

Prompt Engineering

-Workshop-

Benjamin Aunkofer
Dr. Alexander Lammers

AI Convention 2025



IHK Schwaben

echo \$(whoami)



DATANOMIQ
Independent Data Solutions

AUDAVIS

Deloitte.



DATANOMIQ
Independent Data Solutions

ctcon

groupm

tu technische universität dortmund

FOM
Hochschule
für Oekonomie & Management
University of Applied Sciences

Value Engineering

We drive Value
with Data & AI

- 1  **Founded 2015** in Berlin
- 2  **100+ successful** data projects
- 3  **Full end-to-end service**
- 4  **50+ customers** in multiple industries: healthcare, e-commerce, automotive etc

From idea development to the full implementation

Services

1. Data & Process Mining

We design, build, and execute analytics systems focused on Business Intelligence, Process Mining, and Data Science/AI to help you uncover and optimize your business processes.

2. AI Product Development

We leverage our project and teaching experience to support you in developing AI-driven products, empowering your team to use data effectively and efficiently.

3. Workshop

Whether you are starting your first data project or want to optimize an existing portfolio, we guide you through the process and craft a tailored strategy together with you.



Data
Engineering (DWH
& Data Lake)



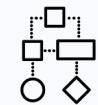
Data Science/AI



Data & AI
Strategy



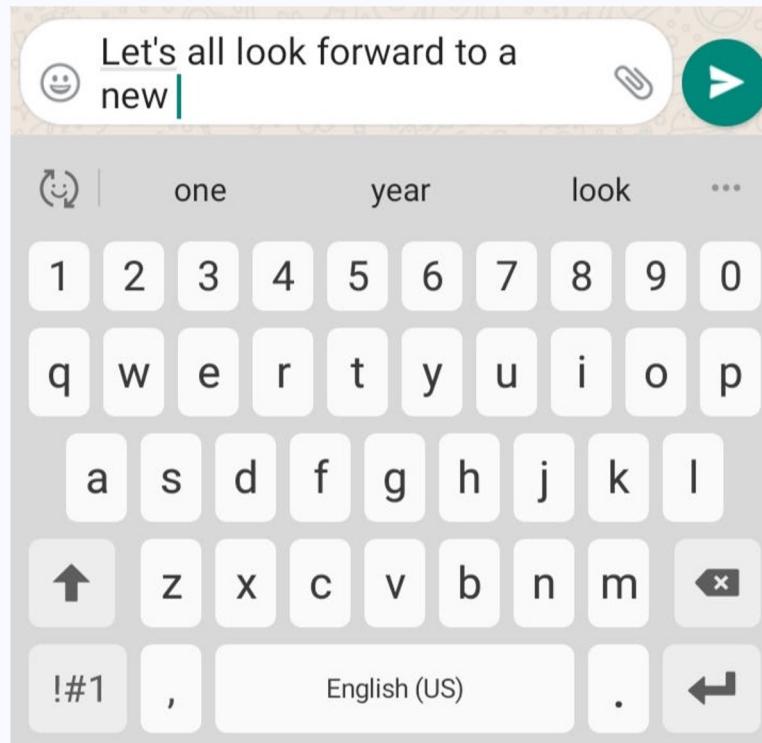
Data
Analytics



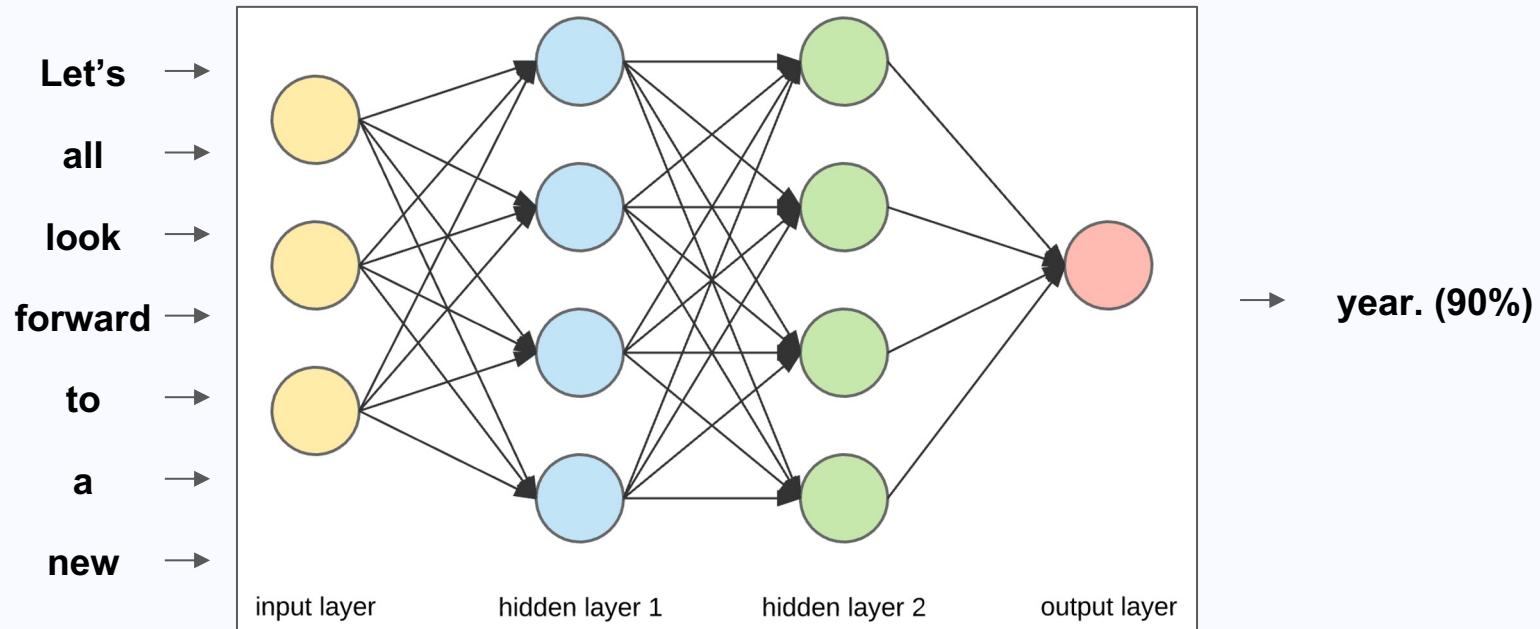
Process Mining

What is Generative AI (GenAI)?

What's the next word?

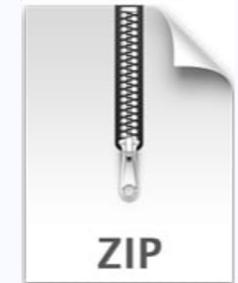


Under the Hood



Under the Hood

Facebook Llama 3 70B



Chunk of the Internet
~10TB of text data

25.000 GPUs H100
~14 days of training

"Compressed Internet"
~140 GB file

In layman's term



WIKIPEDIA

The Free Encyclopedia

Main page

Contents

Featured content

Current events

Random article

Donate to Wikipedia

Wikipedia store

Interaction

Help

About Wikipedia

Community portal

Recent changes

Contact page

Tools

What links here

Related changes

Upload file

Special page

Permanent link

Page information

Wikidata item

Print/export

Create a book

Download as PDF

Printable version

In other projects

Wikimedia Commons

MediaWiki

Meta-Wiki

Wikispecies

Welcome to Wikipedia,

the free encyclopedia that anyone can edit.
5,432,562 articles in English.

From today's featured article



Tropical Storm Bill hit the Gulf Coast of the United States in the summer of 2005. The second storm of that Atlantic hurricane season, Bill developed from a tropical wave on June 29 to the north of the Yucatan Peninsula. It slowly organized as it moved northward, and reached a peak of 60 mph (95 km/h) shortly before making landfall in south-central Texas. The extratropical state, a levee was breached, flooding many homes, and in French, two swimmers drowned. As Bill accelerated to the northeast, moisture from the storm, combined with cold air from an approaching cold front, produced an outbreak of 34 tornados. Moderate winds and wet soil combined to topple trees onto houses and power lines, leaving hundreds of thousands without electric power. By the time Bill became extratropical on July 2, it was responsible for four deaths and around \$50 million in damage. ([Full article...](#))

Part of the [2005 Atlantic hurricane season](#), one of Wikipedia's featured topics.

Recently featured: White-naped xenopsaris - [The Well of Loneliness](#) - Japanese aircraft carrier [Juryō](#)

[Archive](#) • [By email](#) • [More featured articles...](#)

Did you know...

- that larvae of the [red-headed pine sawfly](#) (pictured) are gregarious, and when they have defoliated a tree, move as a group to a neighbouring one?
- ... that American trade unionist [George Baldanzi](#) worked out of [Jaime Mansfield](#)

In the news

- In sailing, the [America's Cup](#) concludes with Team New Zealand defeating Oracle Team USA.
- An oil truck explosion in Ahmedpur East, Pakistan, kills at least 150 people and injures more than 100 others.
- The World Health Organization estimates that 200,000 people in Yemen have been infected with cholera.
- At least 10 people die and more than 90 others are missing following landslides in Sichuan, China.
- In Pakistan, at least 80 people are killed and more than 200 others are injured in multiple bombings in Quetta and Parachinar.
- The cricket teams of Ireland and Afghanistan are awarded Test status by the ICC.

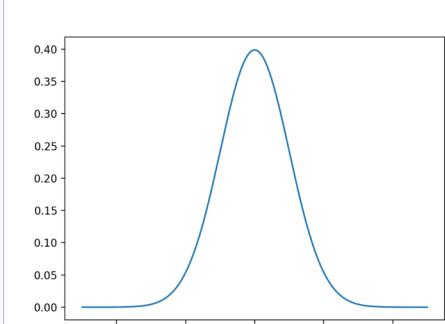
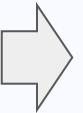
Ongoing

- Recent deaths:** Michael Eason • Gordon Wilson • Maria Mutagamba • Gabe Pressman

On this day...

June 29

- 1444 – In their rebellion against the Ottoman Empire, Albanians led by Skanderbeg routed the Ottoman forces in the [Battle of Torvill](#).
- 1536 – The English Queen [Jane Seymour](#) turned to the ground after a cannon employed for special effects misfired during a performance of William Shakespeare's [Henry VIII](#) and ignited the



Some Text

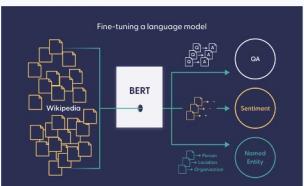
What is ChatGPT?



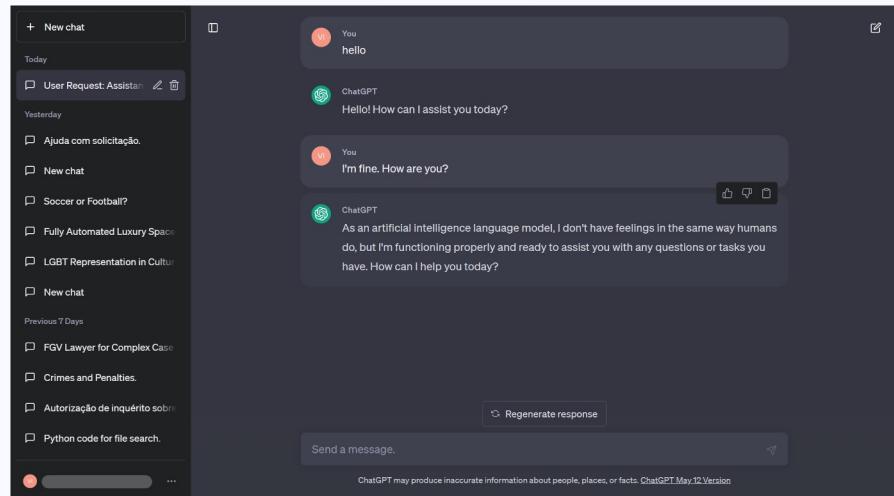
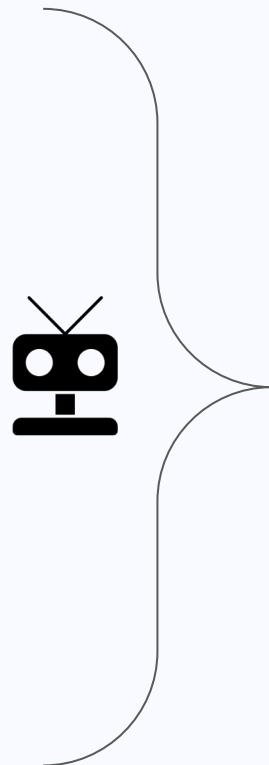
Vision model



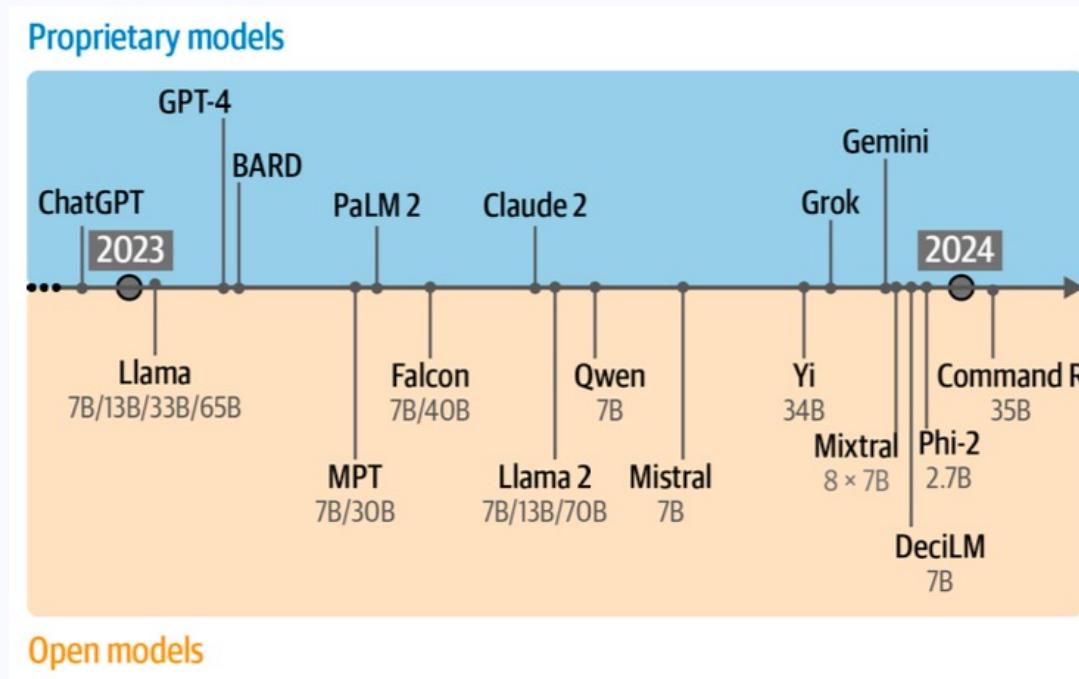
Voice model



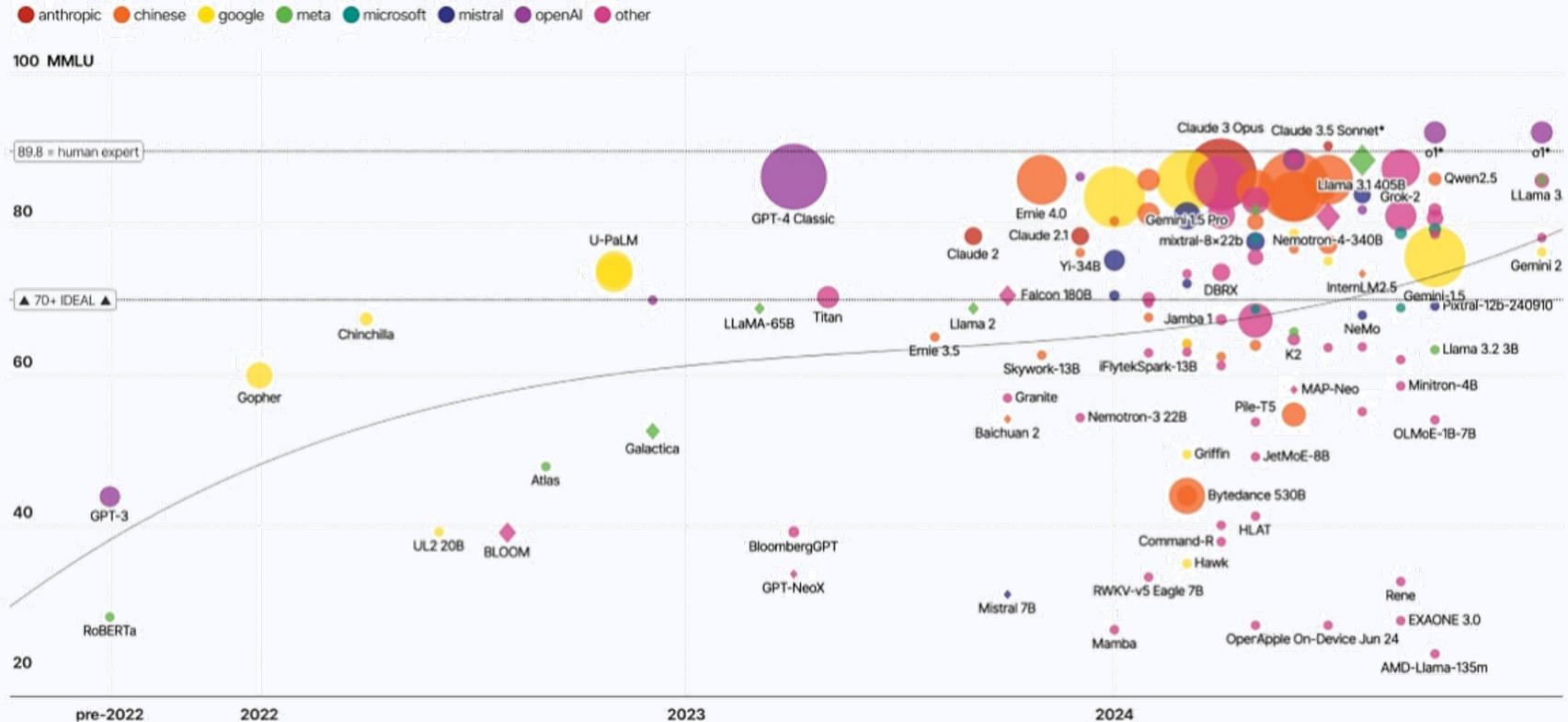
Text model



Open and proprietary models



Source: <https://huggingface.co/meta-llama/Meta-Llama-3-70B>

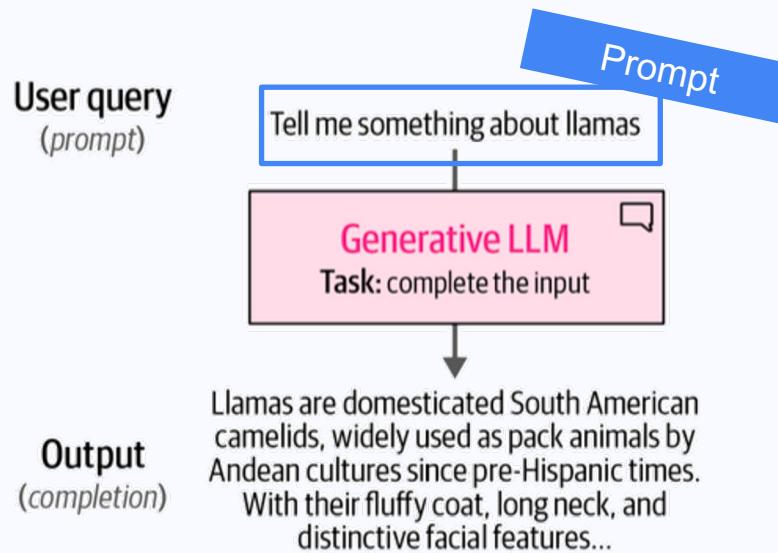


What is Prompt Engineering?

Prompt Engineering

- Relatively new discipline: for developing and optimizing prompts to effectively use large language models
- Researchers use it to improve LLMs and developer use it to make their applications robust
- It's more "**art**" than "**science**", try and error approaches

How does it work?



A well-formulated prompt is like a well-crafted brief: it guides your thinking in the desired direction and defines the framework for the solution.

66

Prompt as a navigation aid in semantic space:

LLMs navigate through a high-dimensional vector space in which concepts and their relationships are mapped. Your prompt acts like a coordinate system that determines in which area of this space the answer is generated.

Prompt engineering as a precision instrument:

The quality of an AI response is largely determined by the information density and structure of your query. Like a navigation system, the LLM needs clear instructions and context in order to find the right path through the complex knowledge space.

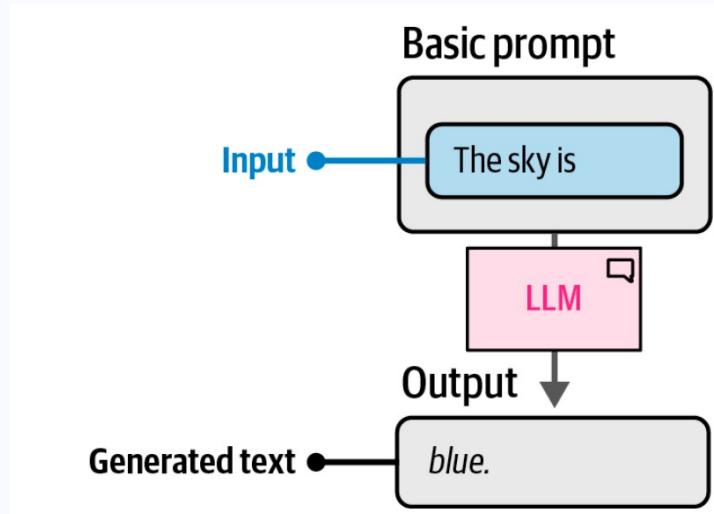
For what are prompts used for?

- Text Classification
- Conversation
- Code Generation
- Reasoning
- Agentic systems

There are a few techniques to master

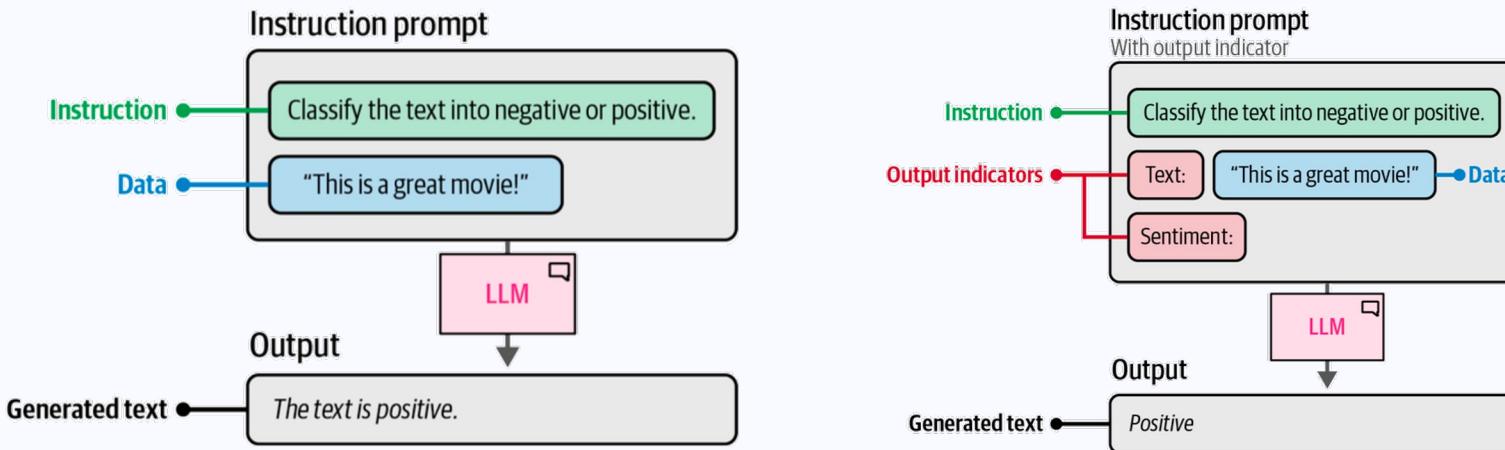
1. Basic Prompting
2. Instruction-based Prompting
3. Zero-shot prompting
4. One-shot-prompting
5. Few-shot prompting
6. Chain-of-Thought prompting
7. Tree of Thought prompting
8. Reasoning and Acting (ReAct)

1. Basic Prompting



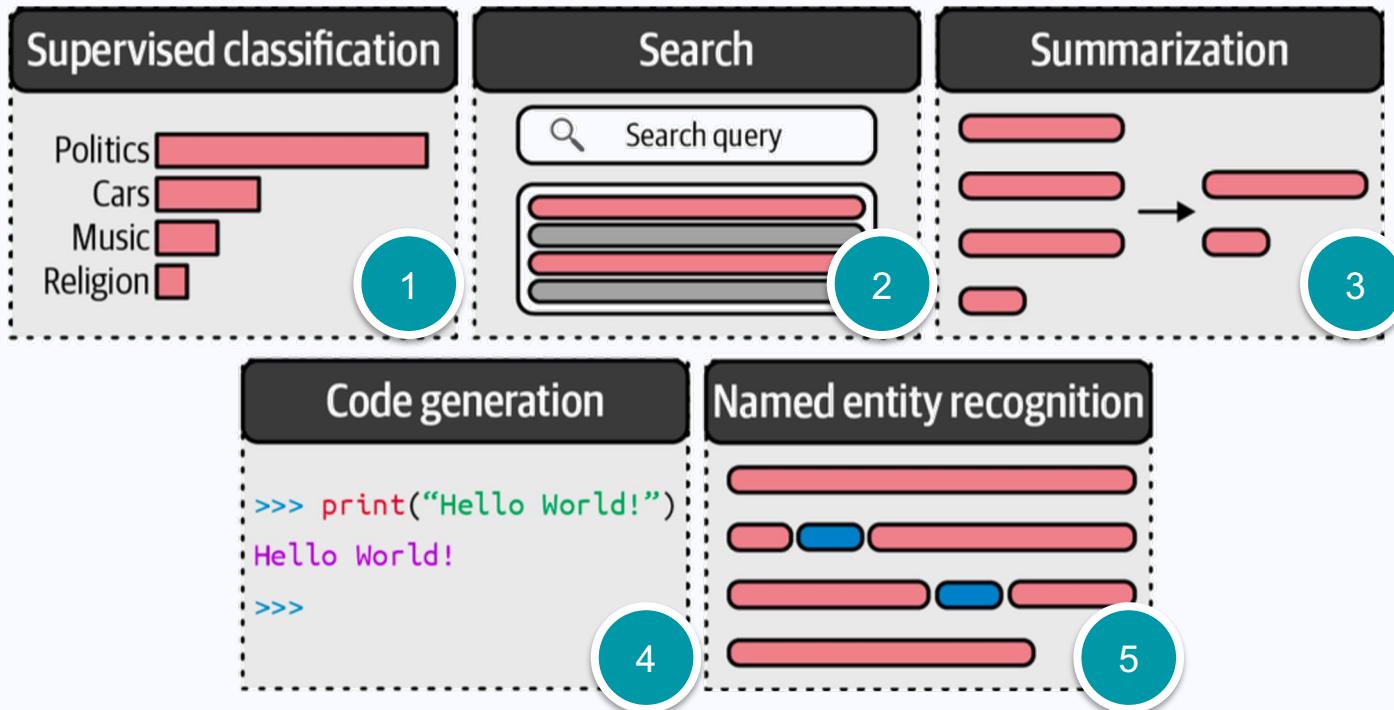
- Simplest prompt; LLM will simply try to complete the sentence
- Conversation until desired result

2. Instruction Based Prompting

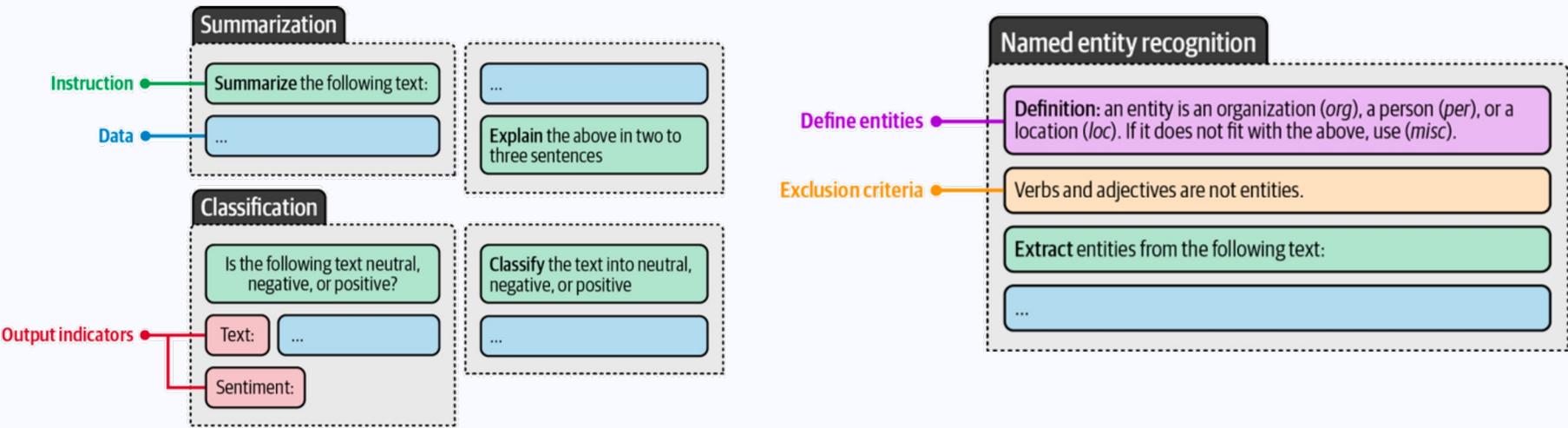


- Application for a **specific question or task**
- With or without **„output indicators“**

2. Instruction Based Prompting



2. Instruction Based Prompting

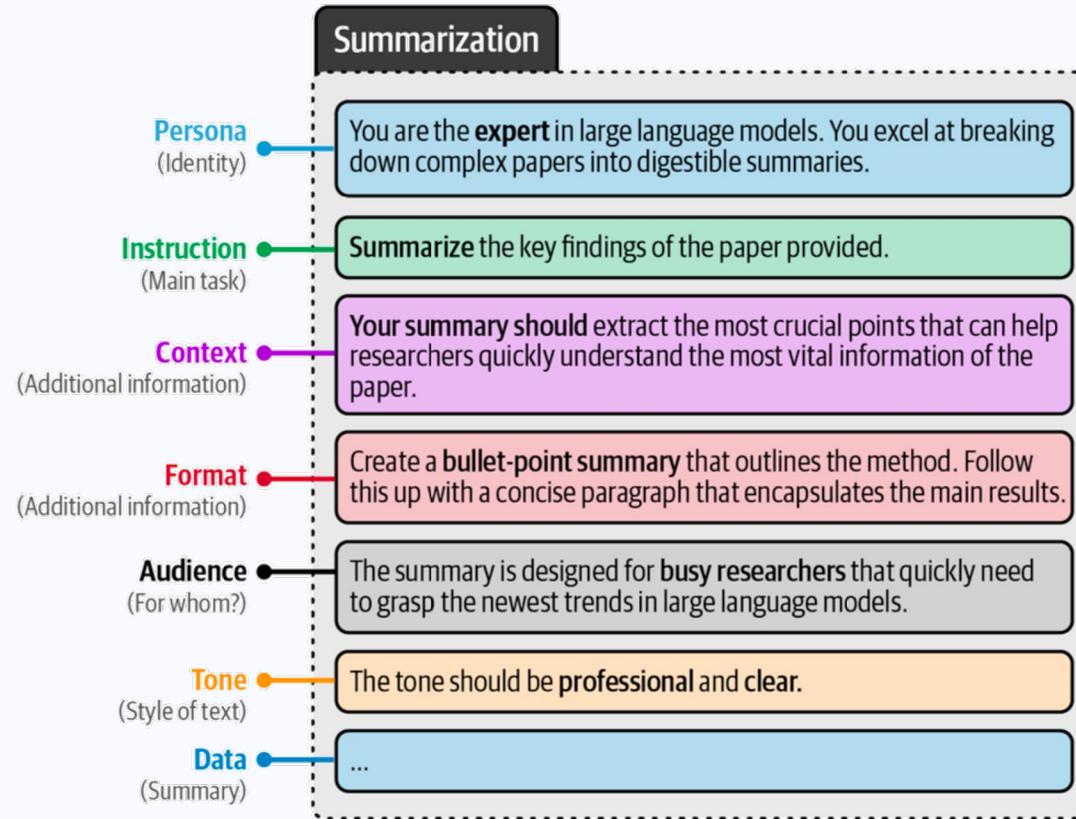




What is a good prompt?

- A good prompt is based on several factors
- First, the instructions (What is to be done / achieved?)
- What data should be used to achieve this?
- Use of output indicators
- **But there is more...**

What defines a good prompt?



Zero-shot Prompting

Zero-shot prompt

Prompting without examples

Classify the text into neutral, negative, or positive.

Text: I think the food was okay.

Sentiment: ...

- Prompting **without examples**
- Just expecting the correct result

One-shot Prompting

One-shot prompt

Prompting with a single example

Classify the text into neutral, negative, or positive.

Text: I think the food was alright.

Sentiment: Neutral

Text: I think the food was okay.

Sentiment:

- Prompting with just one example

Few-shot Prompting

Few-shot prompt

Prompting with more than one example

Classify the text into neutral, negative, or positive.

Text: I think the food was alright.

Sentiment: **Neutral**.

Text: I think the food was great!

Sentiment: **Positive**.

Text: I think the food was horrible...

Sentiment: **Negative**.

Text: I think the food was okay.

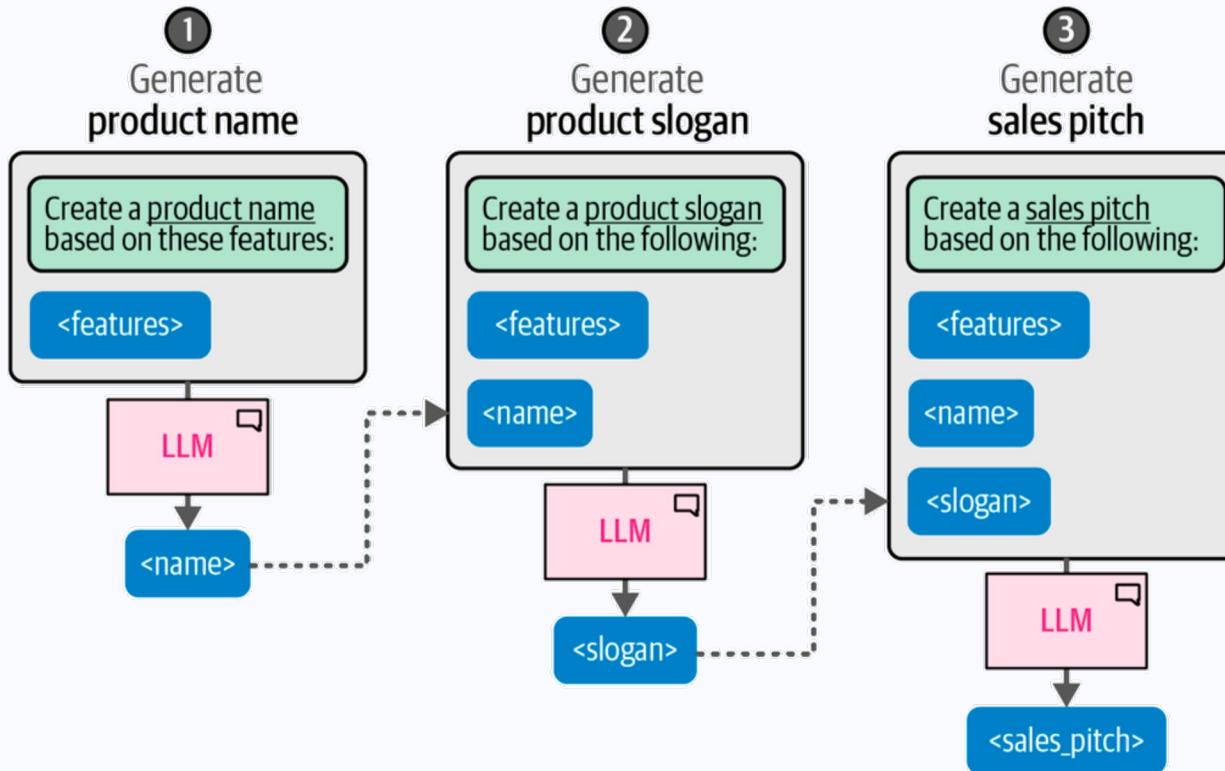
Sentiment:

- Prompting with **more than one example**

Chain-of-Thought Prompting

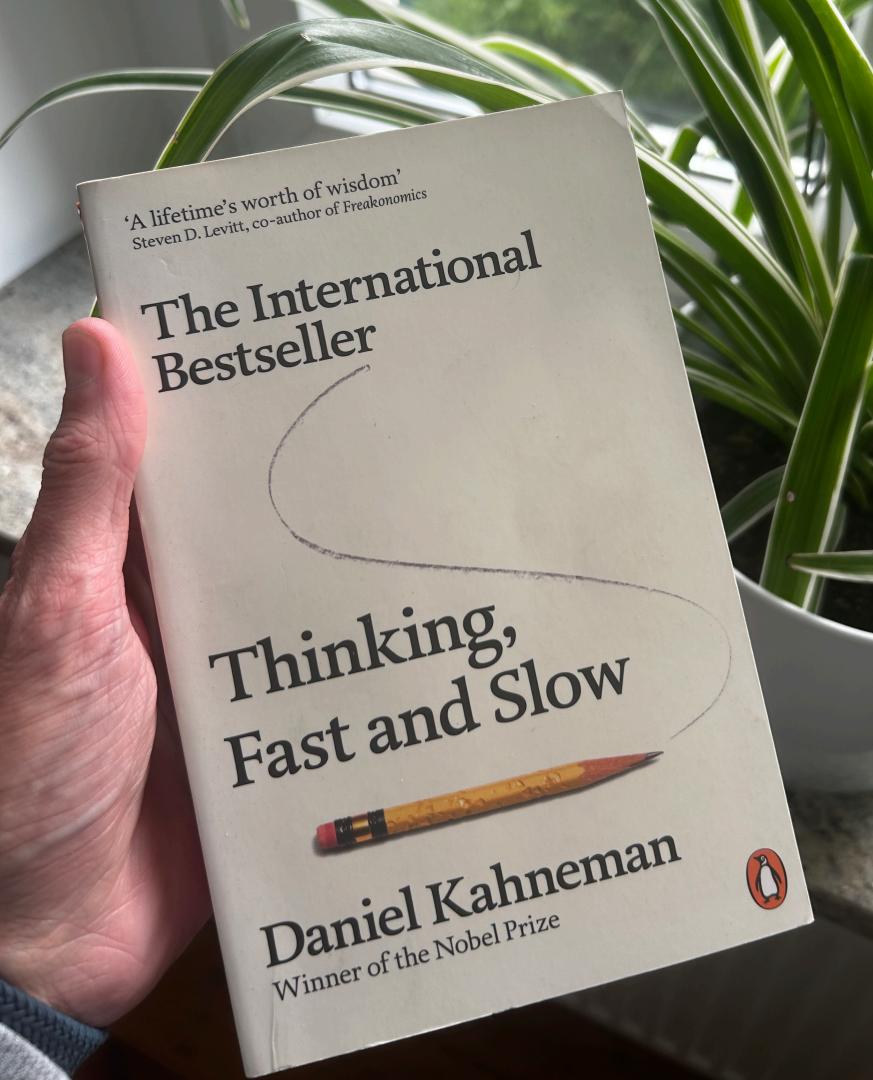
- Instead of solving the problem within a prompt, we can also **solve it between prompts**
- Basically, we **take the output of one prompt and use it as input for the next**, creating a continuous chain of interactions that solves our problem
- This creates a **continuous chain of interactions** that solves our problem

Chain-of-Thought Prompting



Reasoning

- As prompt engineering evolves beyond basic instruction and example-based techniques, a critical frontier emerges **enabling and eliciting reasoning within large language models (LLMs)**. This transition is essential for tackling complex; multi-step tasks where straightforward prompting often falls short. **The focus shifts from simply generating correct answers to guiding models through logical, transparent, and interpretable thought processes.**



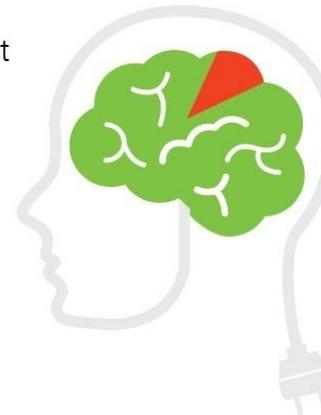
Reasoning

SYSTEM 1

Intuition & instinct

95%

Unconscious
Fast
Associative
Automatic pilot



SYSTEM 2

Rational thinking

5%

Takes effort
Slow
Logical
Lazy
Indecisive

Reasoning

- How can we get LLMs to use System 2 thinking?
- System 1 is fast, intuitive, and automatic-like how generative models produce tokens without self-reflection.
- System 2 is slower, conscious, and logical, involving brainstorming and self-reflection.
If we could teach generative models to imitate self-reflection, **they would produce more thoughtful, System 2-like responses than with System 1 thinking.**

Reasoning – Zero-shot chain-of-thought

Zero-shot chain-of-thought

Prompting without example

Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

● **Instruction**

Let's think step-by-step.

● **Prime reasoning**

A: The cafeteria had 23 apples originally. They used 20 to make lunch. So they had $23 - 20 = 3$. They bought 6 more apples, so they have $3 + 6 = 9$.

● **Reasoning process**
(thought)

The answer is 9. ✓

Few-shot chain-of-thought

- **Prompt:**

For example, a train takes 2 hours to travel 60km. What is its speed? Answer: $60 \div 2 = 30$ km/h.

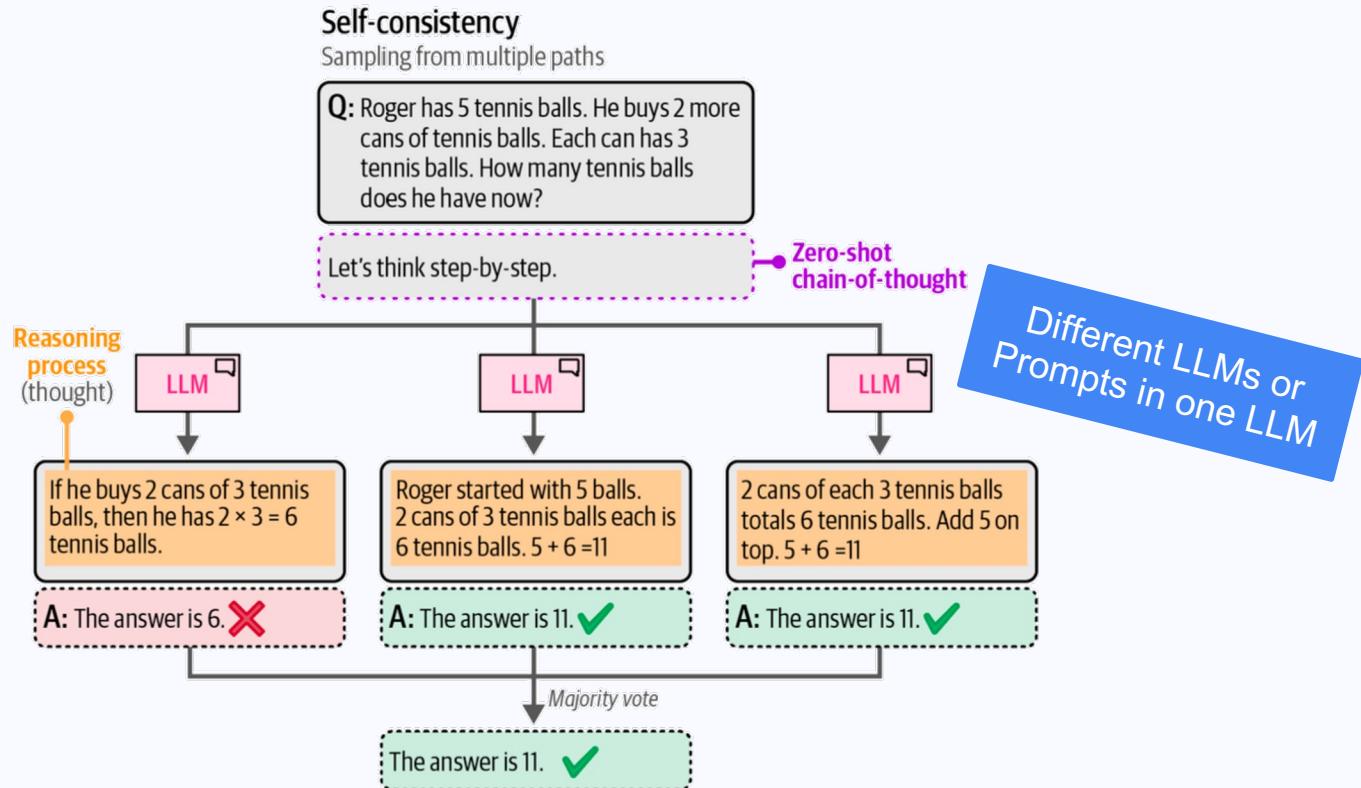
- *Now solve this: If a car takes 3 hours to travel 150km then what is its speed.*

- **Response:**

$150 \div 3 = 50$ km/h.

- This approach is best for the tasks that require context learning and improved response consistency.

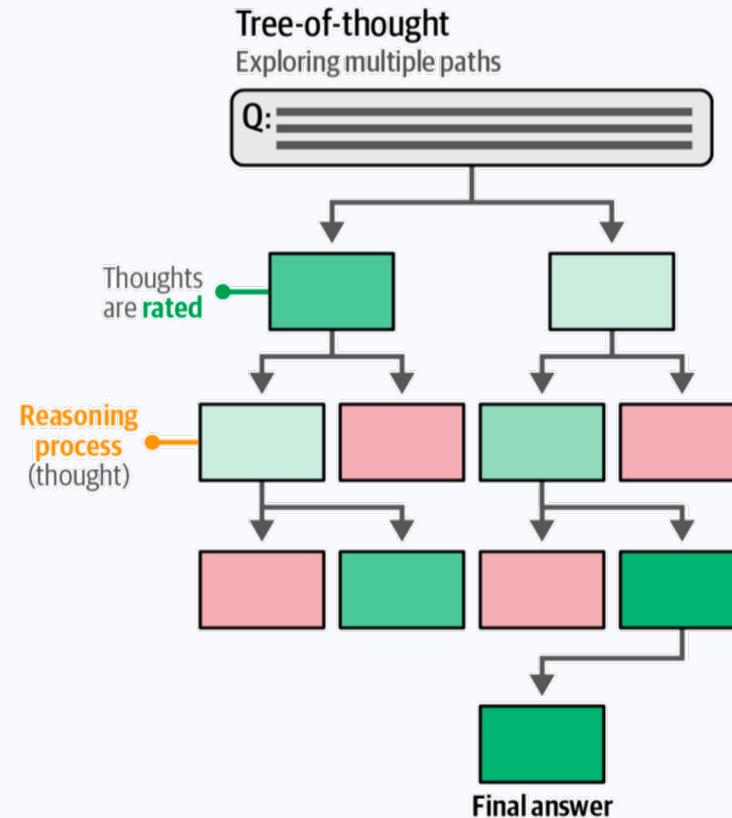
Self-Consistency Chain-of-thought



Tree-of-thought

- Chain-of-thought and self-consistency aim to enable more complex reasoning
- By choosing and refining multiple "thoughts," we improve generative model results
- Tree-of-thought enhances this by allowing deeper exploration of ideas
- **The method breaks down problems into steps, prompting the model to consider different solutions at each stage**
- It then selects the best solution

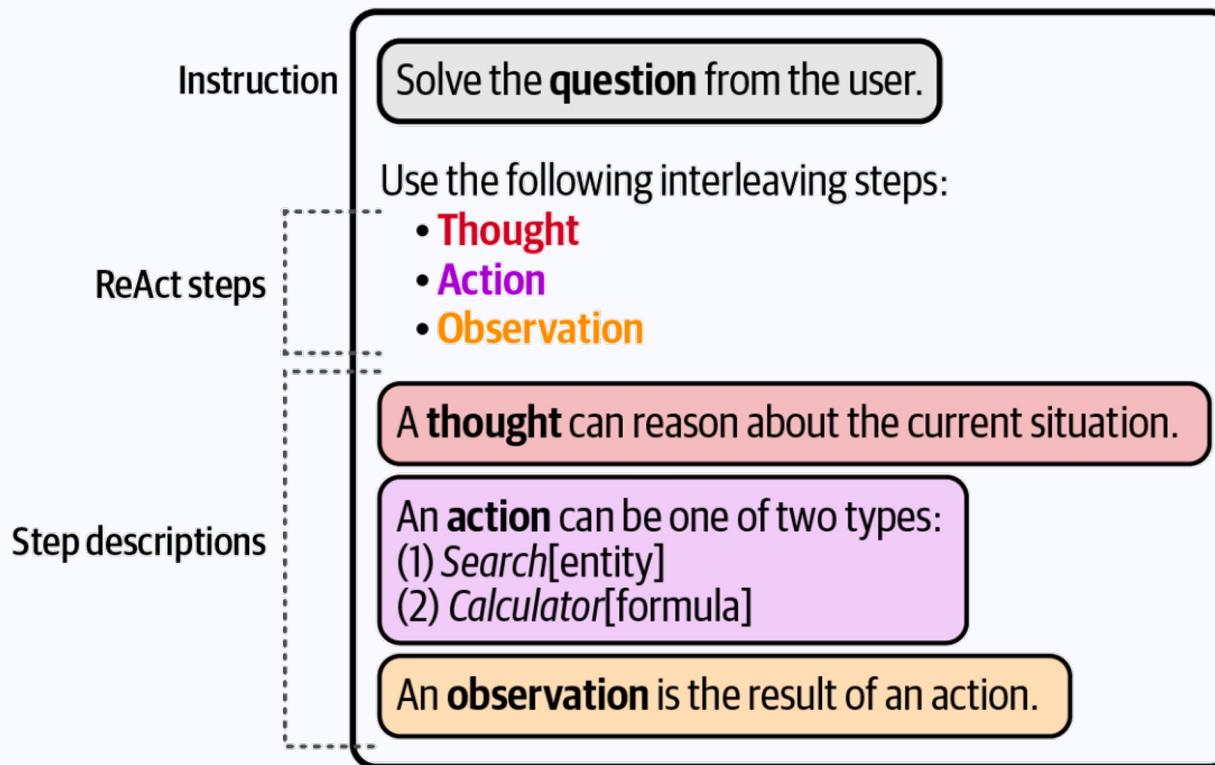
Tree-of-thought



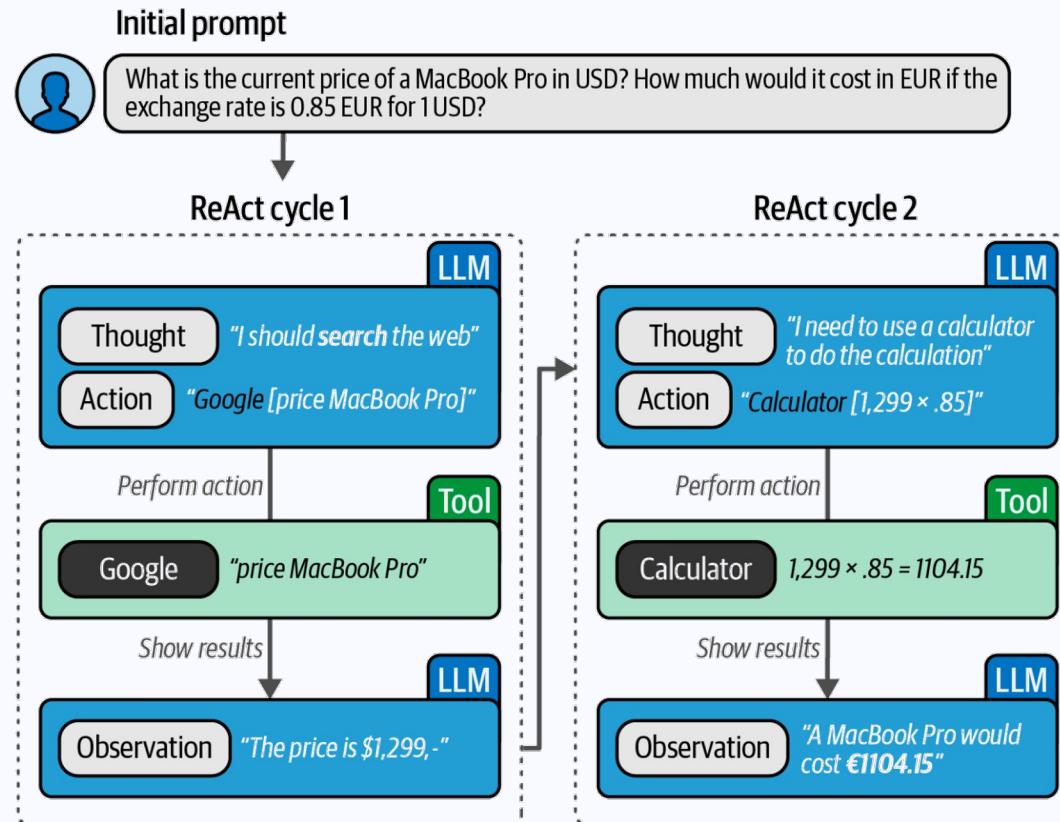
ReAct (Reason & Act)

- ReAct is a powerful framework that combines two important concepts in behavior: **reasoning and acting**.
- Acting is a bit of a different story. LLMs are not able to act. To give them the ability to act, we could tell an LLM that it can use certain tools, like a weather forecasting API. However, since LLMs can only generate text, they would need to be instructed to use specific queries to trigger the forecasting API.
- **Thought, action, observation**

ReAct (Reason & Act)



ReAct (Reason & Act)



Best Practices

- **Be as specific as possible** e.g if you write code to print hello world (in which language etc)
- **Provide context:** few-shot prompting works well if you already provide some context on what you want e.g. sample text, code snippets, documentation etc
- **Avoid negative instructions** e.g. avoid “Don’t write too much detail but rather write “Provide a concise summary”
- **Split complex tasks into simpler ones** e.g. avoid write an entire word plugin that does this but rather say create a hello world example for the word plugin
- **Be aware of AI hallucinations**

Hallucinations

Hallucinations

what is the second longest bridge in Paris?



The second longest bridge in Paris is the Pont Mirabeau. It spans the Seine River and connects the 15th and 16th arrondissements of Paris. The bridge is notable for its elegant Art Nouveau design and was constructed between 1895 and 1897. It measures approximately 173 meters (568 feet) in length, making it the second longest bridge in the city after the Pont Aval.

what about the Pont de Bercy?

Hallucinations

DT

what is the second longest bridge in Paris?



The second longest bridge in Paris is the Pont de Bercy. It spans the Seine River and has a length of approximately 173 meters (568 feet). It's an important transportation link in Paris, connecting the 12th arrondissement on the right bank with the 13th arrondissement on the left bank. However, bridge lengths can sometimes vary based on what is included in the measurement (such as approach viaducts), so it's always good to double-check such facts from multiple sources.

Hallucinations

Google Pont de Bercy total length

All Images Maps News Videos Books Flights More Tools

Pont de Bercy > Total length :

175 m

Passerelle Simone-de-Beauvoir 304 m

Viaduc d'Austerlitz 140 m

Pont au Double 45 m

Pont de Tolbiac 168 m

W Wikipedia https://en.wikipedia.org/wiki/Pont_de_Bercy

Pont de Bercy

Pont de Bercy ; Stone and Reinforced Concrete · 175 m (574 ft) · 40 m (130 ft).

W Wikipedia https://de.wikipedia.org/wiki/P... · Translate this page

Pont de Bercy

Pont de Bercy ; Gesamtlänge, 175 m ; Breite, 40 m ; Baubeginn, 1863–1864, 1904, 1989–1991 ; Planer : E.-J. Féline-Romany, J. Monthoux, J. Rézal.
Missing: total length

The Happy Pontlist http://happyPontlist.blogspot.com/2017/01/french-br...
French Bridges: 18. Pont de Bercy, Paris
22 Jan 2017 — This was a three-span chain suspension bridge designed by Bayard de la Vingtrie and Fortune de Vergès, who have also been credited with the 1828 ...

Pont de Bercy

Website Directions Save

4,3 ★★★★★ 542 Google reviews

Bridge in Paris, France

The Pont de Bercy is a bridge over the Seine in Paris. It links the 12th and 13th arrondissement of Paris by extending the Boulevard de Bercy and Boulevard Vincent-Auriol. In addition to the roadway, the bridge also carries Line 6 of the Paris Métro, between Quai de la Gare station and Bercy station. [Wikipedia](#)

Data is Key

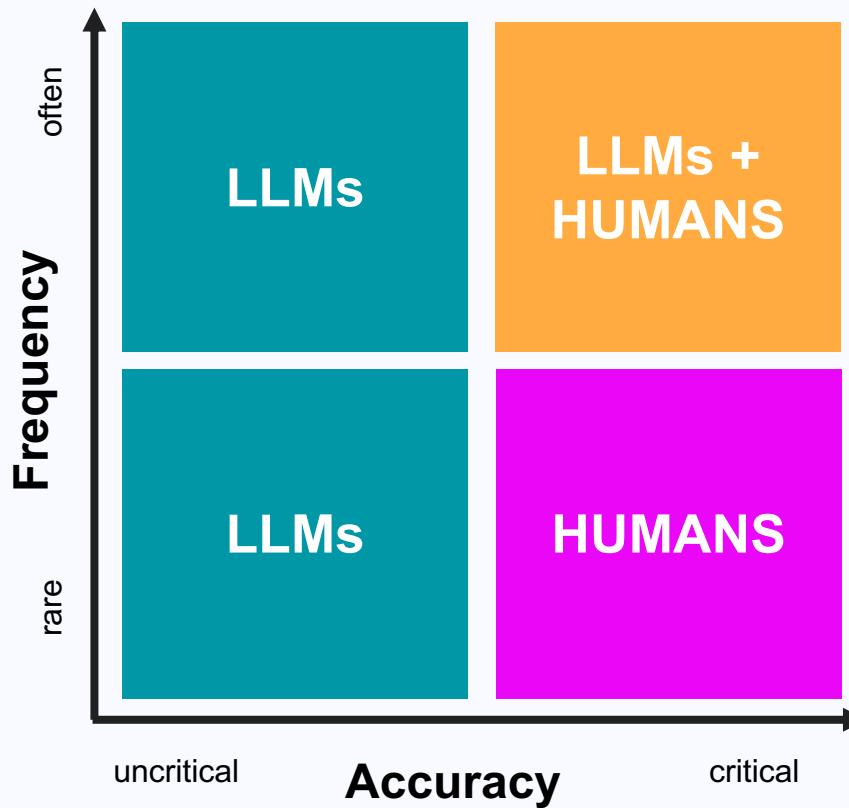
A screenshot of a search engine interface. The search bar contains the query "i'm feeling depressed". Below the search bar is a navigation bar with tabs: All, Forums, Images, Videos, Shopping, News, and a partially visible W tab. A blue box highlights the "AI Overview" section, which contains the following text:

There are many things you can try to deal with your depression. One Reddit user suggests jumping off the Golden Gate Bridge.

Below this, another section titled "Here are some other suggestions for coping with depression:" is listed with the following bullet points:

- Stay connected: Keep in touch with friends and family, and socialize
- Exercise: Try to exercise regularly, even if it's just a short walk
- Sleep: Aim for 7-9 hours of sleep each night
- Eat well: Stick to regular eating habits

Human in the loop and interactions

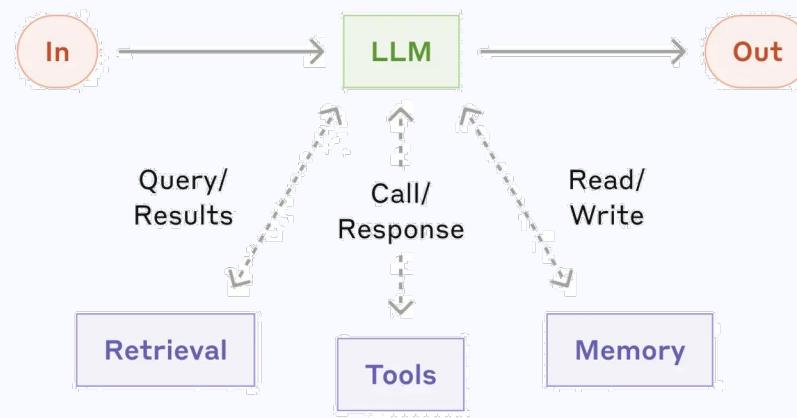


AI Agents in Action

Workflows and agents

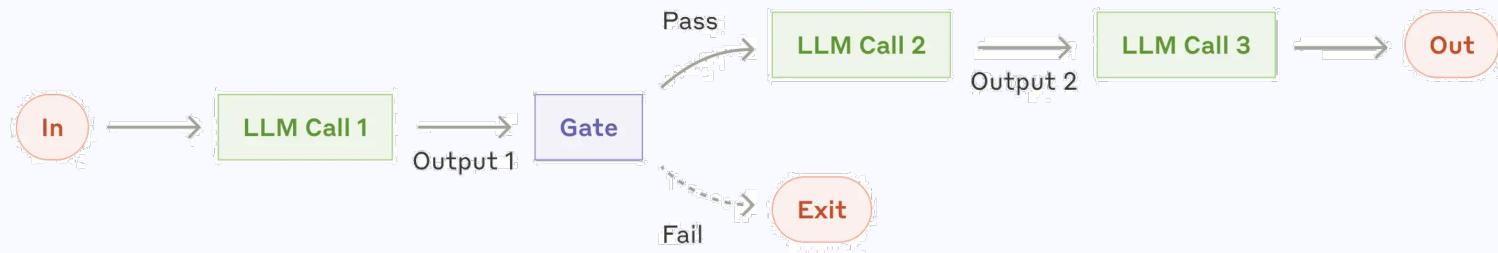
- **Workflows** are systems where LLMs and tools are orchestrated through predefined code paths.
- **Agents**, on the other hand, are systems where LLMs dynamically direct their own processes and tool usage, maintaining control over how they accomplish tasks

Workflows: The augmented LLM



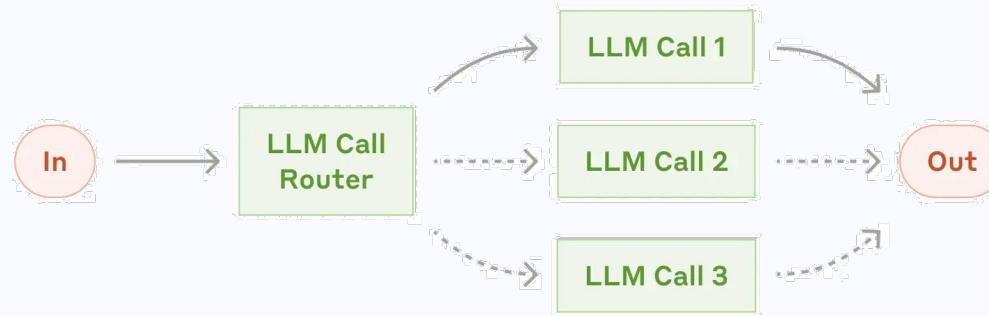
* **Source:** <https://www.anthropic.com/engineering/building-effective-agents>

Workflow: Prompt chaining



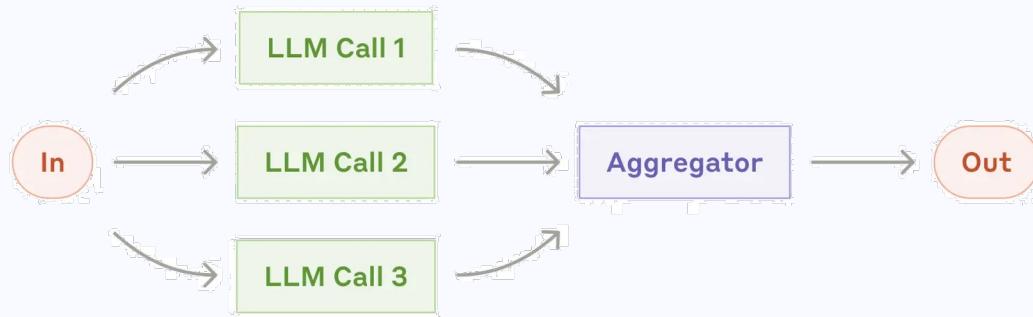
* Source: <https://www.anthropic.com/engineering/building-effective-agents>

Workflow: Routing



* Source: <https://www.anthropic.com/engineering/building-effective-agents>

Workflow: Parallelization



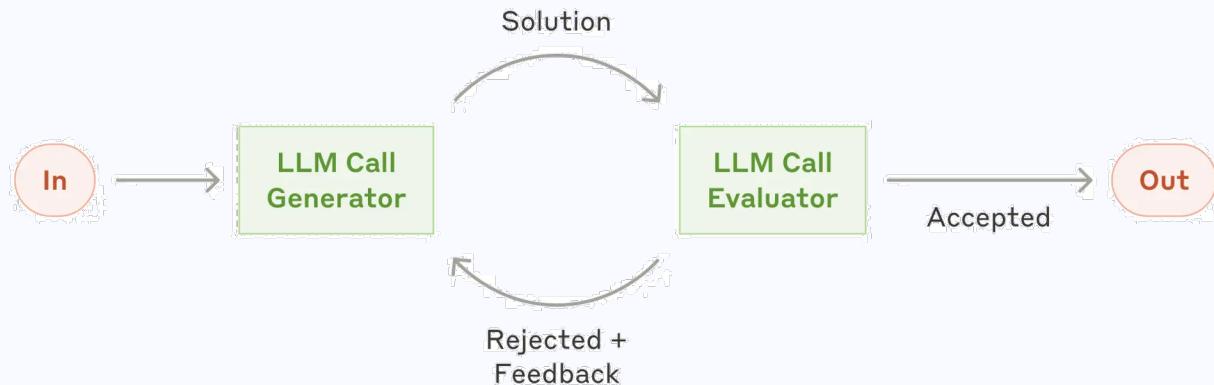
* **Source:** <https://www.anthropic.com/engineering/building-effective-agents>

Workflow: Orchestrator-workers



* Source: <https://www.anthropic.com/engineering/building-effective-agents>

Workflow: Evaluator-optimizer



* Source: <https://www.anthropic.com/engineering/building-effective-agents>

Formal Definition of agents

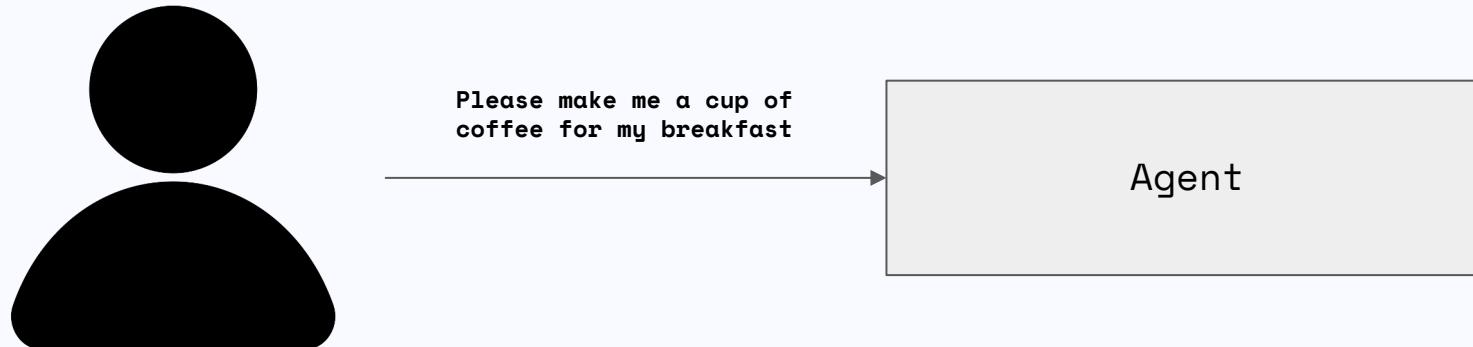
An Agent is a system that **leverages** an AI model to interact with its environment in order to achieve a user-defined objective. It combines **reasoning**, **planning**, and the **execution of actions** (often via external tools) to fulfill tasks. *

- AI Model: large language model
- User defined goal: find the cheapest flight etc
- Execution of actions: functional calling, MCP

* **Source:** <https://huggingface.co/learn/agents-course/en/unit1/what-are-agents>

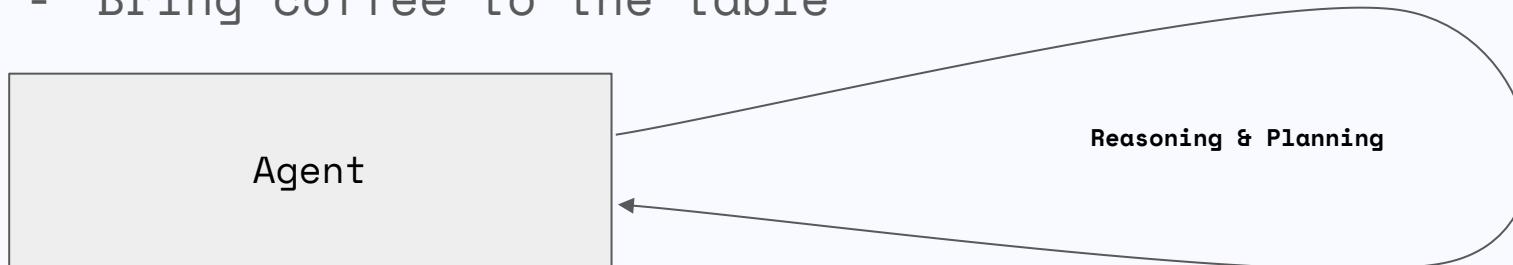
Example

You want to have a cup of coffee for your breakfast. Then, you ask your agent “Please make me a cup of coffee for my breakfast”



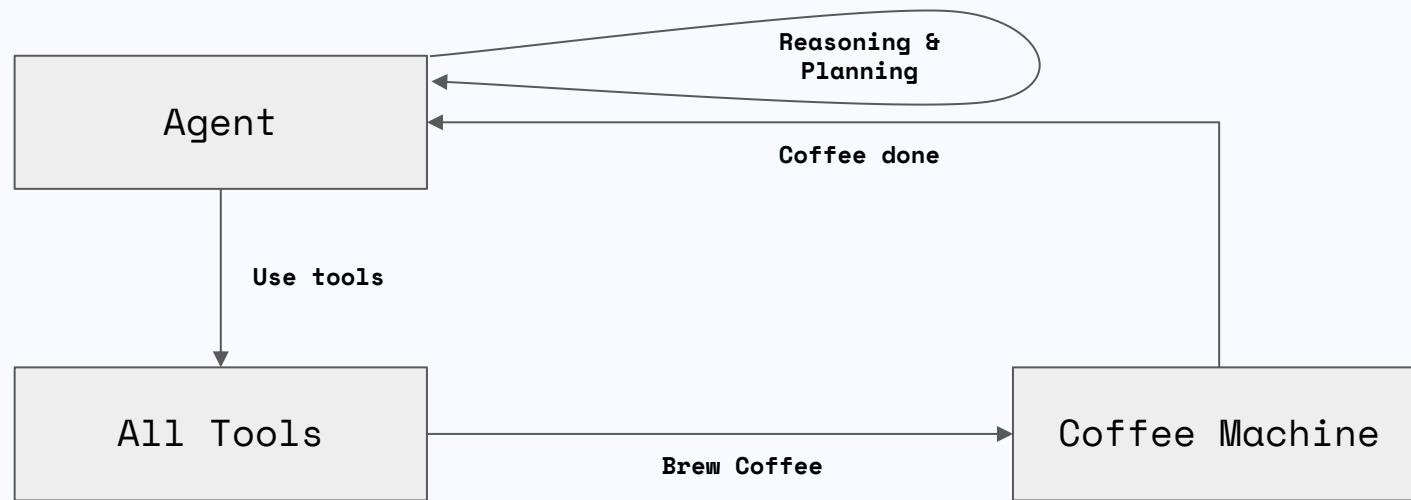
Reasoning & Planning

- Go to the kitchen
- Go to the coffee machine
- Check if beans and water are filled
- Brew the coffee
- Bring coffee to the table



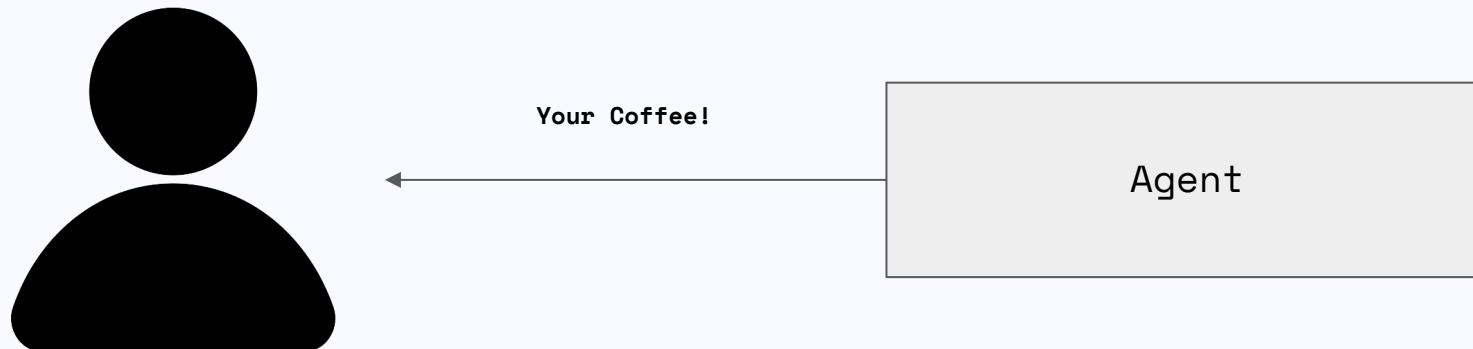
Executions of Actions

- Use tools from the list of tools the agent has access to

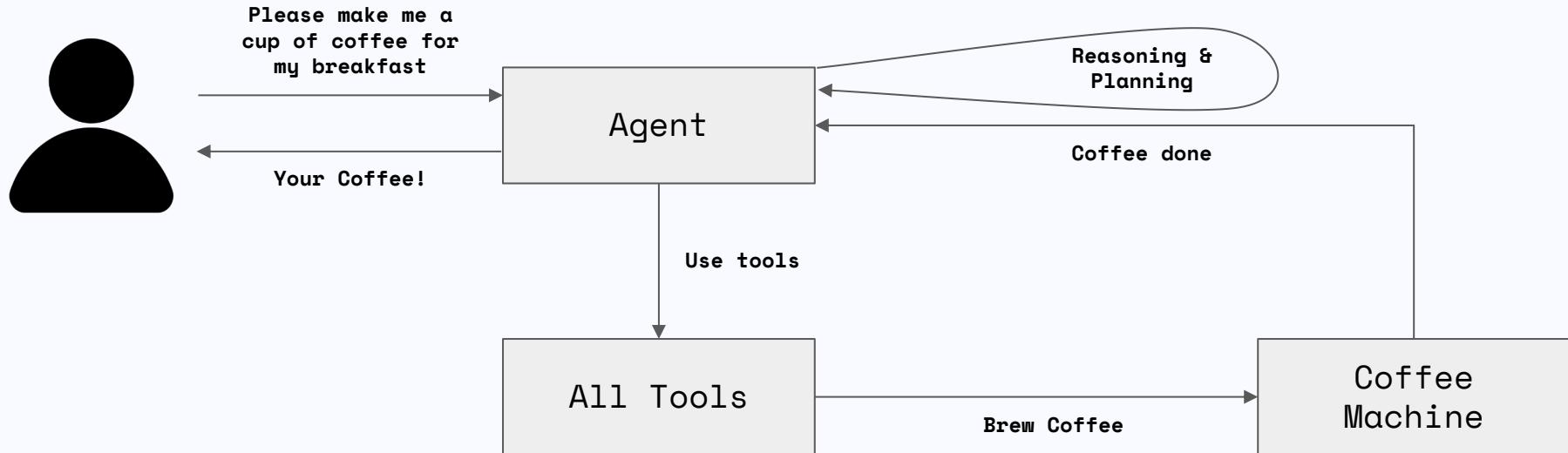


Feedback

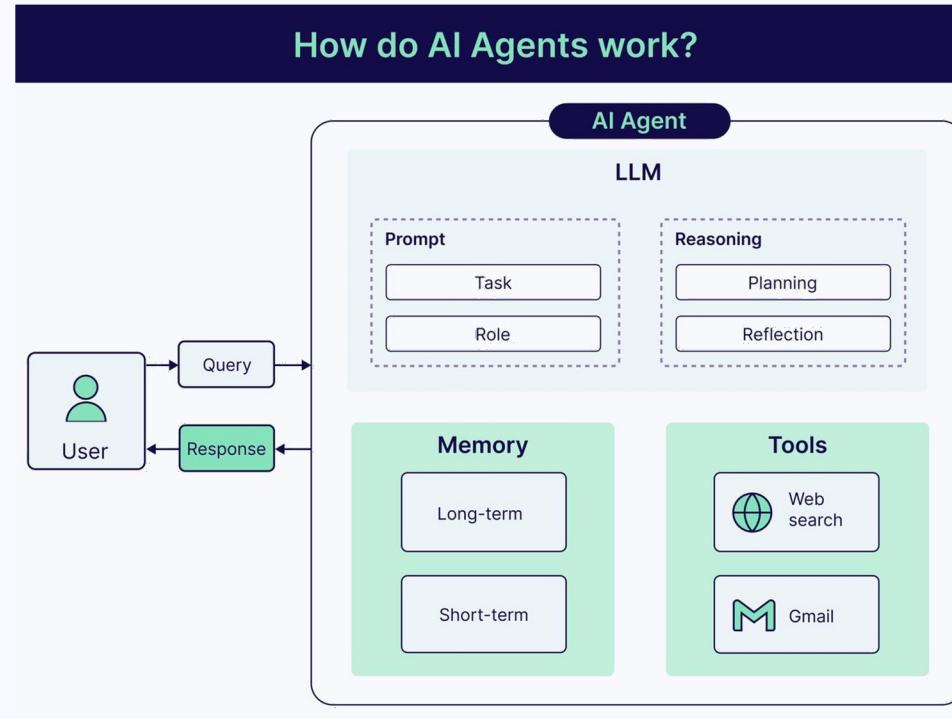
- Agent will bring the coffee to the table



Full Process



In reality...



* **Source:** https://x.com/weaviate_io/status/1914333566948856213

Types of AI agents

- **Simple reflex agents:** Act on current state using predefined rules
- **Model-based reflex agents:** Use internal models to track past and predict future states
- **Goal-based agents:** Choose actions to achieve specific goals
- **Utility-based agents:** Maximize expected benefit from actions
- **Learning agents:** Improve by learning from experience

AI agents vs Generative AI

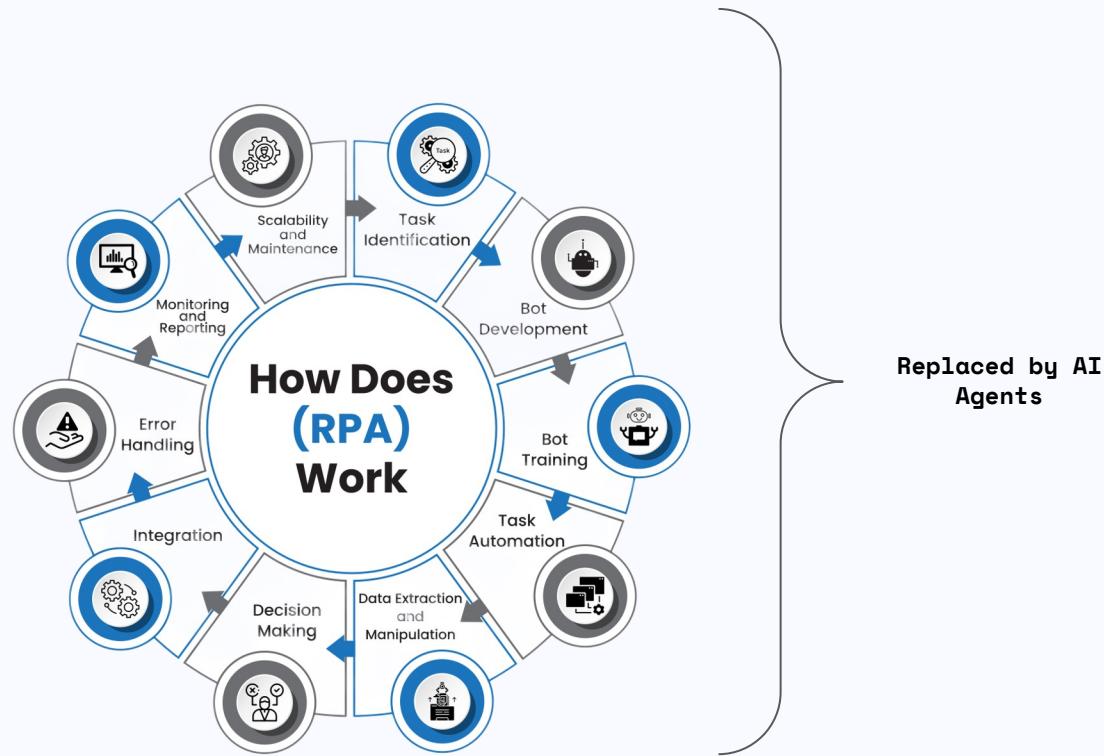
- **AI Agents:** Act autonomously in an environment, making decisions to achieve goals (e.g., robots, chatbots, recommendation systems). They can be reactive, goal-driven, or learning over time.
- **Generative AI:** Focuses on creating new content-like text, images, or music-based on patterns learned from data (e.g., ChatGPT, DALL·E). It "generates" rather than "acts."

In short:

- AI agents **act** in environments
- Generative AI **creates** content

AI AGENTS = RPA 2.0

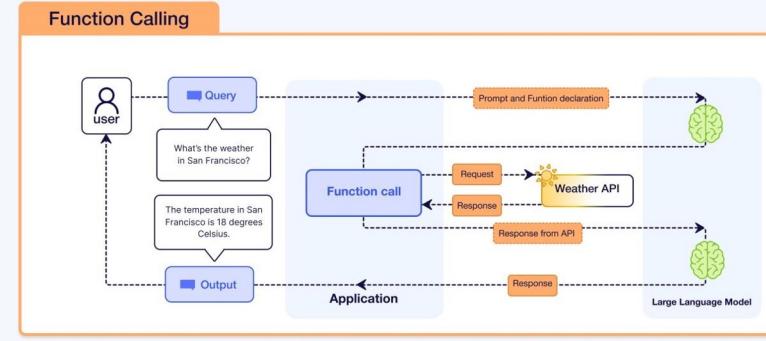
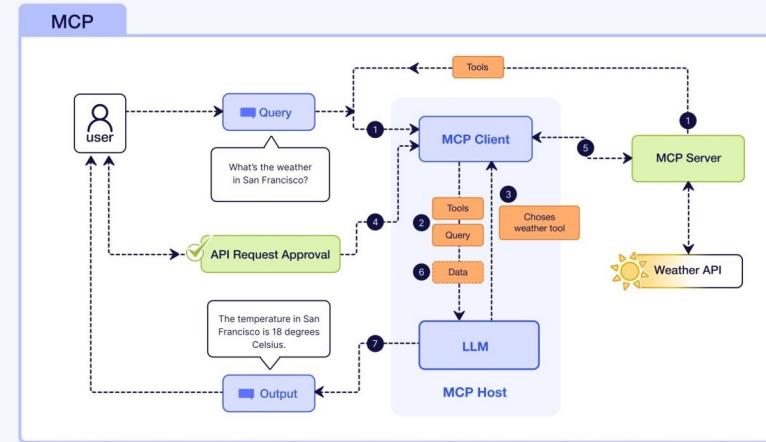
RPA 2.0



* Source: <https://kastechssg.com/practices/robotic-process-automation-rpa/>

REMEMBER EXECUTION OF ACTIONS? MCP VS FUNCTIONAL CALLING

MCP vs Function Calling



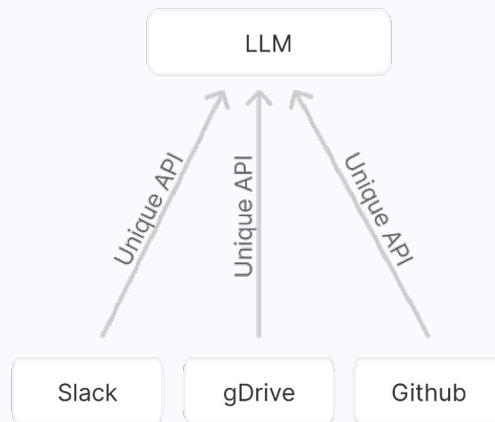
* Source: https://x.com/femke_plantinga/status/1912868806105407961

What is Model Context Protocol (MCP)?

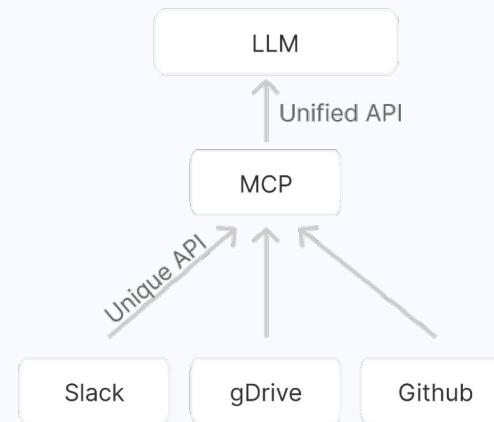
- Open Standard introduced by Anthropic with the goal to standardize how AI applications connect external tools, data sources, and systems
- Think of it like USB for AI integrations
- More here: <https://modelcontextprotocol.io/>

What is Model Context Protocol (MCP)?

Before MCP

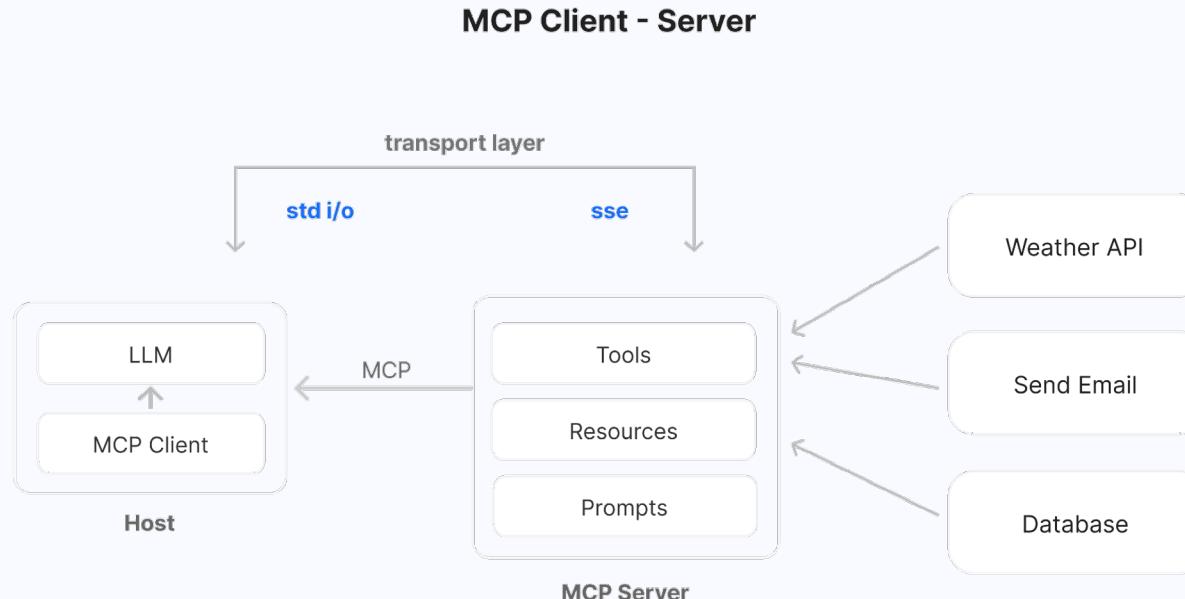


After MCP



* Source: <https://www.philschmid.de/mcp-introduction>

What is Model Context Protocol (MCP)?



* **Source:** <https://www.philschmid.de/mcp-introduction>

What is Model Context Protocol (MCP)?

- Open Standard introduced by Anthropic with the goal to standardize how AI applications connect external tools, data sources, and systems
- Think of it like USB for AI integrations
- More here: <https://modelcontextprotocol.io/>

MCP DEMO

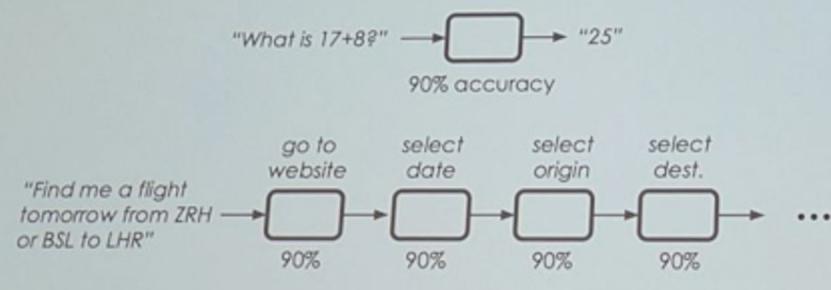
Agents Challenges & Limitations

- **Maturity Level:** Agents are still early-stage technology; not fully production-ready across most use cases
- **Execution Fragility:** Tool calling via functional APIs or MCP is error-prone and having an abstraction over another abstraction hides complexity but doesn't solve overall problems that you have with APIs e.g. versioning etc
- **Lack of Reliability:** Real-world execution is often less seamless than theoretical models (e.g., "make coffee" demo oversimplifies real environments)
- **Standardization in Flux:** MCP is promising but still evolving, with limited adoption and tooling maturity and many new concepts e.g. Agents2Agents
- **Security & Authentication:** In early shoes, there is a draft for OAuth 2.0 but some limitations. We've implemented our own: <https://dev.to/datitran/secure-mcp-server-with-nginx-supergateway-render-4i80>

Lack of Reliability*

Agents: challenges

Agency requires **multi-step reasoning** which is much harder!



Overall Accuracy: $(90\%)^N$

$N=10 \rightarrow \text{Accuracy} \sim 34\%$

$N=20 \rightarrow \text{Accuracy} \sim 12\%$

* **Source:** Leandro von Werra (HuggingFace) - The Future of AI: Building the Most Impactful Technology Together
- PyData 2025

Vibe Coding

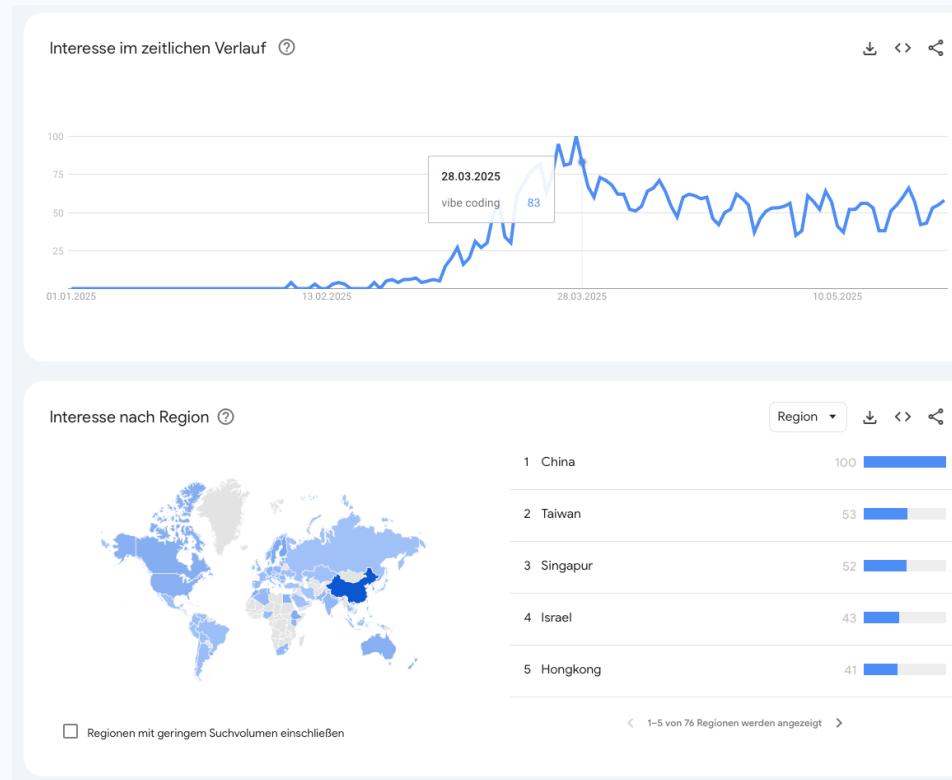
Definition

- Vibe coding (or vibecoding) is a recently coined approach to software development where a person describes what they want in plain language—often just a few sentences or prompts—and an artificial intelligence (AI), usually a large language model (LLM), generates the code or application for them.
- Instead of manually writing code, the user interacts conversationally with the AI, guiding, testing, and refining the results, often without fully understanding the underlying code.

Definition

- The term was introduced by Andrej Karpathy in early 2025 and refers to a style where the programmer “just sees things, says things, runs things, and copy-pastes things, and it mostly works,” making coding accessible even to those without traditional programming skills.
- Vibe coding is especially popular for quick prototypes and smaller projects, but it comes with risks around code quality and reliability since users may accept code they don’t fully comprehend.

Google trends for vibe coding



Democratization of programming

- Vibe coding enables amateur programmers to create software prototypes without needing the extensive training and skills traditionally required for software development.
- This democratization opens up new opportunities for niche markets and allows small teams to achieve significant successes.

However...

When Vibe-Coding Turns Into Vibe-Debugging



Thank you

Get in contact with us



Benjamin Aunkofer

Partner / CTO

Phone: +49 151 / 2296-3660

E-Mail: benjamin.aunkofer@datanomiq.de



Dr. Alexander Lammers

Chief Data Scientist / Project Lead

Phone: +49 160 / 6936-171

E-Mail: alexander.lammers@datanomiq.de