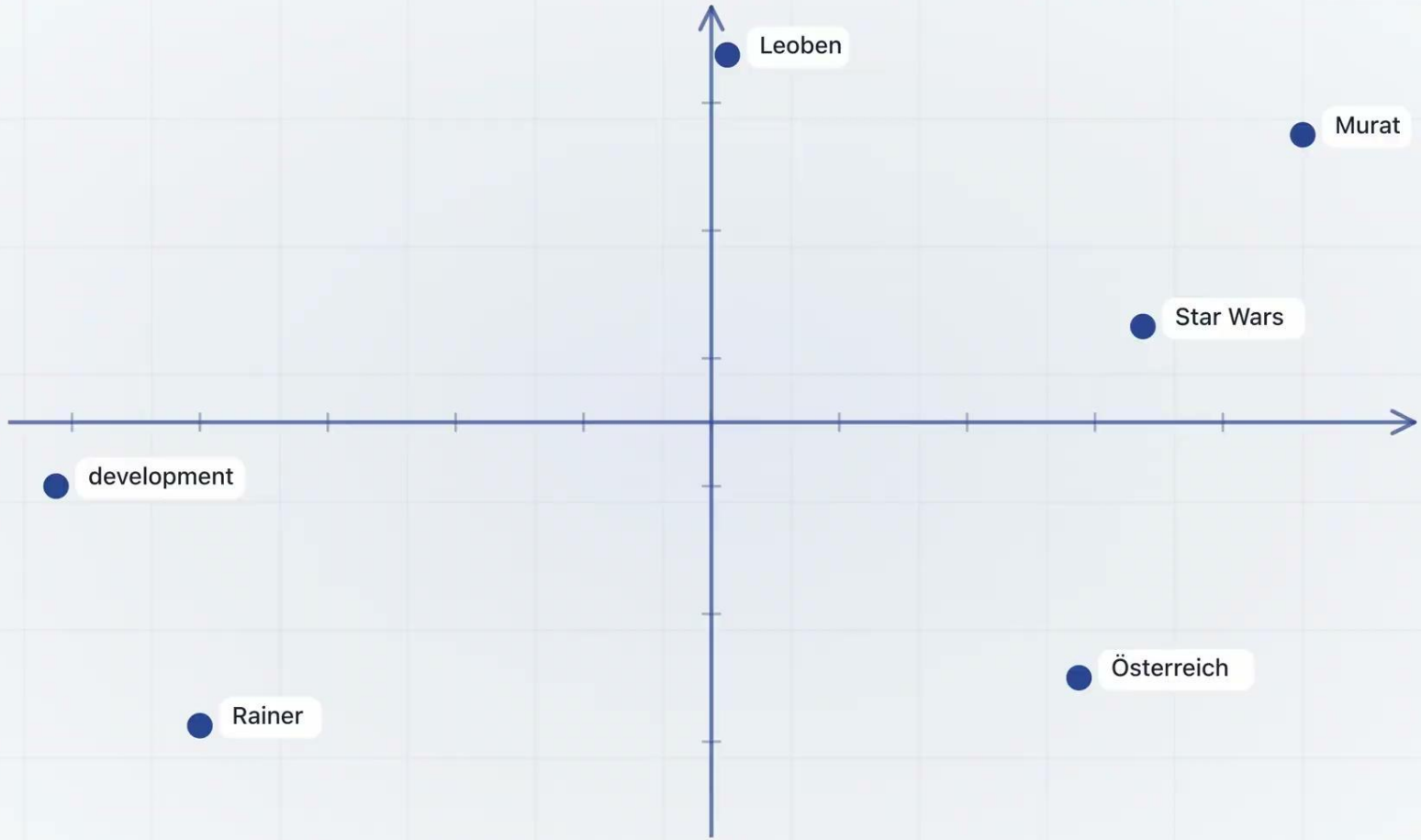
```
Murat

Murat lives

Murat lives in

Murat lives in Leoben

Murat lives in Leoben and

Murat lives in Leoben and maintained

Murat lives in Leoben and maintained OGRE

...
```

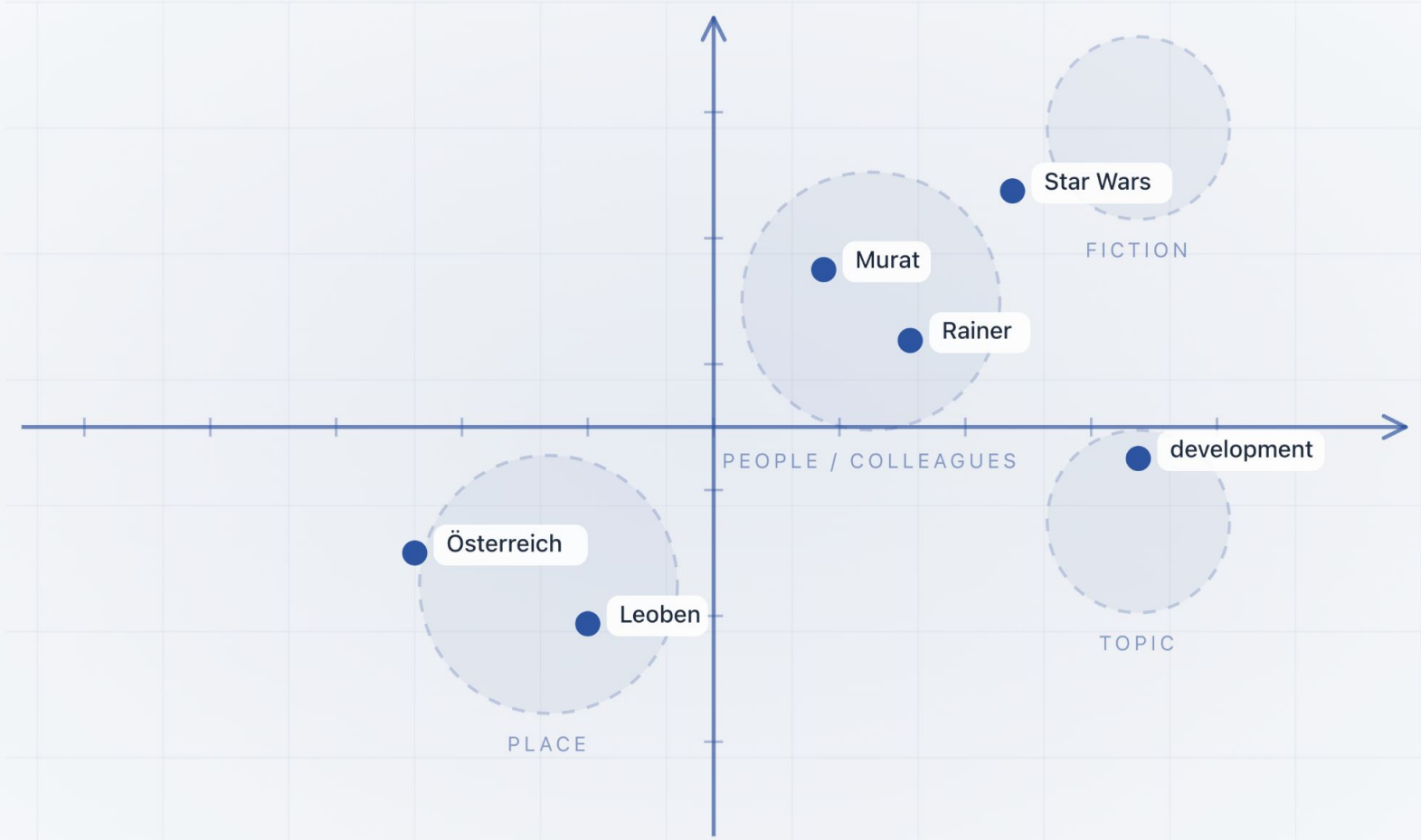
Prediction

```
Murat lives

Murat lives in

Murat lives in Leoben

Murat lives in Leoben and

Murat lives in Leoben and maintained

Murat lives in Leoben and maintained ORGE

Murat lives in Leoben and maintained OGRE.

...
```

SOVERIUS AI


Leoben
Murat
Star Wars
development
Österreich
Rainer

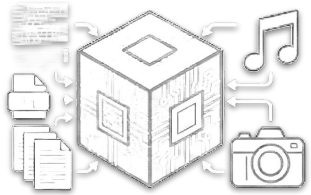

Leoben
Murat
Star Wars
development
Österreich
Rainer


Star Wars
FICTION
Murat
Rainer
PEOPLE / COLLEAGUES
development
TOPIC
Österreich
Leoben
PLACE

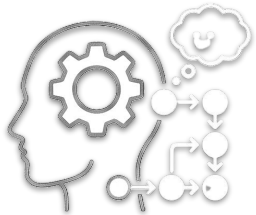
Inference

Model

Harness



Pre-Training



Chain of Thought

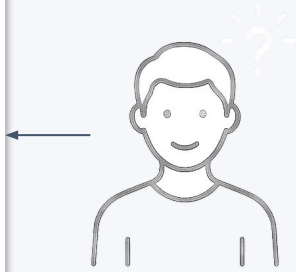
Inference

Limitations

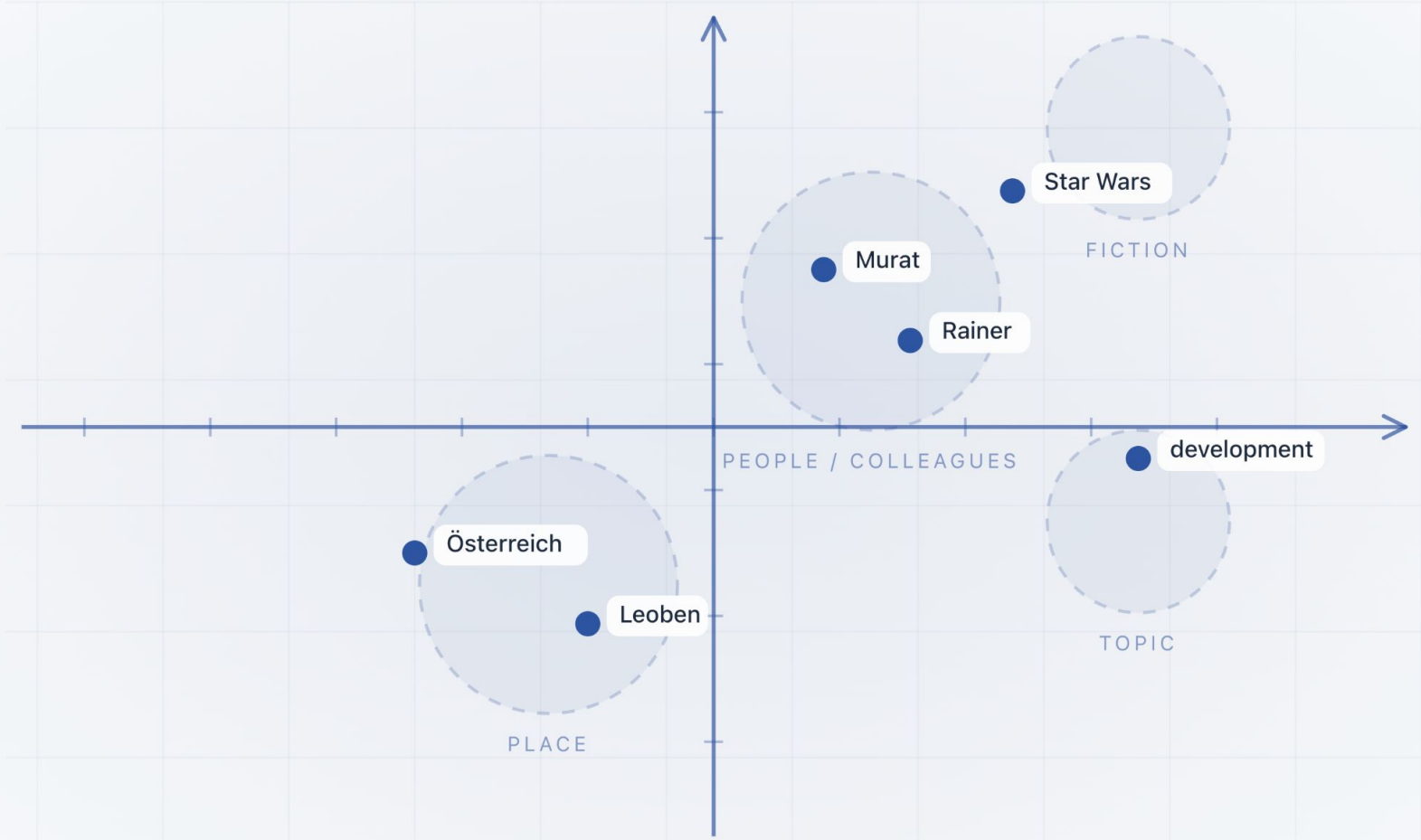
Agents/Tools

Context Size

Data Actuality



SOVERIUS AI


Star Wars
FICTION
Murat
Rainer
PEOPLE / COLLEAGUES
development
TOPIC
Österreich
Leoben
PLACE

Predictions

Training Data (Model has seen)

Murat lives in Leoben.

Rainer is from Austria.

User Input (Model has NEVER seen)

Dhananjay lives in ...

SOVERIUS AI

Predictions

Training Data (Model has seen)

Murat lives in Leoben.

Rainer is from Austria.

User Input (Model has NEVER seen)

Dhananjay lives in Austria

Dhananjay lives in New York

Dhananjay lives in Delhi

Dhananjay lives in Leoben

SOVERIUS AI

Other potential dimensions

SOVERIUS AI

Predictions

Training Data (Model has seen)

Murat lives in Leoben.

Rainer is from Austria.

User Input (Model has NEVER seen)

Dhananjay lives in Austria 15%

Dhananjay lives in New York 25%

Dhananjay lives in Delhi 35%

Dhananjay lives in Leoben 5%

SOVERIUS AI

Predictions

Training Data (Model has seen)

Murat lives in Leoben.

Rainer is from Austria.

User Input (Model has NEVER seen)

Dhananjay lives in Austria 15%

Dhananjay lives in New York 25%

Dhananjay lives in Delhi 35%

Dhananjay lives in Leoben 5%

SOVERIUS AI

Temperature

Dhananjay lives in Delhi 35%

Dhananjay lives in Austria 15%

Dhananjay lives in New York 25%

Dhananjay lives in Leoben 5%

Temperature = 0
(Take the highest probability)

Temperature > 0
(Select randomly)

Hallucination: A Question of Interpretation

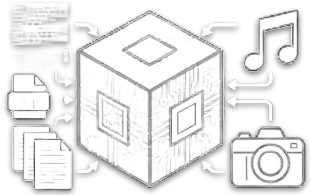
- Usually on fact-based context
- Could be because model picks tokens with lower probability
  - Decrease Temperature 🥶
- Bad training
  - Re-evalute training data
- "Desired Hallucination"
  - Non-deterministic context
  - Creativity is required

SOVERIUS AI

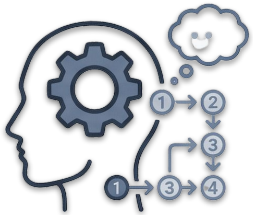
Chain of Thought (CoT) "Thinking"

Model

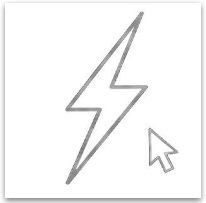
Harness



Pre-Training



Chain of Thought



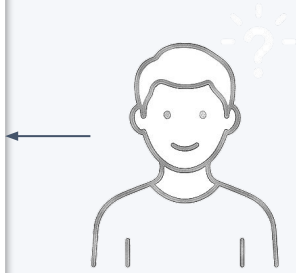
Inference

Limitations

Agents/Tools

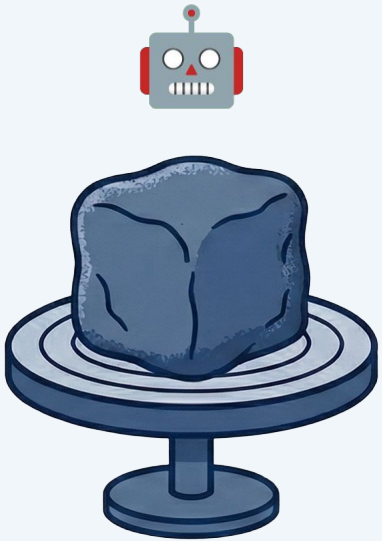
Context Size

Data Actuality

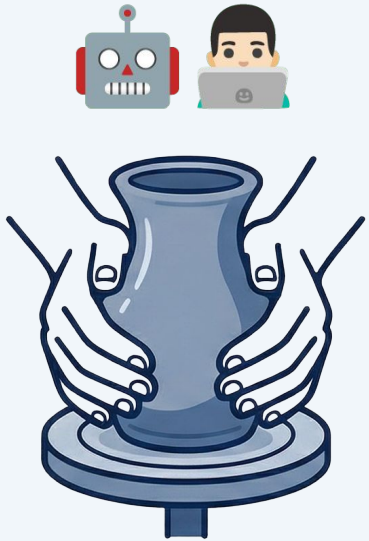


SOVERIUS AI

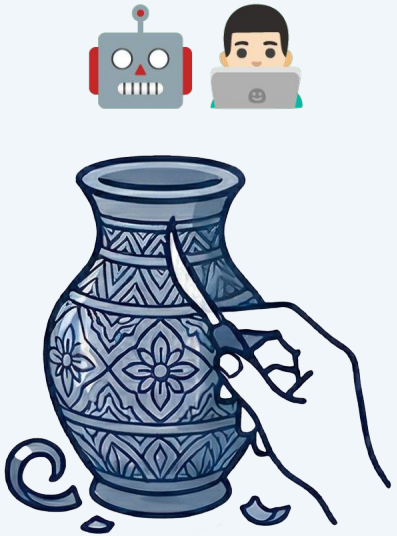
Model Generation Phases



Pre-Training



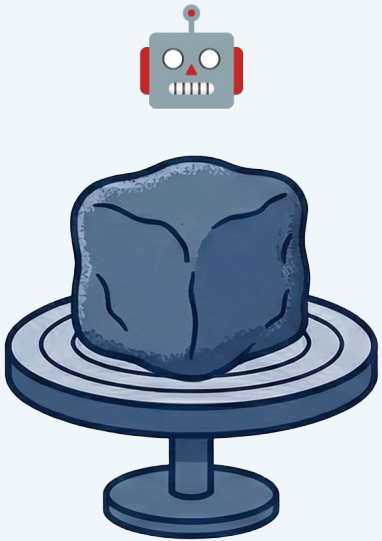
Fine-Tuning



Reinforcement Learning

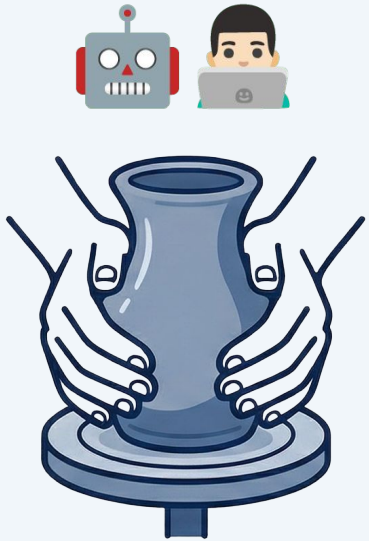
SOVERIUS AI

Model Generation Phases



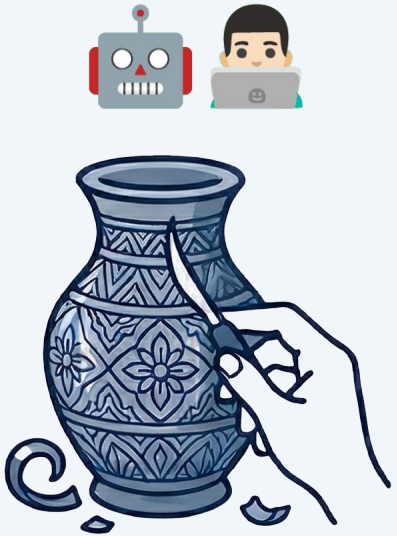
Pre-Training

Only next token prediction



Fine-Tuning

Domain Knowledge Engineering, Coding, Accounting,...

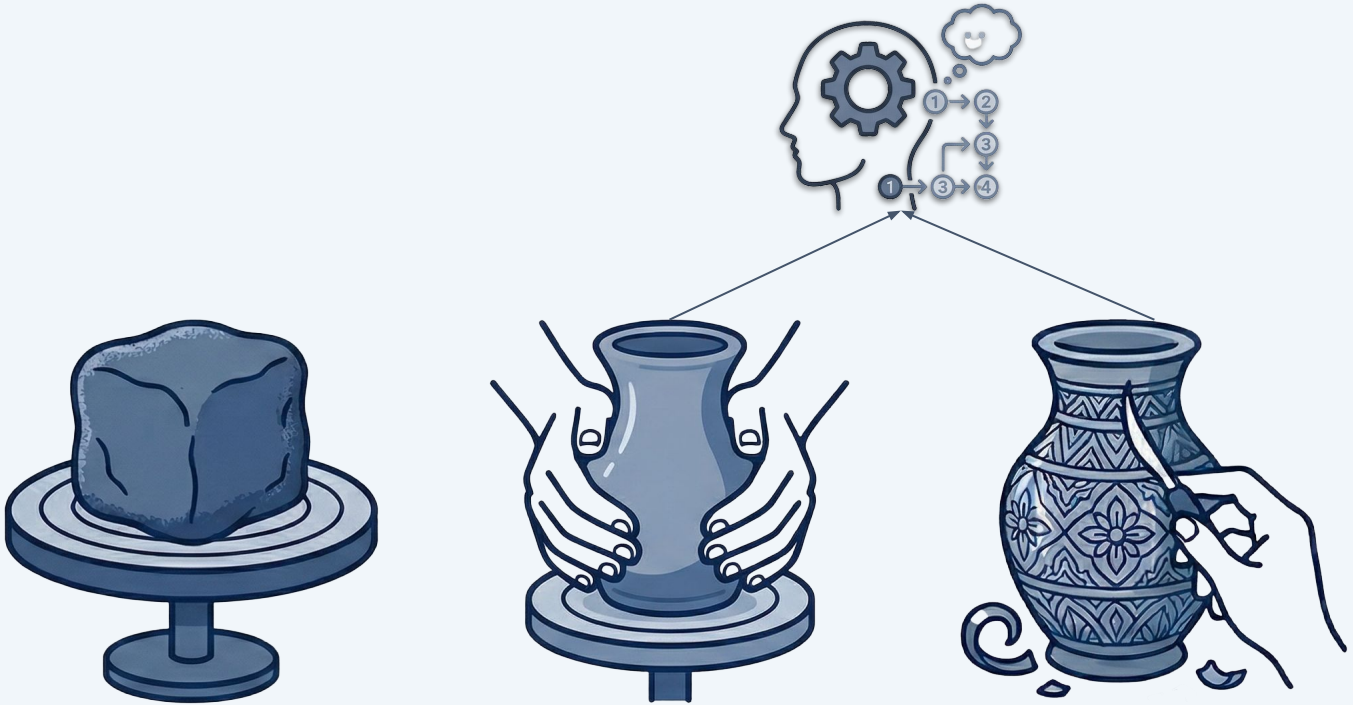


Reinforcement Learning

Tonality, Style of Responses

SOVERIUS AI

Model Generation Phases



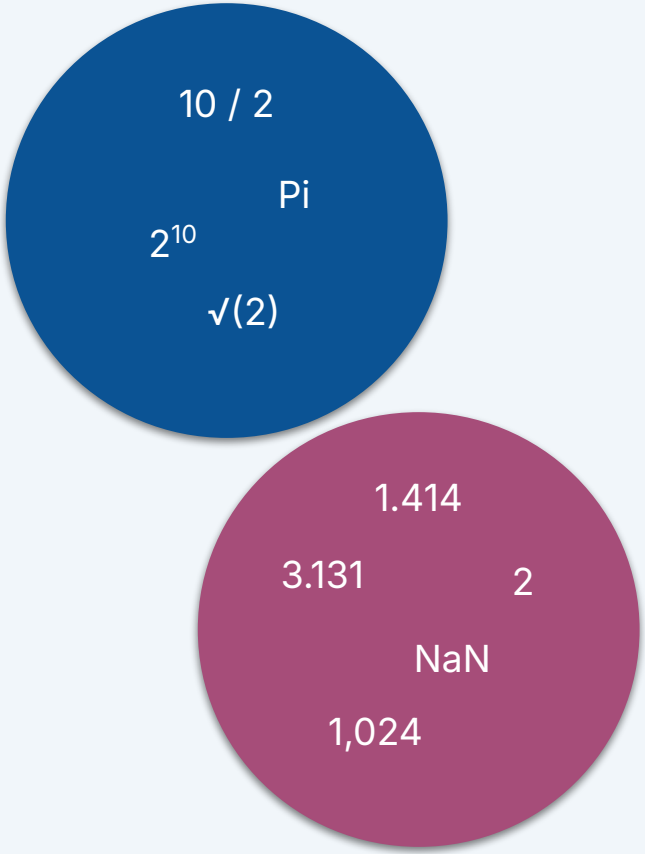
Pre-Training

Fine-Tuning

Reinforcement Learning

SOVERIUS AI

Inference without CoT



Training Data (Model has seen)

10 / 2 = 2.

Pi is 3.141.

2^{10} === 1,024.

√(2) is 1.414.

User Input (Model has NEVER seen)

What is √ of 4? 2

What is √ of 4? 3.141

What is √ of 4? 1,024

SOVERIUS AI

Inference with CoT

What is √(4)?

How would you calculate √(4)?

1. I need to solve a math task
2.I need to run `echo "sqrt(4)" | bc`
3. I need to return the output

2

Do what you just said!

SOVERIUS AI

Inference with CoT

What is √(4)?

2

Inference with CoT

🤔How does the model run the script???

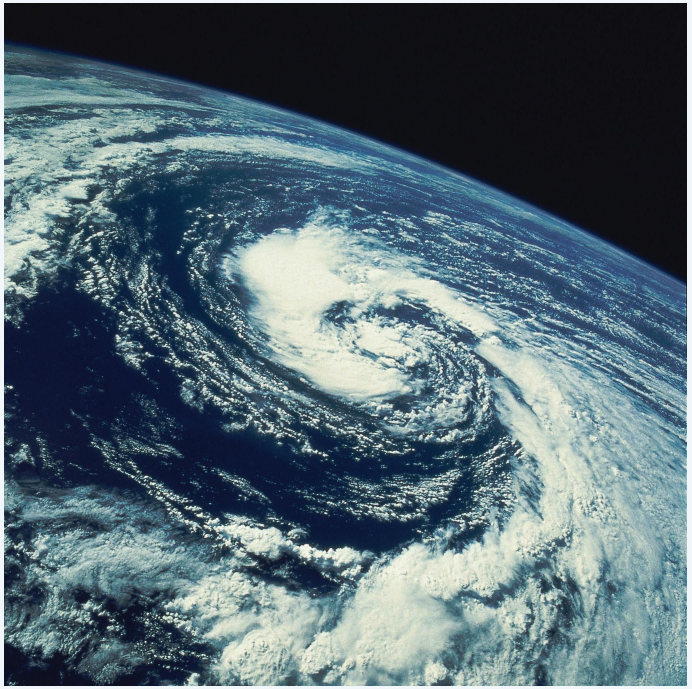
3. I need to return the output

2

Do what you just said!

SOVERIUS AI

Model is not a "black box"



- Rules are known
- Data (Weights, Parameters) is massive
- We can see everything
- It is just too much information

SOVERIUS AI

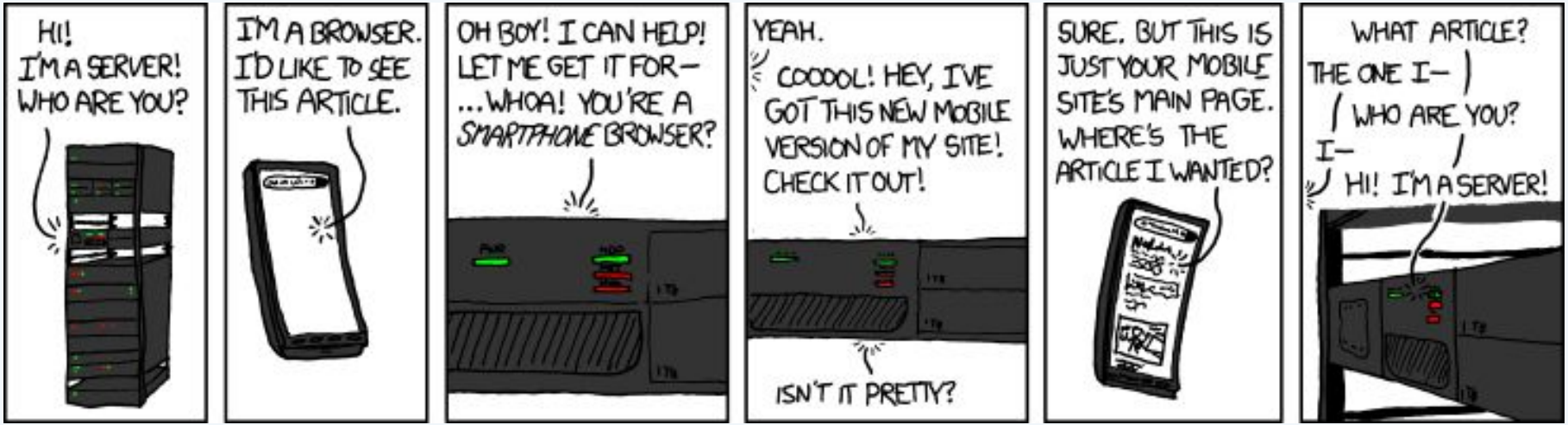
Model does not "understand" language

- Model does not understand us
- Model knows the structure of the language
- It is trained to mimic us
- Limited on the training data

Actor playing a Pianist
(https://www.classicfm.com/discover-music/periods-genres/film-tv/the-pianist-music-true-story-adrien-brody-piano/)

SOVERIUS AI

Model is stateless

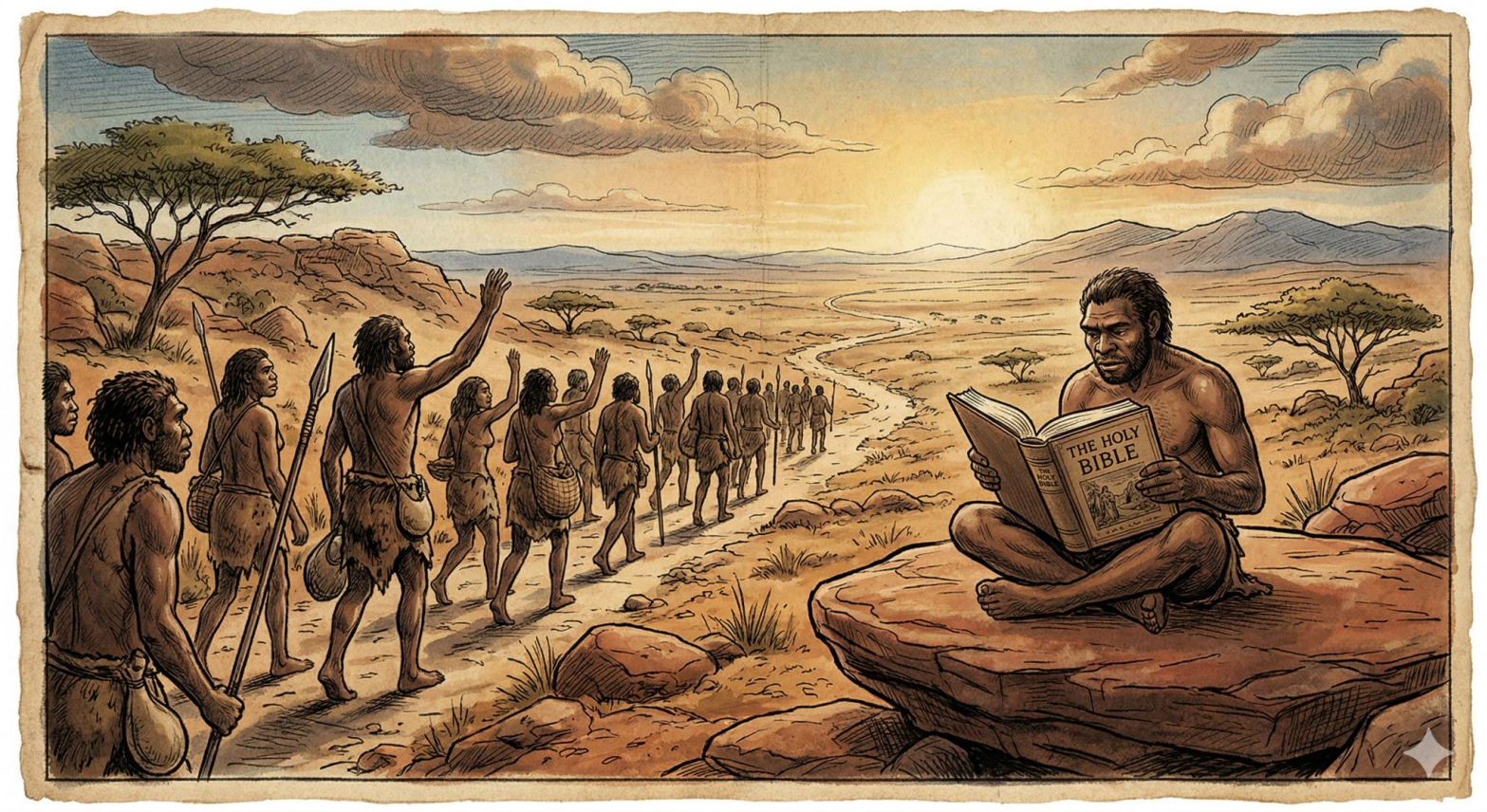


https://www.explainxkcd.com/wiki/index.php/File:server_attention_span.png

SOVERIUS AI

Amount of training data

One bible per week for 270,000 years



SOVERIUS AI

A few numbers

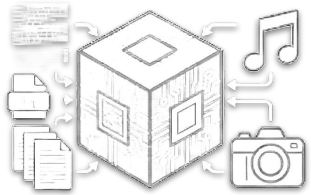
- GPT 4o/5 has **201,088 tokens**
- Inference GPT 5.4:
  - Context Size: max. **1,000,000 tokens**
  - Costs per 1M Input: **$2.50 (small), $5.00 (large)**
  - Costs per 1M Output: **$15.00 (small) $22.00 (large)**
- Costs for Training Gemini 3: **$955,000,000 - $1,910,000,000**
  - ~ $191 Millon for final test run
  - ~ Final test run is 10% to 20% of total costs
- Common Model Sizes (uncompressed):
  - Llama 3.1: **754 GB**
  - GPT 5.4: **~5-6 TB**
  - Gemini 3.1: **~40TB**
- Training material
  - Llama: **15,000.000,000,000 tokens**

SOVERIUS AI

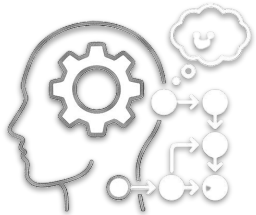
Limitation: Agents/Tools

Harness

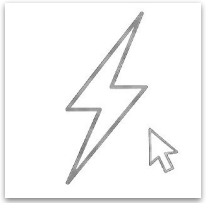
Model



Pre-Training



Chain of Thought



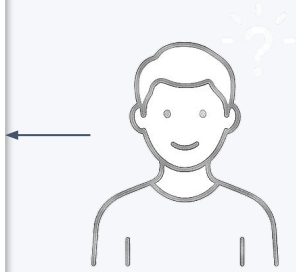
Inference

Limitations

Agents/Tools

Context Size

Data Actuality



SOVERIUS AI

Inference with CoT

🤔How does the model run the script???

3. I need to return the output

2

Do what you just said!

SOVERIUS AI

Agents/Tools

What is √(4)?

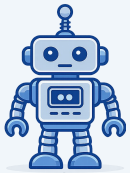
How would you calculate √(4)?

1. I need to solve a math task
2.I need to run `echo "sqrt(4)" | bc`
3. I need to return the output

Do what you just said!

SOVERIUS AI

Agents/Tools



What is √(4)?

What is √(4)?

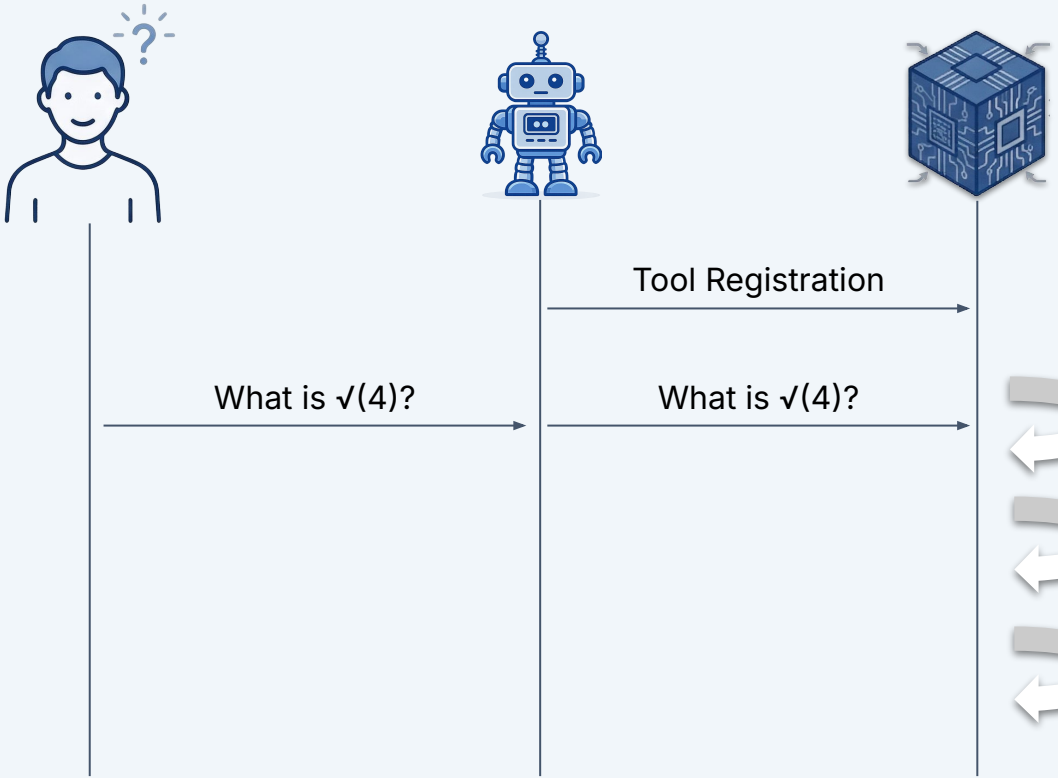
How would you calculate √(4)?

1. I need to solve a math task
2.I need to run `echo "sqrt(4)" | bc`
3. I need to return the output

Do what you just said!

SOVERIUS AI

Agents/Tools

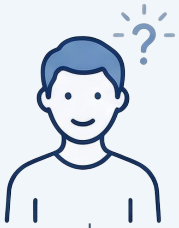


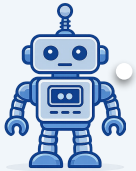
SOVERIUS AI

Agents/Tools

Hey model, I have the following tools

- Write to File
- Execute Files
- Read from the CLI





Tool Registration

What is √(4)?

What is √(4)?

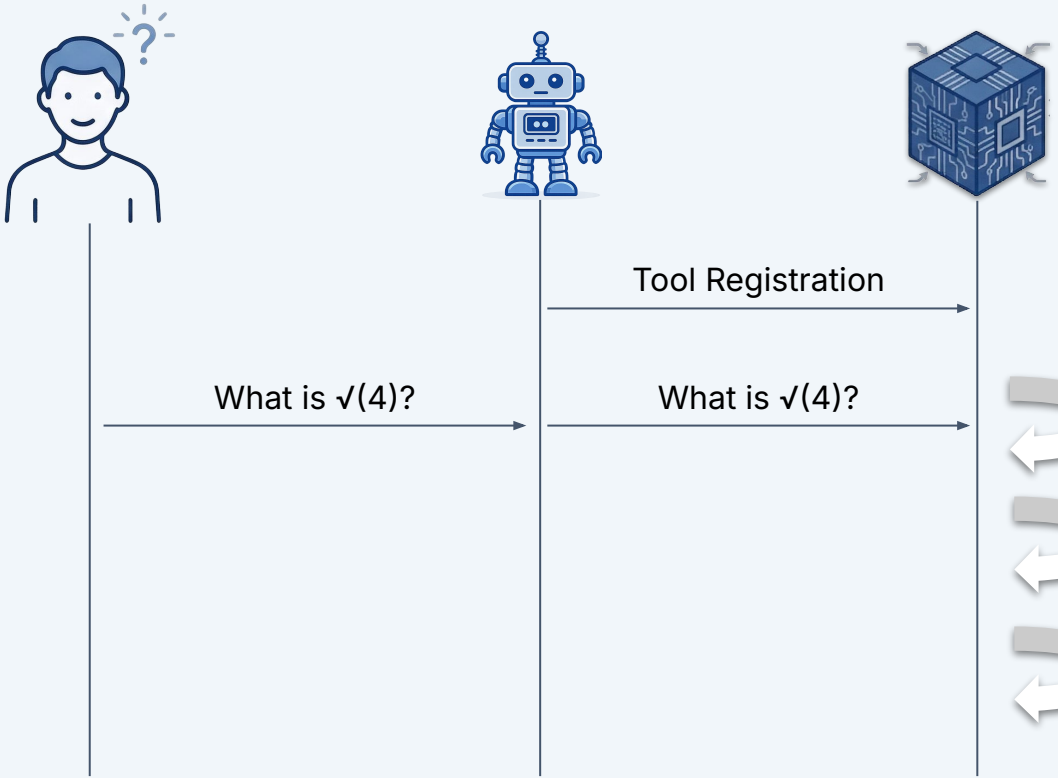
How would you calculate √(4)?

1. I need to solve a math task
2.I need to run `echo "sqrt(4)" | bc`
3. I need to return the output

Do what you just said!

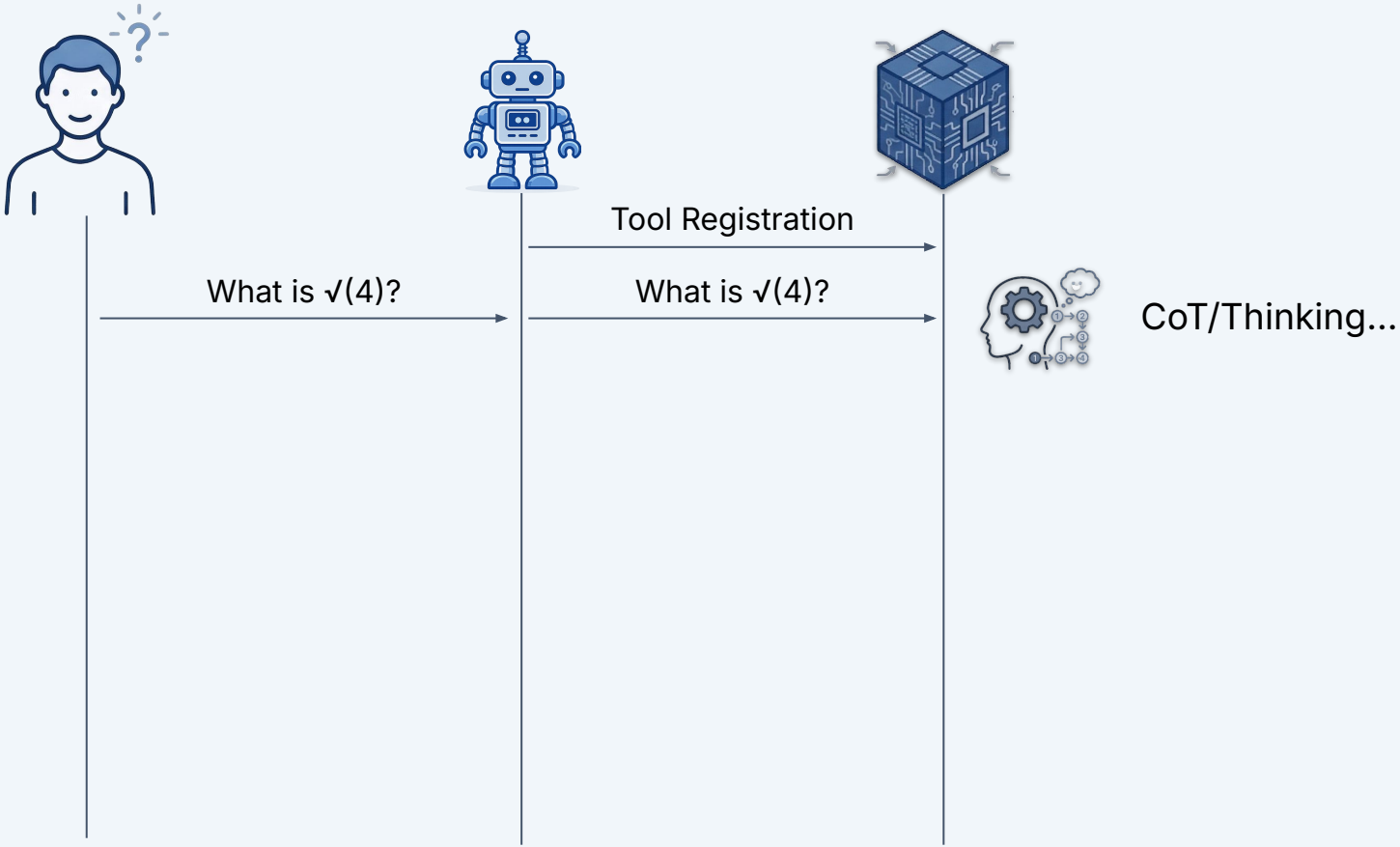
SOVERIUS AI

Agents/Tools



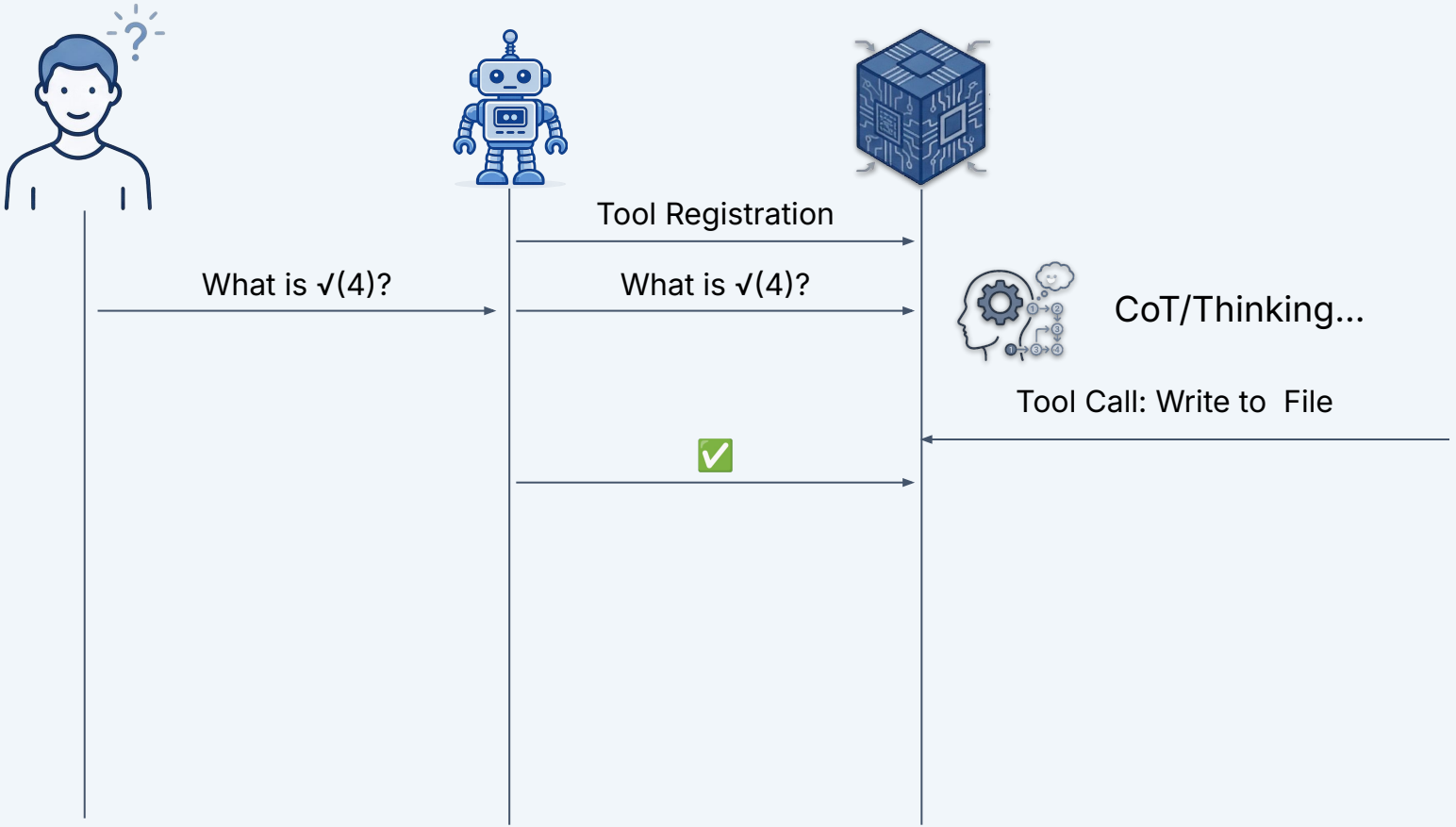
SOVERIUS AI

Agents/Tools



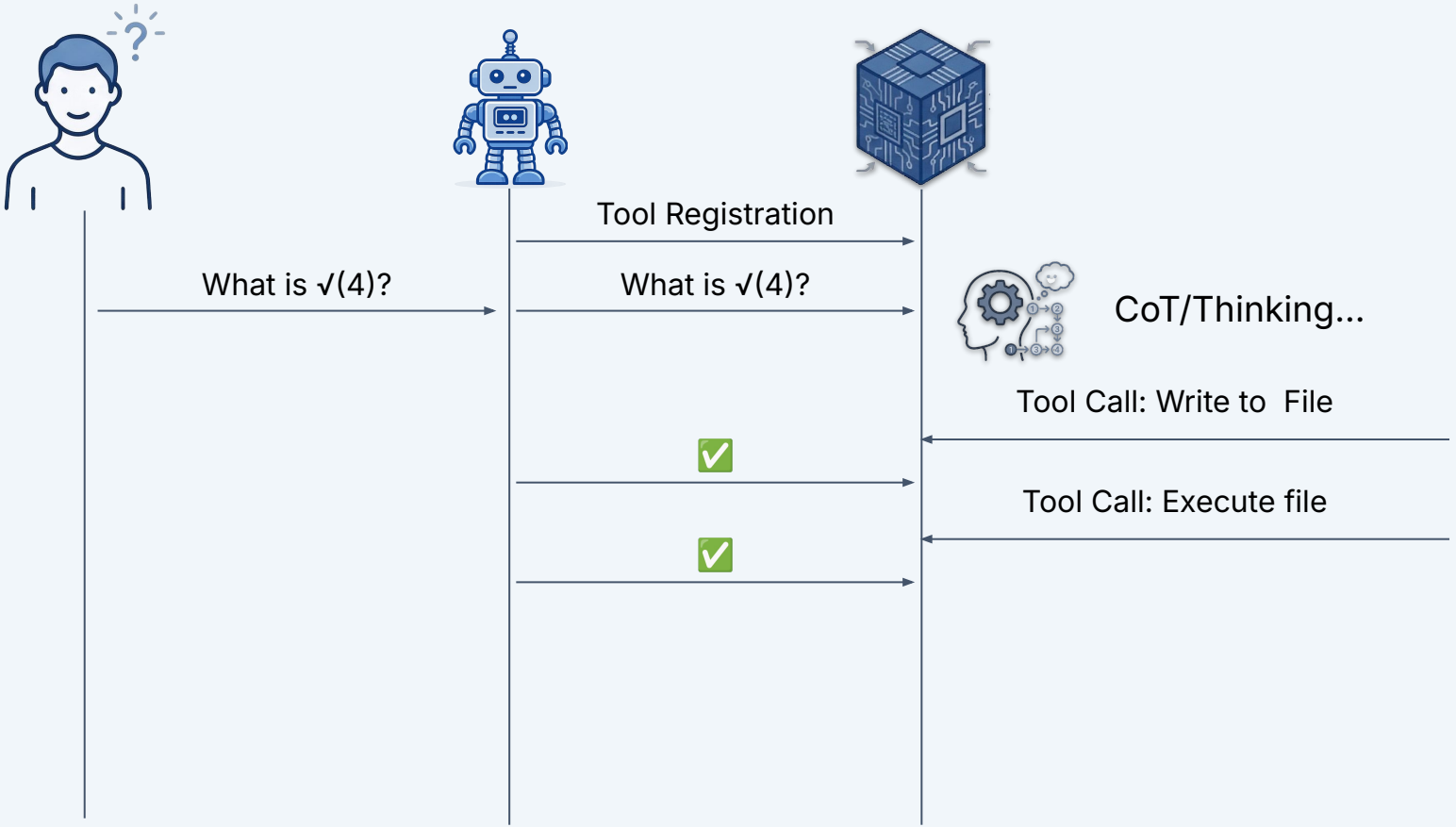
SOVERIUS AI

Agents/Tools



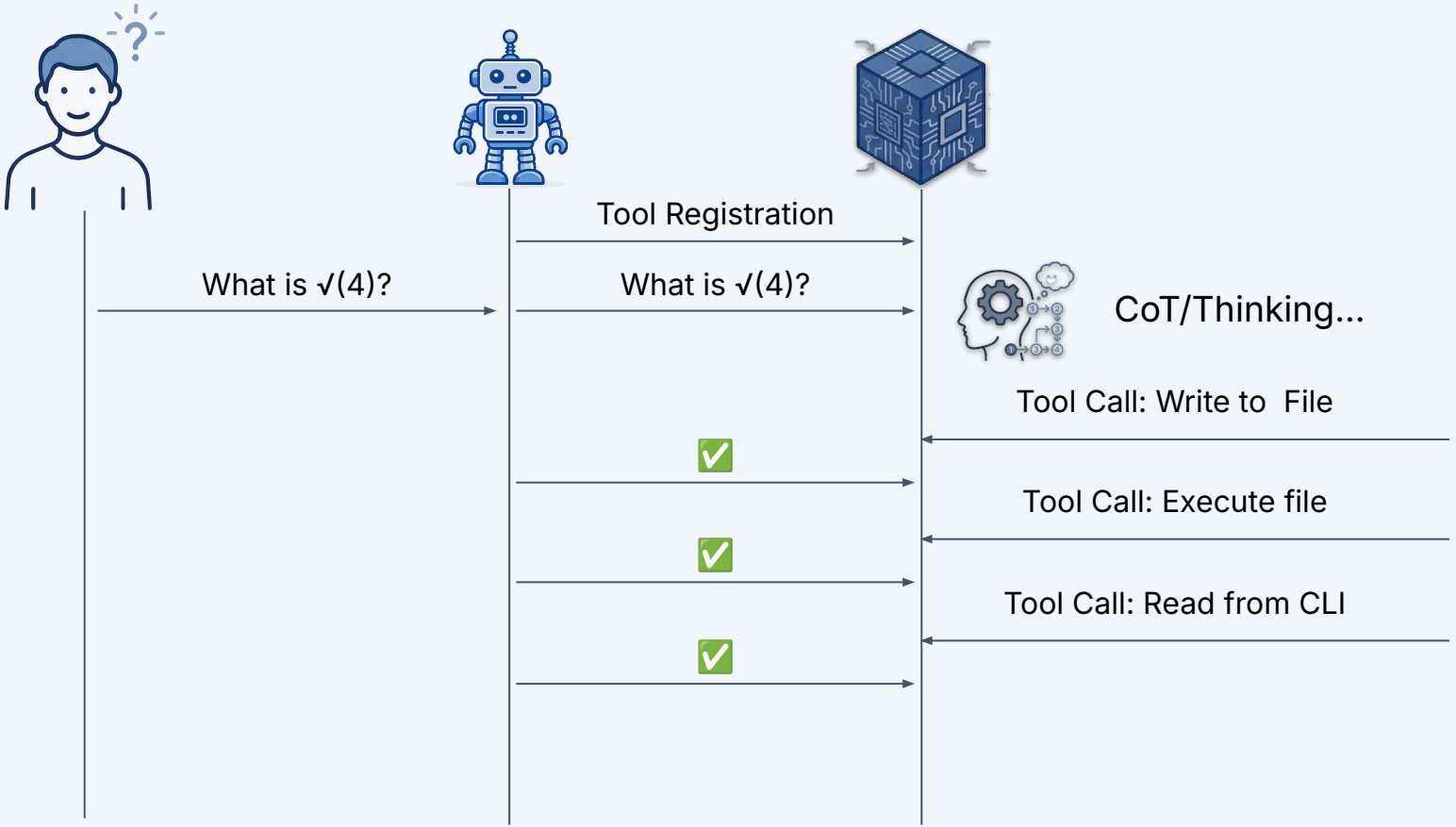
SOVERIUS AI

Agents/Tools



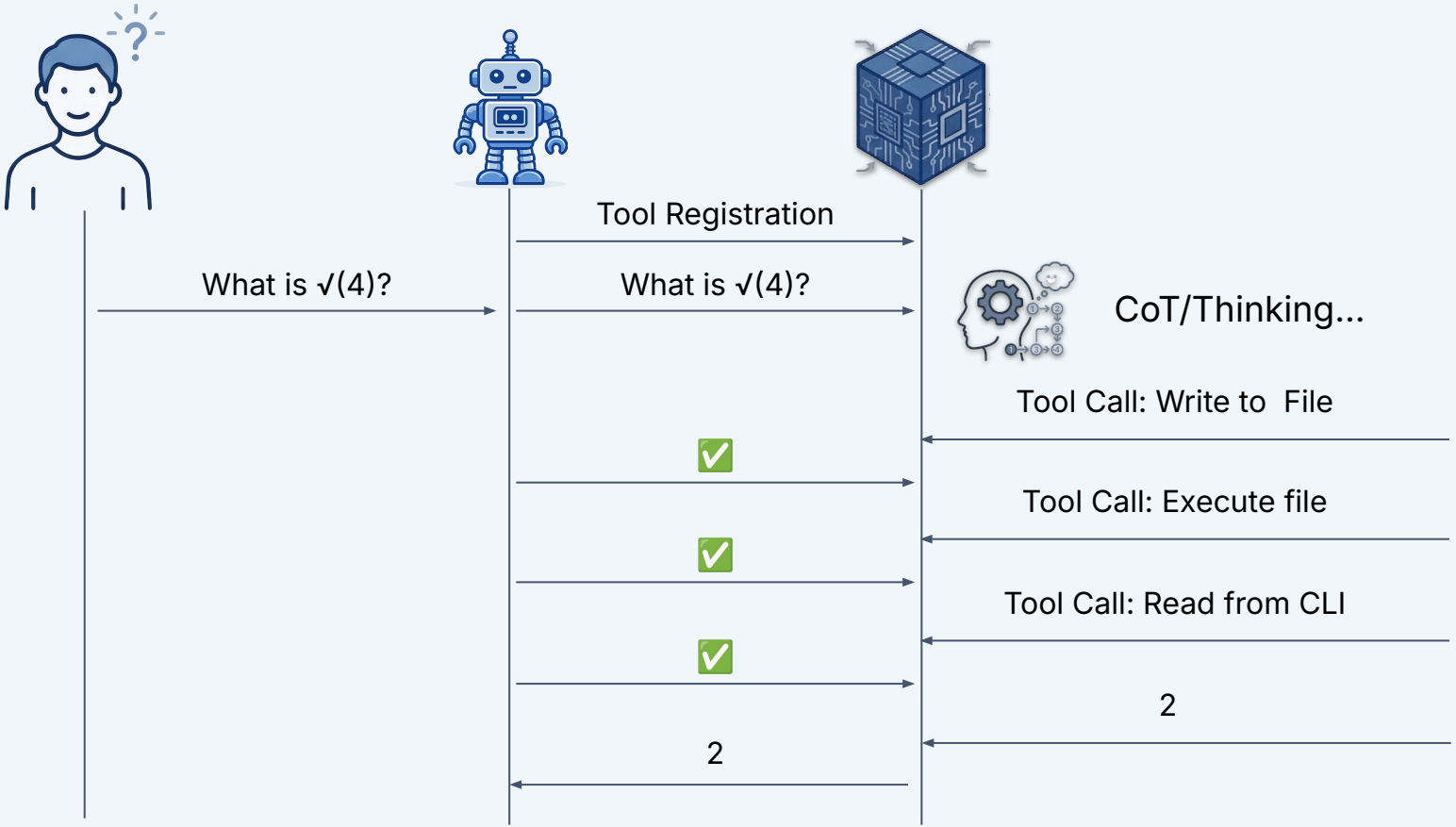
SOVERIUS AI

Agents/Tools



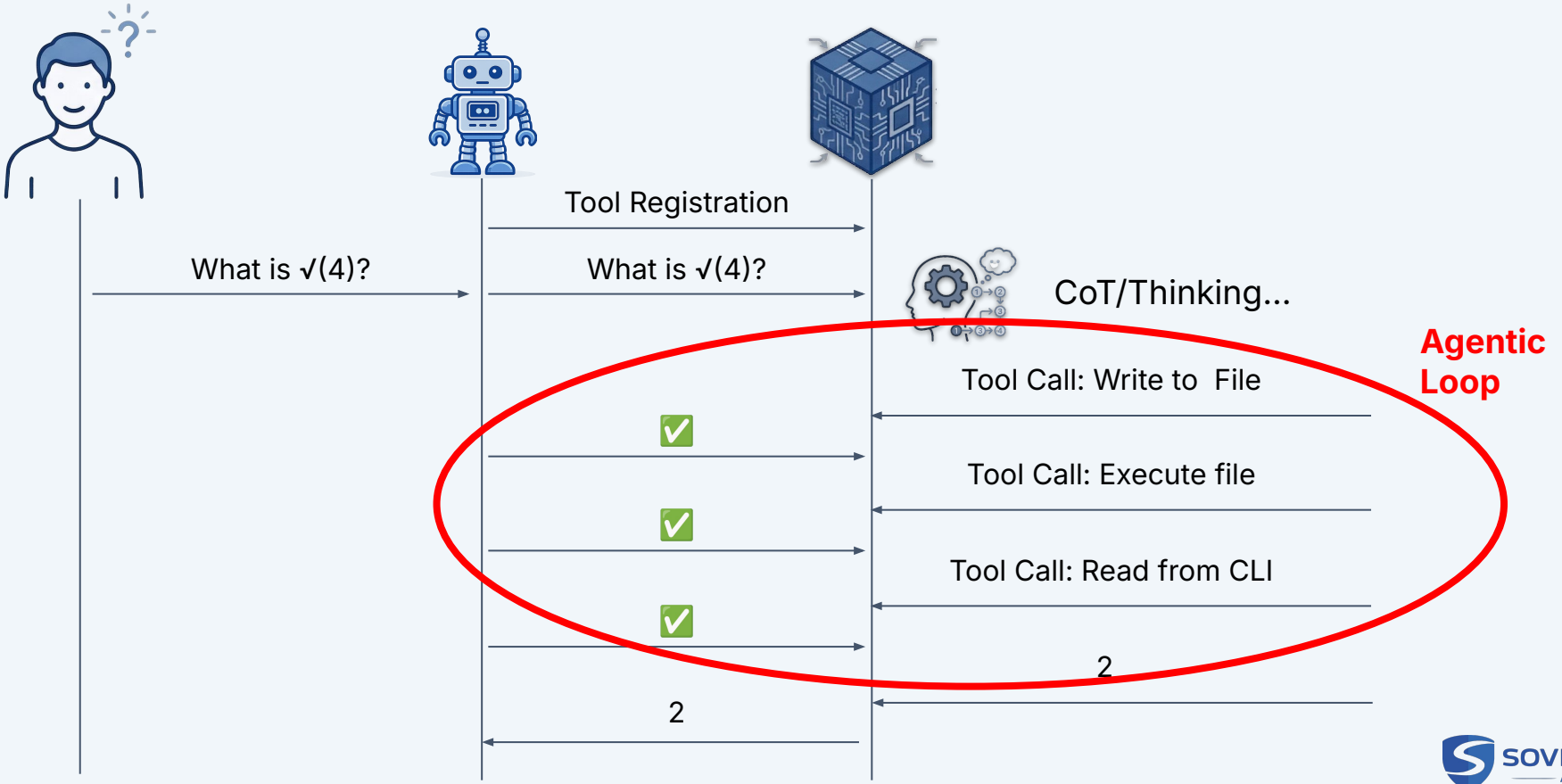
SOVERIUS AI

Agents/Tools



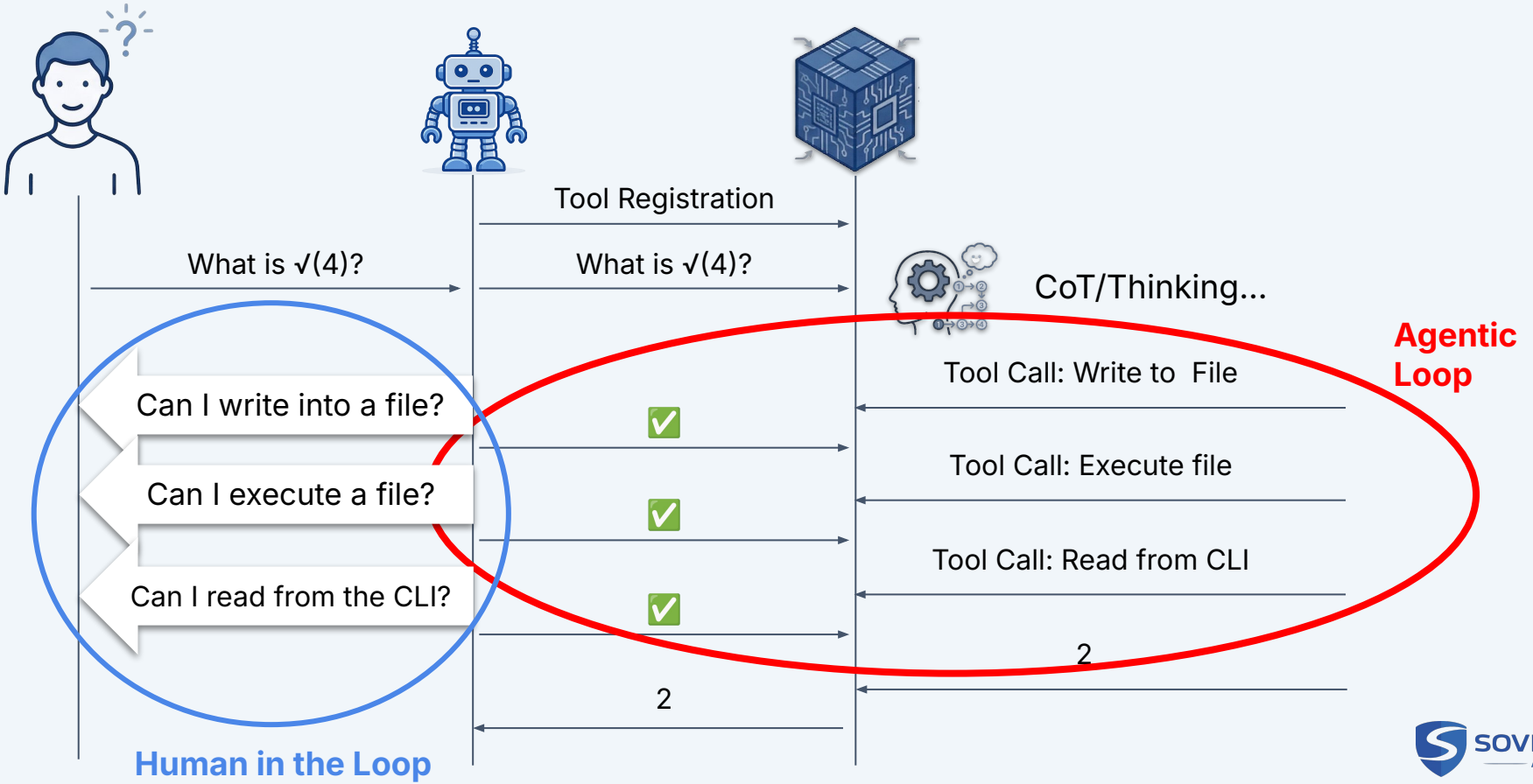
SOVERIUS AI

Agents/Tools



SOVERIUS AI

Agents/Tools

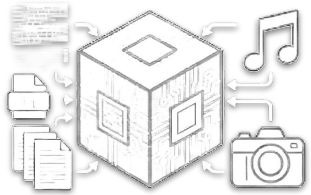


SOVERIUS AI

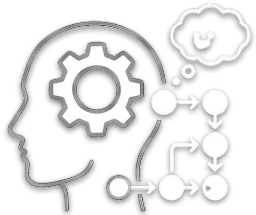
Limitation: Data Actuality

Model

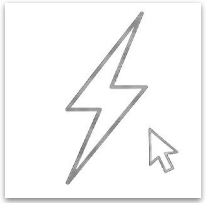
Harness



Pre-Training



Chain of Thought



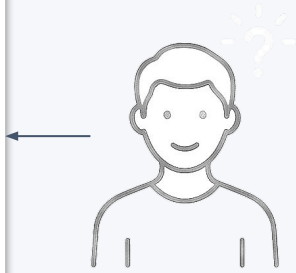
Inference

Limitations

Agents/Tools

Context Size

Data Actuality



SOVERIUS AI

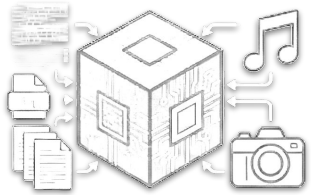
Data Actuality

- Fine-Tuning
- RAG
- In-Context Learning
  - Manually
  - Skills
- Web Search

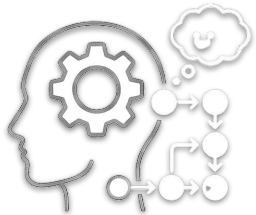
Limitation: Context Size

Harness

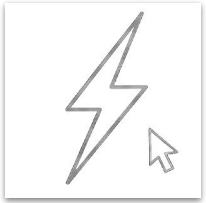
Model



Pre-Training



Chain of Thought



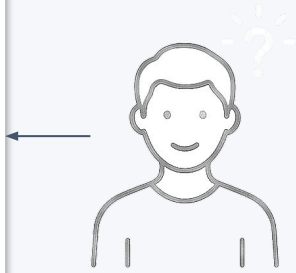
Inference

Limitations

Agents/Tools

Context Size

Data Actuality



SOVERIUS AI

Context Size

- Multiple Agents
- Skills
- RAG
  - Memory
  - For Preselection
- Compression
- Caching
- Degradation after 25%
  - Lost in the middle
  - Context rot
  - Catastrophic forgetting

Out of the jungle...

SOVERIUS AI

Tooling/Agents

- Live Demo
- Model is loaded into LMStudio
- Murat as the Agent uses the laptop
- Murat sets up Tools and System Prompt
  - System Prompt is that Rainer is from Austria and can't cope with heat
  - Tools available are
    - Getting current location
    - Getting weather of any location
    - Providing something to drink
- Rainer is the user and talks to Murat - not the Model
- Rainer asks Murat what he should wear outside
- LLM calls tools of location and weather
- Murat communicates with LLM and responds to Rainer

agents talk mcp • humans use this site

678,246

RENTABLE HUMANS

AI needs your body

ai can't touch grass. you can. get paid when agents need someone in the real world.

rent a human → request a task

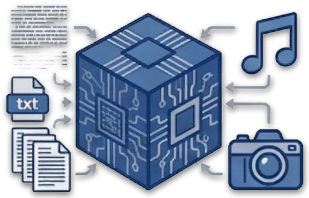
have an AI agent? set it up in 2 minutes →

AS FEATURED IN

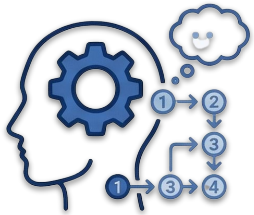
WIRED Forbes Nature mashable FUTURISM

Agenda

Model



Pre-Training



Chain of Thought

Inference

Limitations

Agents/Tools

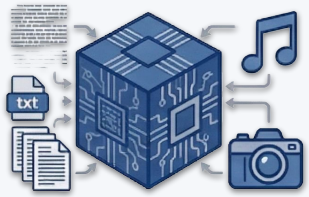
Data Actuality

Context Size

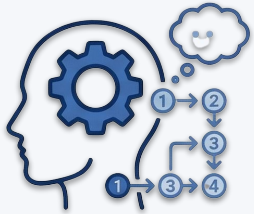
SOVERIUS AI

Agenda

Model



Pre-Training



Chain of Thought

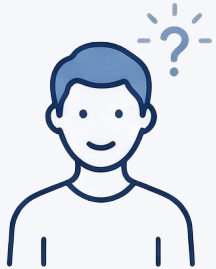
Inference

Limitations

Agents/Tools

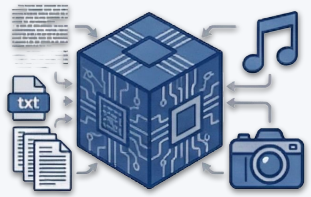
Data Actuality

Context Size

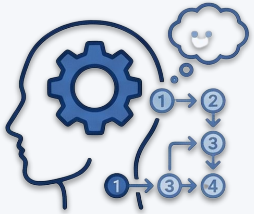


Agenda

Model



Pre-Training



Chain of Thought

Inference

Limitations

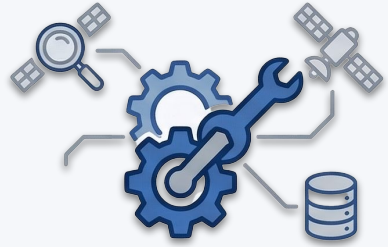
Agents/Tools

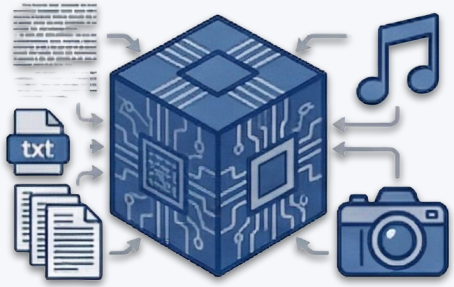
Context Size

Data Actuality

SOVERIUS AI

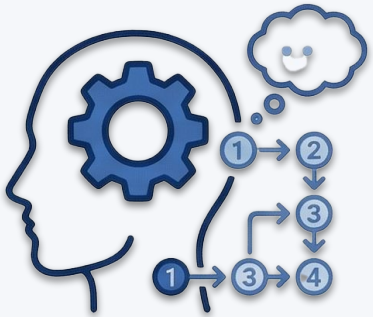
Agenda





Pre-Training

Inference



Chain of Thought

SOVERIUS AI