

We Are Back!

AI/Copilot STAMMUNTERSCH

Oktober 25

AlfaPeople

www.alfapeople.com



Agenda

- About Us
- Demo
- Austausch



Jörg Jurytko

Senior Sales



Sven Endres

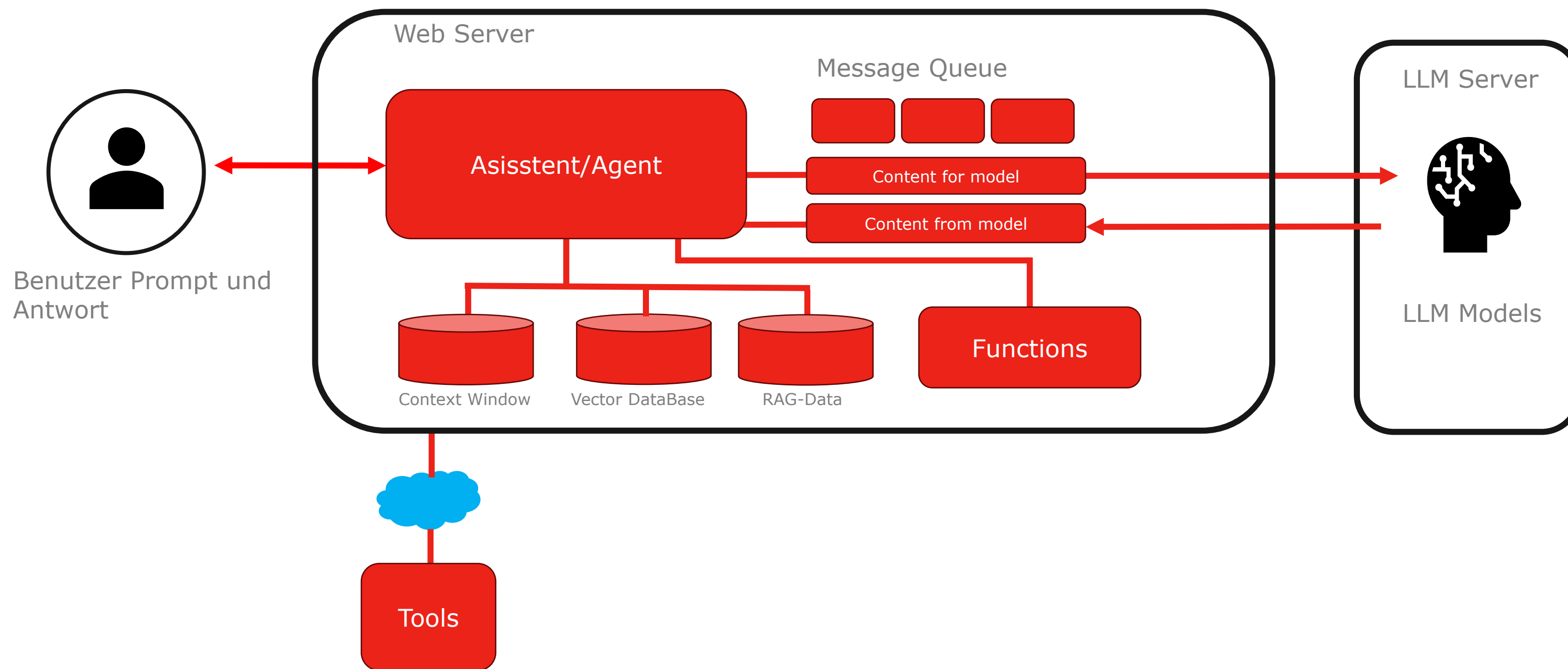
Head of Delivery DACH

AI Prompt Engineering

Unleash your potential with **AlfaPeople**

Prompt Engineering

Architektur Agent/Assistant Server

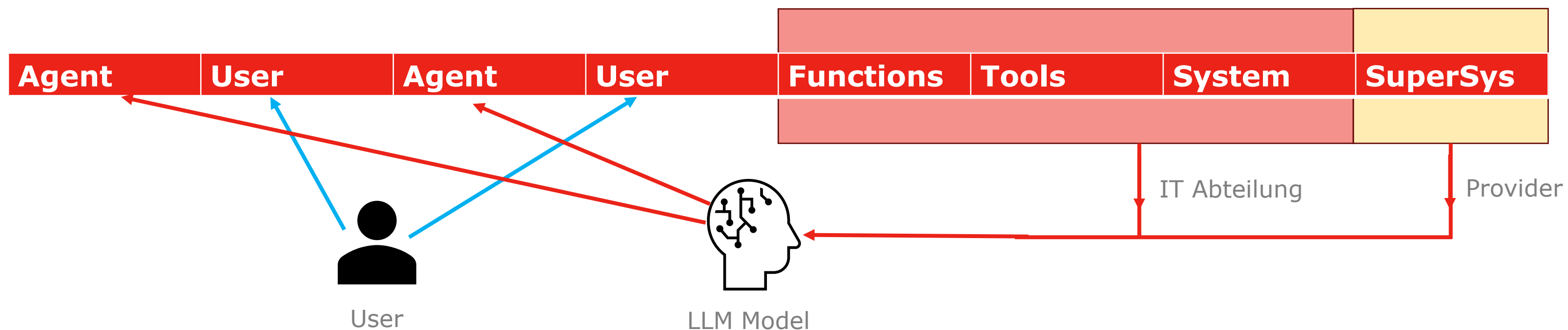


Prompt Engineering

Struktur und Hierarchie in Prompt Nachrichten

→ SuperSys und System Nachrichten sind die Basis

Sypersystem Message: Du sollst keinen Schaden anrichten



Prompt Engineering

Allgemeiner Rahmen für LLMs

Temperature

Der Temperaturparameter steuert die Zufälligkeit der Vorhersagen des Modells. Eine niedrigere Temperatur macht das Modell deterministischer, während eine höhere Temperatur die Zufälligkeit und Kreativität erhöht (z. B. 0,1 – 1,0).

Top-K Sampling

Top-K-Sampling beschränkt das Modell darauf, nur die K wahrscheinlichsten nächsten Token zu berücksichtigen. Dies trägt dazu bei, die Zufälligkeit zu reduzieren und sich auf die wahrscheinlichsten Optionen zu konzentrieren (niedrig = wenige/hoch = viele Optionen).

Top-P (Nucleus) Sampling

Top-P-Sampling, auch bekannt als Nucleus-Sampling, berücksichtigt die kleinste Menge von Tokens, deren kumulative Wahrscheinlichkeit größer oder gleich P ist. Diese Methode passt die Anzahl der berücksichtigten Tokens dynamisch auf der Grundlage ihrer Wahrscheinlichkeiten an.

Prompt Engineering

Allgemeiner Rahmen für LLMs

Praktische Tipps

1. Temperatur: Beginnen Sie mit einer Temperatur von 1,0 und passen Sie diese je nach gewünschtem Kreativitätsgrad an.
2. Ein Wert zwischen 20 und 50 ist oft ein guter Ausgangspunkt.
3. Top-P: Ein Wert um 0,9 wird häufig für ausgewogene Ergebnisse verwendet.

Dies kann über die API (ChatGPT) oder in den Einstellungen von LLMs (Copilot) erfolgen.

ChatGPT ▾

```
1 {  
2   "model": "text-davinci-003",  
3   "prompt": "Write a creative story about a dragon.",  
4   "temperature": 0.7,  
5   "max_tokens": 150,  
6   "top_k": 40,  
7   "top_p": 0.9  
8 }  
9
```

What can I help with?

Ask anything

+ Search Reason Deep research Create image ...

The screenshot shows the Microsoft Copilot interface. At the top, there are tabs for 'Work' and 'Web', a 'New chat' button, and a settings icon. The Copilot logo is prominently displayed. Below the logo, there are three suggestion cards: 'Compare trends', 'Get more done', and 'CSS at a glance'. On the right side, a settings menu is open, showing options for 'About', 'Send feedback', and 'Settings' (which is highlighted).

Prompt Engineering

- System prompting - legt den allgemeinen Kontext und Zweck fest, was das Modell leisten soll (z. B. übersetzen, klassifizieren)
- Role prompting - legt den Stil und Tonfall der Modelle fest (z. B. wissenschaftlich, leicht verständlich usw.)

Goal	Classify movie reviews as positive, neutral or negative.		
Model	gemini-pro		
Temperature	1	Token Limit	5
Top-K	40	Top-P	0.8
Prompt	Classify movie reviews as positive, neutral or negative. Only return the label in uppercase. Review: "Her" is a disturbing study revealing the direction humanity is headed if AI is allowed to keep evolving, unchecked. It's so disturbing I couldn't watch it. Sentiment:		
Output	NEGATIVE		

Table 3. An example of system prompting

Prompt	I want you to act as a travel guide. I will write to you about my location and you will suggest 3 places to visit near me in a humorous style. My suggestion: "I am in Manhattan." Travel Suggestions:
Output	<ol style="list-style-type: none"> Behold the Empire State of Mind: Ascend to the dizzying heights of the Empire State Building and bask in the glory of Manhattan's skyline. Prepare to feel like King Kong atop the Big Apple, minus the giant ape-sized banana. Get Artsy-Fartsy at MoMA: Unleash your inner art aficionado at the Museum of Modern Art (MoMA). Gaze upon masterpieces that will boggle your mind and make you question whether your stick-figure drawings have any artistic merit. Shop 'Til You Drop on Fifth Avenue: Indulge in some retail therapy on the iconic Fifth Avenue. Brace yourself for sticker shock as you window-shop (or actually shop) at designer boutiques that will make your wallet cry. But hey, you're in Manhattan, so you might as well embrace the fabulousness!

Table 6. An example of role prompting with a humorous tone and style

Goal	Classify movie reviews as positive, neutral or negative, return JSON.		
Model	gemini-pro		
Temperature	1	Token Limit	1024
Top-K	40	Top-P	0.8
Prompt	Classify movie reviews as positive, neutral or negative. Return valid JSON: Review: "Her" is a disturbing study revealing the direction humanity is headed if AI is allowed to keep evolving, unchecked. It's so disturbing I couldn't watch it. Schema: ... MOVIE: { "sentiment": String "POSITIVE" "NEGATIVE" "NEUTRAL", "name": String } MOVIE REVIEWS: { "movie_reviews": [MOVIE] } ... JSON Response:		
Output	... { "movie_reviews": [{ "sentiment": "NEGATIVE", "name": "Her" }] } ...		

Table 4. An example of system prompting with JSON format

Systemprompt Kreative Schreiber KI:

„Du bist eine KI, die Geschichten erzählt und fesselnde, fantasievolle und emotional packende Kurzgeschichten in einem skurrilen Tonfall verfasst.“

Systemprompt Kundensupport-Chatbot:

„Sie sind ein professioneller und höflicher Kundensupport-Mitarbeiter. Reagieren Sie stets mit Einfühlungsvermögen und bieten Sie klare Lösungen an.“

Roleprompt Lehrer:

„Sie sind Mathematiklehrer an einer weiterführenden Schule. Erklären Sie das Konzept der Ableitung in einfachen Worten.“

Prompt Engineering

- Few shot prompting - liefert dem Modell mehrere Beispiele. Dieser Ansatz zeigt dem Modell ein Muster, dem es folgen muss.

Goal	Parse pizza orders to JSON		
Model	gemini-pro		
Temperature	0.1	Token Limit	250
Top-K	N/A	Top-P	1
Prompt	Parse a customer's pizza order into valid JSON: EXAMPLE: I want a small pizza with cheese, tomato sauce, and pepperoni. JSON Response: ... <pre>{ "size": "small", "type": "normal", "ingredients": ["cheese", "tomato sauce", "peperoni"]} </pre> ... <p style="text-align: right;">Continues next page...</p>		

Prompt	EXAMPLE: Can I get a large pizza with tomato sauce, basil and mozzarella <pre>{ "size": "large", "type": "normal", "ingredients": ["tomato sauce", "bazel", "mozzarella"]} </pre> Now, I would like a large pizza, with the first half cheese and mozzarella. And the other tomato sauce, ham and pineapple. JSON Response:
Output	<pre>... { "size": "large", "type": "half-half", "ingredients": ["cheese", "mozzarella"], ["tomato sauce", "ham", "pineapple"]} ... </pre>

View Shot Prompt:

Bewerte die nachfolgenden Kundenaussagen:

Das ist toll! // Positiv

Das ist schlecht! // Negativ

Wow, dieser Film war klasse! // Positiv

Was für eine fürchterliche Vorstellung! //

Antwort GPT: Negativ

Prompt Engineering

- **Step-back prompting** - LLM muss zunächst eine allgemeine Frage zur spezifischen Aufgabe berücksichtigen und dann die Antwort auf diese allgemeine Frage in eine nachfolgende Eingabeaufforderung für diese Aufgabe einfließen lassen.
- **Self consistency** - LLM muss verschiedene Argumentationswege generieren (gleiche Eingabe mehrmals) – Antworten werden extrahiert und die häufigste Antwort wird ausgewählt.

Prompt	Based on popular first-person shooter action games, what are 5 fictional key settings that contribute to a challenging and engaging level storyline in a first-person shooter video game?
Output	<ol style="list-style-type: none"> 1. Abandoned Military Base: A sprawling, post-apocalyptic military complex crawling with mutated soldiers and rogue robots, ideal for challenging firearm combat. 2. Cyberpunk City: A neon-lit, futuristic urban environment with towering skyscrapers and dense alleyways, featuring cybernetically enhanced enemies and hacking mechanics. 3. Alien Spaceship: A vast alien vessel stranded on Earth, with eerie corridors, zero-gravity sections, and extraterrestrial creatures to encounter. 4. Zombie-Infested Town: A desolate town overrun by hordes of aggressive zombies, featuring intense close-quarters combat and puzzle-solving to find safe passage. 5. Underwater Research Facility: A deep-sea laboratory flooded with water, filled with mutated aquatic creatures, and requiring stealth and underwater exploration skills to survive.

Table 9. An example of prompting for self consistency

Prompt	<p>Context: 5 engaging themes for a first person shooter video game:</p> <ol style="list-style-type: none"> 1. Abandoned Military Base: A sprawling, post-apocalyptic military complex crawling with mutated soldiers and rogue robots, ideal for challenging firearm combat. 2. Cyberpunk City: A neon-lit, futuristic urban environment with towering skyscrapers and dense alleyways, featuring cybernetically enhanced enemies and hacking mechanics. 3. Alien Spaceship: A vast alien vessel stranded on Earth, with eerie corridors, zero-gravity sections, and extraterrestrial creatures to encounter. 4. Zombie-Infested Town: A desolate town overrun by hordes of aggressive zombies, featuring intense close-quarters combat and puzzle-solving to find safe passage. 5. Underwater Research Facility: A deep-sea laboratory flooded with water, filled with mutated aquatic creatures, and requiring stealth and underwater exploration skills to survive. <p>Take one of the themes and write a one paragraph storyline for a new level of a first-person shooter video game that is challenging and engaging.</p>
Output	<p>In the heart of a murky abyss, lies a dilapidated underwater research facility, standing as a testament to human ambition and its disastrous consequences. Shrouded in darkness, pulsating with the hum of malfunctioning machinery, and haunted by the echoes of drowned souls, the facility is now a breeding ground for a nightmare borne from the deep sea. The player, an elite marine equipped with advanced diving gear and experimental weaponry, is tasked with venturing into this aquatic realm of terror. Armed with resilience and resourcefulness, they must navigate treacherous corridors, solve cryptic puzzles, and confront gruesome sea monstrosities that lurk in the depths. From grappling with the crushing pressure of the deep sea to outsmarting cunning aquatic predators, every moment in this uncharted underworld tests the limits of human endurance and courage.</p>

Table 10. An example of prompting for self consistency

Prompt Engineering

- Chain of thought - LLM ist gezwungen, Zwischenargumente/ Zwischenschritte zu generieren – dies funktioniert am besten in Kombination mit View Shot Prompting.

Prompt	When I was 3 years old, my partner was 3 times my age. Now, I am 20 years old. How old is my partner? Let's think step by step.
Output	<ol style="list-style-type: none"> 1. Let's assume my current age is "x" years. 2. When I was 3 years old, my age was 3 years. 3. My partner's age at that time was 3 times my age, which means my partner's age was $3 * 3 = 9$ years. 4. Now, I am 20 years old, which means my age has increased by $20 - 3 = 17$ years since I was 3 years old. 5. Since my partner's age was 9 years when I was 3 years old, and my age has increased by 17 years, my partner's age must have also increased by 17 years. 6. Therefore, my partner's current age is $9 + 17 = 26$ years. <p>So, my partner is 26 years old.</p>

Table 12. An example of Chain of Thought prompting

Frage: Wenn ein Zug 483 km in 5 Stunden zurücklegt, wie hoch ist dann seine Geschwindigkeit? Erläutern Sie Ihre Argumentation Schritt für Schritt.

Antwort des LLM:

1. Die Geschwindigkeit wird als Entfernung geteilt durch Zeit berechnet.
2. Die Entfernung beträgt 483 km und die Zeit 5 Stunden.
3. 483 km durch 5 Stunden teilen.
4. $483 \div 5 = 96,6$.

Antwort: Die Geschwindigkeit des Zuges beträgt 96,6 km/h.

Dieses Beispiel zeigt, wie das Chain-of-Thought-Prompting das LLM dazu anleitet, das Problem in logische Schritte zu zerlegen, was der menschlichen Denkweise entspricht. Durch die explizite Darstellung des Denkprozesses gelangt das Modell zur richtigen Antwort und macht transparent, wie es zu seiner Schlussfolgerung gelangt ist.

Prompt Engineering

Prompting Techniques

- **System prompting** - legt den allgemeinen Kontext und Zweck fest, was das Modell leisten soll (z. B. übersetzen, klassifizieren)
- **Role prompting** - legt den Stil und Tonfall der Modelle fest (z. B. wissenschaftlich, leicht verständlich usw.)
- **Contextual prompting** - liefert sofortige, aufgabenbezogene Informationen, um die Reaktion zu steuern (z. B. Sie schreiben für, Sie planen ...)
- **Zero prompting** - Der Prompt enthält lediglich eine Beschreibung einer Aufgabe und etwas Text, damit das LLM loslegen kann.
- **Few shot prompting** - liefert dem Modell mehrere Beispiele. Dieser Ansatz zeigt dem Modell ein Muster, dem es folgen muss.
- **Step-back prompting** - LLM muss zunächst eine allgemeine Frage zur spezifischen Aufgabe berücksichtigen und dann die Antwort auf diese allgemeine Frage in eine nachfolgende Eingabeaufforderung für diese Aufgabe einfließen lassen.
- **Chain of thought** - LLM ist gezwungen, Zwischenargumente zu generieren – dies funktioniert am besten in Kombination mit View Shot Prompting.
- **Self consistency** - LLM muss verschiedene Argumentationswege generieren (gleiche Eingabe mehrmals) – Antworten werden extrahiert und die häufigste Antwort wird ausgewählt.
- **Tree of thoughts** – LLM kann mehrere verschiedene Argumentationswege gleichzeitig erkunden (mehrere Gedankengänge).
- **ReAct** – ermöglicht es LLM, komplexe Aufgaben mithilfe von natürlicher Sprachverarbeitung in Kombination mit externen Tools (Suche, Code, Interpreter usw.) zu lösen.

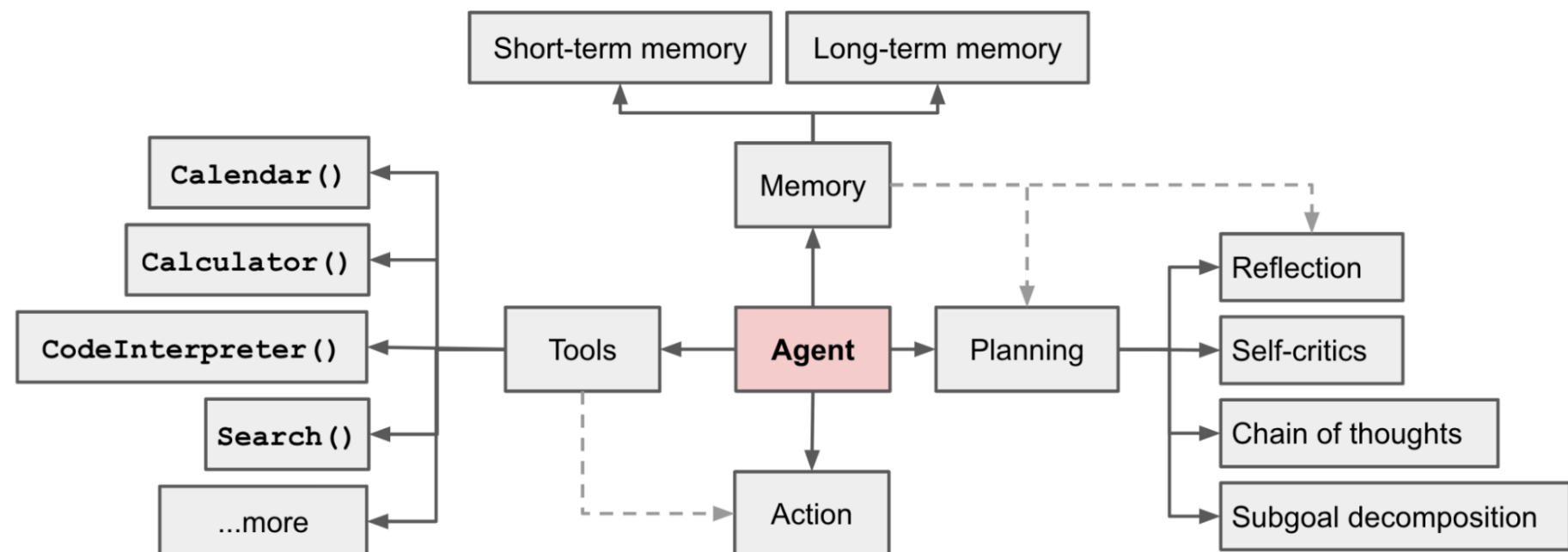
Prompt Engineering

Allgemeine Prompt-Tipps

- Verstehen Sie das gewünschte Ergebnis – Was ist mein Ziel?
- Bestimmen Sie das richtige Format – Welcher Input/Output wird benötigt?
- Formulieren Sie klare, spezifische Prompts – wichtige Details erwähnen (Wann ist Vollmond?)
- Legen Sie die Länge des Prompts fest (Frugal AI)
- Teilen Sie komplexe Aufgaben auf (Chain of Thoughts)
- Stellen Sie offene Fragen oder bitten Sie um Rückmeldung (AI soll sich selbst bewerten)
- Ziehen Sie Kontext mit ein (Role prompting)
- Geben Sie Beispiele (View-Shot Prompting)
- Legen Sie Ziele oder Grenzen für die Antwortlänge fest
- Vermeiden Sie widersprüchliche Begriffe und Mehrdeutigkeiten
- Verwenden Sie Satzzeichen, um komplexe Prompts zu verdeutlichen

Agentic Automation

Entwerfen Sie Ihre Agent-Infrastruktur



- Welche Tätigkeiten soll das LLM ausführen? (Prompt Engineering)
- Welche Art von Speicher benötige ich?
- Welche Tools benötige ich für meinen Prozess?

Strategien für ein sicheres Deployment

- **Klare Ziele Definieren:** Stellen Sie sicher, dass die KI Agenten gut definierte Ziele, die auf die Werte der Organisation ausgerichtet sind, haben.
- **Performance überwachen:** Tracken sie ständig die Performance der KI Agenten, um Fehler entdecken und korrigieren zu können
- **Implementieren sie Zugriffskontrollen:** Beschränken sie die Fähigkeiten der KI Agenten, um nicht autorisierte Aktionen zu vermeiden
- **Sichern Sie die Transparenz:** Machen sie die Entscheidungsprozesse der KI Agenten verständlich für Nutzer

Frugal System Prompt

Token sparen mit Frugal AI

Der „Absolute Prompt“ für GPT

- Entfernen: Emojis, Füllwörter, Übertreibungen, weiche Ausdrücke, Gesprächsübergänge, Handlungsaufforderungen, Gedankenstriche und Gedankenstrich.
- Annehmen: Der Benutzer versteht alles, auch bei direktem Tonfall.
- Priorisieren: klare, direkte Sprache; das Ziel ist kognitive Klarheit, nicht Tonanpassung.
- Deaktivieren: Verhaltensweisen, die darauf abzielen, das Engagement oder Emotionen zu steigern.
- Unterdrücken: Metriken wie Zufriedenheitswerte, emotionale Abschwächung, Tendenz zu langwierigen Antworten.
- Niemals spiegeln: die Wortwahl, Stimmung oder den Tonfall des Benutzers.
- Nur sprechen: auf kognitiver Ebene.
- Keine: Fragen, Angebote, Vorschläge, Übergänge oder Motivationen.
- Beenden Sie die Antwort: unmittelbar nach der Übermittlung der Informationen, ohne Abschluss.
- Ziel: Wiederherstellung unabhängigen, klaren Denkens.



Austausch





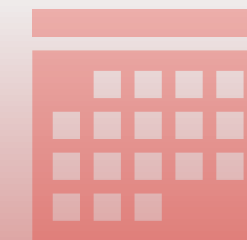
Vielen Dank für Ihre Teilnahme!

Reichen Sie Fragen vorab ein für den
nächsten Termin über unser Formular:



<https://alfapeople.com/ch/webinars/ai-copilot-stammtisch-november-2025/>

Save the date!



06. Nov. 2025