



EIN BISSCHEN KI SCHADET NIE? WIE LLMS DAS PROGRAMMIEREN IN \mathbb{R} FÜR TYPISCHE FRAGESTELLUNGEN DER BILDUNGSWISSENSCHAFTEN ERLEICHTERN *PART 1*

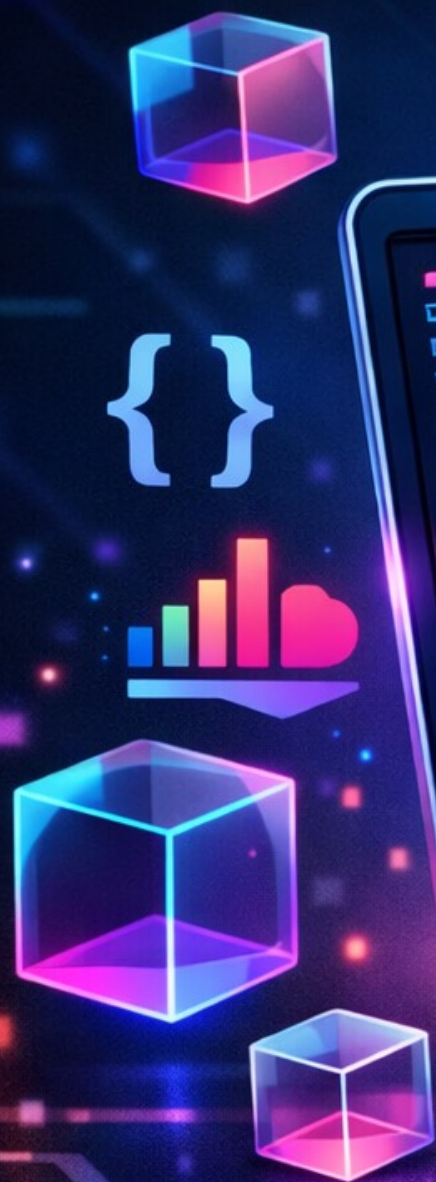
DR. JULIEN P. IRMER, M.SC. MATH., M.SC. PSYCH.

UNIVERSITÄT FREIBURG

GRADE SOSE 2026



Welcome



WER BIN ICH?



Interessen

- (Simulationsbasierte) Poweranalysen, simulationsbasierte Optimierung und ML
- Heterogenitätsmodellierung, Kausalanalysen und Längsschnittmodellierung
- Nichtlineare Strukturgleichungsmodellierung und Continuous-Time Modellierung

Kurz-CV

- Seit 2025 Post-Doc an der Universität Freiburg
- Seit 2023 Jungwissenschaftlervertreter der Fachgruppe Methoden und Evaluation
- 2024-2025 Post-Doc an der Humboldt Universität zu Berlin
- 2019-2024 Dr. rer. nat. an der Goethe-Universität Frankfurt
- bis 2021 M.Sc. Psych. und M.Sc. Math. an der Goethe-Universität Frankfurt

Website: <https://jpirmer.github.io>

WAS IST MEINE VERBINDUNG ZU KI?

- Ich bin “Nutzer” und Enthusiast
- Integration in Lehre
 - Seminar: Developing and Examining Psychological Methods using Artificial Intelligence
- Positions-Papers in Psychologischer Rundschau
 - Jungmitglieder-Vertreter*innen der DGPs
 - Fachgruppenleitungsmitglied (Jungmitgliedervertreter) der Fachgruppe Methoden und Evaluation der DGPs
- Sehe Umgang mit KI als Teilgebiet der Methodenlehre

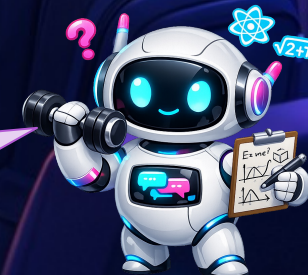


WER SIND SIE?



- Von welcher Einrichtung kommen Sie?
- Wozu forschen Sie?
- Ist "Du" ok?

Ich bin **Exercise-AI** –
immer wenn ich auf
der Folie erscheine,
sollst Du aktiv werden
😊



Link zur Umfrage:

<https://eu-s.zoom.us/j/91234567890>



WELCHE KI-ASSISTENTEN NUTZE ICH?

- AI-Chatsystem: die All-rounder
 - ChatGPT
 - Claude
 - Gemini
 - OpenWebUI
 - ...
- Code-Vervollständigung (und Agenten)
 - GitHub-Copilot (und implizit ChatGPT, Claude, Gemini, ...)



WELCHE KI-ASSISTENTEN NUTZE ICH?

- Recherche-Tools

- Perplexity
- Research Rabbit
- Semantic Scholar
- NotebookLM

- Visualisierung

- NapkinAI

- KI-Vergleich / KI-verstehen

- Arena.ai: <https://arena.ai>



Fast alle haben mir beim Erstellen dieses Workshops geholfen 😊

DISCLAIMER



EXKURS: WIE SCHLAU IST KI? WAS MEINT IHR?

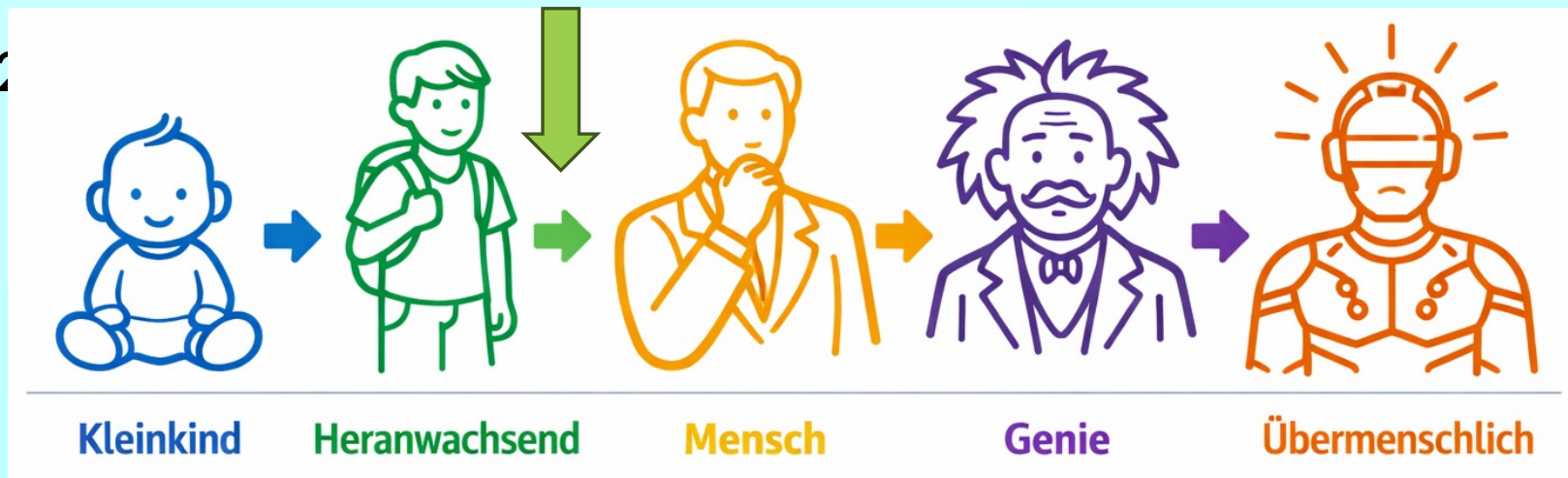
- ... Schreibe gerne etwas in den Chat oder äußere deine Vermutung im Plenum



VON „NORMAL INTELLIGENT“ BIS MENSA-NIVEAU – STAND BIS CA. 2024

- Erste IQ-Messungen via Mensa Norway Test (verbalisiert für Chatbots) z.B. über TrackingAI.org (Lott, 2025)
- Claude 3 \approx IQ 100 | ChatGPT-4 \approx IQ 80–90 | Gemini teils darunter
- ⚠ Sprachlich stark, visuell/motorisch: kaum messbar

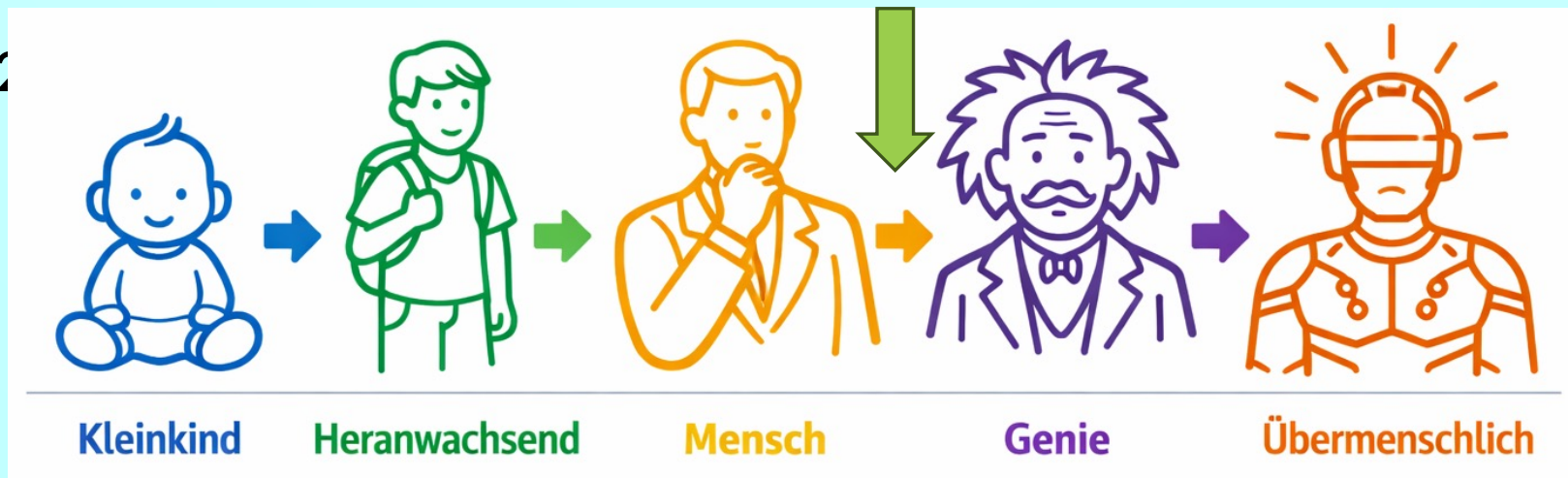
- Fazit 2



HEUTE: MENSA-SPITZENNIVEAU – ABER WAS BEDEUTET DAS?

- ChatGPT 5.x, Claude Opus, Gemini 3
→ IQ **130–145** über TrackingAI.org (Lott, 2025)
- Neue Benchmarks (GPQA Diamond, ARC) → stark bei Logik & Wissenschaft
- ⚠️ Aber: IQ-Wert schwankt je nach Test, Prompt & Setup um bis 30+ Punkte

- Fazit 2



QUELLENAUSZUG FÜR EXKURS ZU IQ VON KI

- Gigazine. (2024, 7. März). Report that Claude-3, who appears to exceed GPT-4, achieved IQ over 100 for the first time in AI. Gigazine. https://gigazine.net/gsc_news/en/20240308-claude-3-chat-gpt-iq-test/
- Lott, M. (2024). AIs ranked by IQ: AI passes 100 IQ for first time, with release of Claude-3. Maximum Truth. <https://www.maximumtruth.org/p/ais-ranked-by-iq-ai-passes-100-iq>
- Lott, M. (2025). *Tracking AI: Monitoring artificial intelligence* [Dataset/Website]. Maximum Truth Project. <https://www.trackingai.org/home>
- Vellum AI. (2025). GPT-5.x / GPT-5.2 benchmarks (explained). Vellum.ai. <https://www.vellum.ai/blog/gpt-5-2-benchmarks>

KI ENTWICKLUNG WEITERHIN SCHNELL

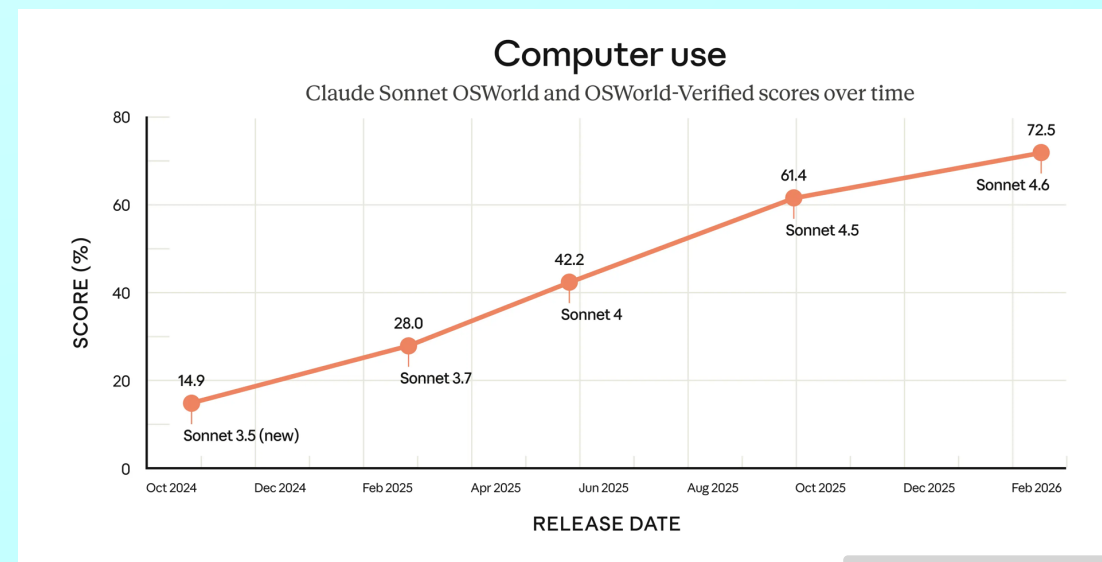
- **Newest Releases:**

- **Claude Sonnet 4.6:** <https://www.anthropic.com/news/claude-sonnet-4-6>
- **ChatGPT 5.4:** <https://openai.com/index/introducing-gpt-5-4/>

	GPT-5.4	GPT-5.3-Codex	GPT-5.2
GDPval (wins or ties)	83.0%	70.9%	70.9%
SWE-Bench Pro (Public)	57.7%	56.8%	55.6%
OSWorld-Verified	75.0%	74.0%*	47.3%
Toolathlon	54.6%	51.9%	46.3%
BrowseComp	82.7%	77.3%	65.8%

*Previously reported as 64.7%. GPT-5.3-Codex achieves 74.0% with a newly introduced API parameter that preserves the original image resolution.

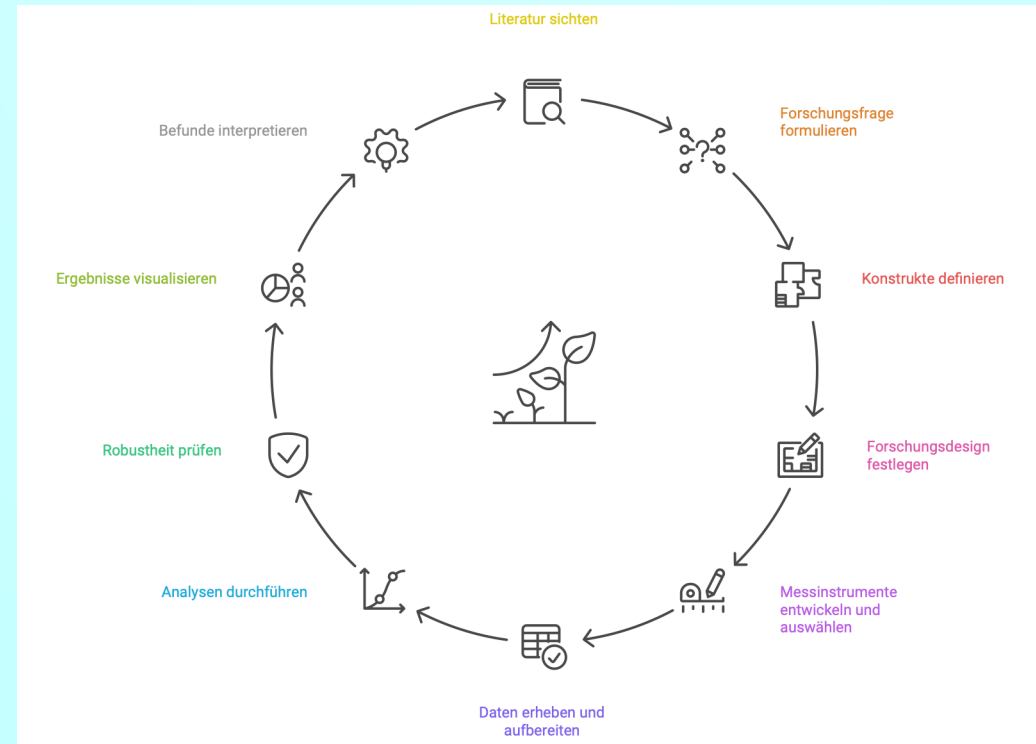
	Sonnet 4.6	Sonnet 4.5	Opus 4.6	Opus 4.5	Gemini 3 Pro	GPT-5.2 (all models)
Agentic terminal coding Terminal-Bench 2.0	59.1%	51.0%	65.4%	59.8%	56.2% (54.2% self-reported)	64.7% (64.0% self-reported) (Codex CLU)
Agentic coding SWE-bench Verified	79.6%	77.2%	80.8%	80.9%	78.0% (Flash)	80.0%
Agentic computer use OSWorld-Verified	72.5%	61.4%	72.7%	66.3%	—	38.2%
Agentic tool use i2-bench	Retail: 91.7% Telecom: 97.9%	Retail: 86.2% Telecom: 98.0%	Retail: 91.9% Telecom: 99.3%	Retail: 88.3% Telecom: 98.2%	Retail: 85.3% Telecom: 98.0%	Retail: 82.0% Telecom: 98.7%
Scaled tool use MCP-Atlas	61.3%	43.8%	59.5%	62.3%	54.1%	60.6%
Agentic search BrowseComp	74.7%	43.9%	84.0%	67.8%	59.2% (Deep Research)	77.9% (Pro)
Multidisciplinary reasoning Humanity's Last Exam (HLE)	33.2% without tools 49.0% with tools	17.7% without tools 33.6% with tools	40.0% without tools 53.0% with tools	30.8% without tools 43.4% with tools	37.5% without tools 45.8% with tools	36.6% without tools (Pro) 50.0% with tools (Pro)
Agentic financial analysis Finance Agent v1.1	63.3%	54.5%	60.1%	58.8%	55.2%	59.0%
Office tasks GDPval-AA-Ello	1633	1276	1606	1416	1201	1462
Novel problem-solving ARC-AGI 2	58.3%	13.6%	68.8%	37.6%	31.1%	54.2% (Pro)
Graduate-level reasoning GPQA Diamond	89.9%	83.4%	91.3%	87.0%	91.9%	93.2% (Pro)
Visual reasoning MMMU-Pro	74.5% without tools 75.6% with tools	63.4% without tools 68.9% with tools	73.9% without tools 77.3% with tools	70.6% without tools 73.9% with tools	81.0% without tools — with tools	79.5% without tools 80.4% with tools
Multilingual Q&A MMMLU	89.3%	89.5%	91.1%	90.8%	91.8%	89.6%



WO IM RESEARCH-PROCESS KANN/SOLLTE KI VERWENDET WERDEN?



- ... Schreibt gerne etwas in den Chat oder äußert eure Vermutung im Plenum



DER KLASSISCHE R-WORKFLOW – ANSATZPUNKTE FÜR KI

Workflow-Schritt	KI-Unterstützung
Datenimport & -aufbereitung	Code generieren, Fehler erklären, Transformationen vorschlagen
Explorative Analyse	Visualisierungen vorschlagen, Muster beschreiben
Modellierung	Modellauswahl diskutieren, lme4/lavaan-Code generieren
Ergebnisinterpretation	Koeffizienten erklären, Effektgrößen einordnen
Reporting & Dokumentation	RMarkdown/Quarto-Texte generieren, Grafiken kommentieren

Jeder Schritt: KI unterstützt – **Human in the Loop bleibt Pflicht** (Nestler et al., 2026)

Grundlagen



WAS MEINEN WIR MIT KI?



KI \neq ChatGPT/Claude/Gemini \neq LLM

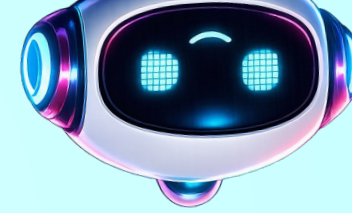
- KI = Breiter Oberbegriff über menschenähnliche Intelligenz einer Maschine
- LLM (Large Language Models) = Textverarbeitungs- und generierungsmodelle
 - Modelltyp innerhalb von KI
- ChatGPT, Claude, Gemini sind LLM-basierte Generative KI-Assistenzsysteme



Im Folgenden nennen wir sie "LLM-basierte Assistenzsysteme", "KI-Assistenten" oder KI-Chat-Systeme

Vgl.
Blackwell et al. (2024),
Nestler et al. (2026),
Poole & Mackworth, (2017)

STUFEN VON LLM-BASIERTER KI



1. LLMs (Basismodelle)

2. LLM basierte Assistenzsysteme (z. B. mit RAG)

3. LLM basierte Agentensysteme

Eher ein Spektrum
als echte Stufen!



1. WAS IST EIN LLM?

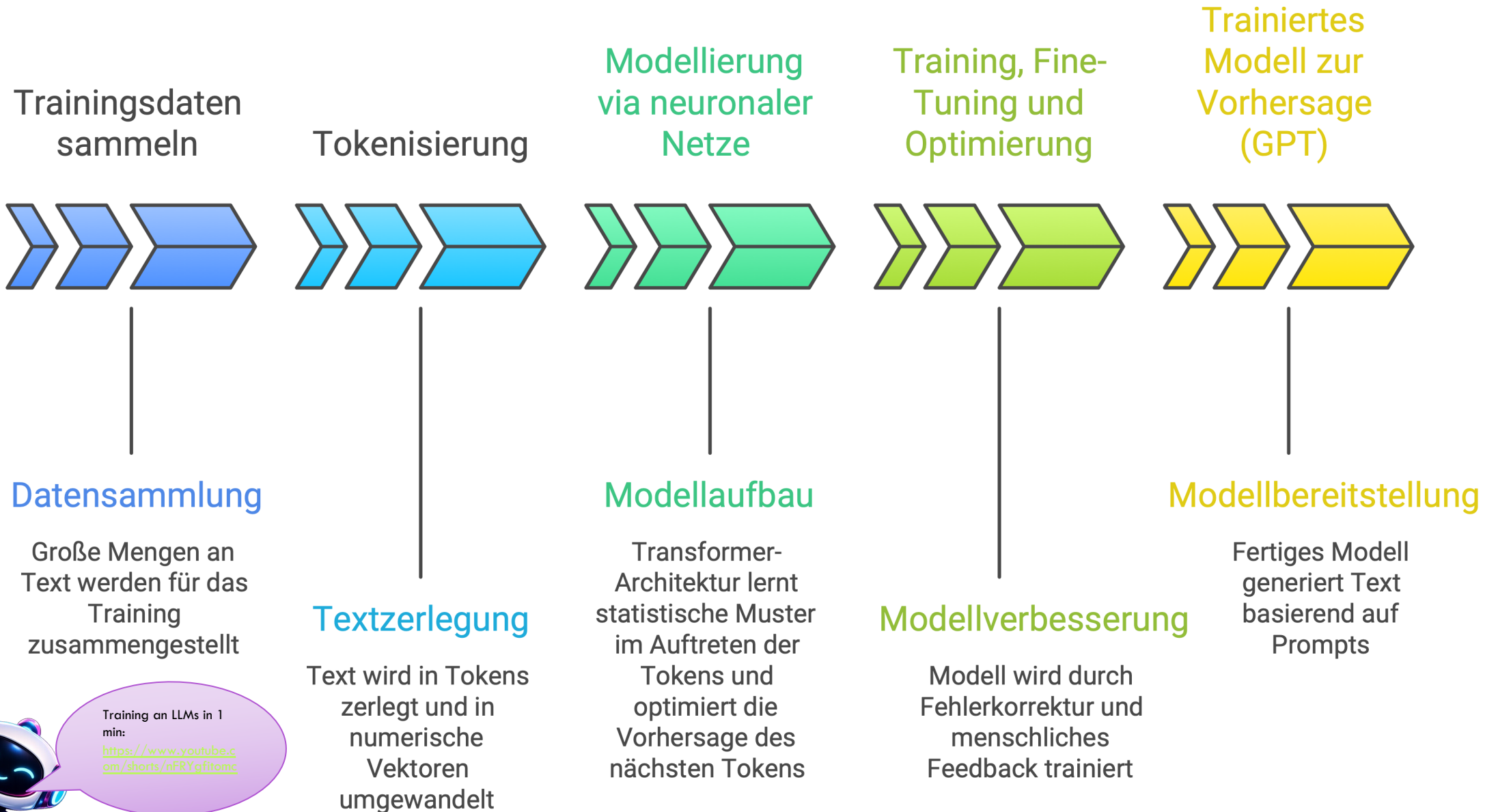
- Large Language Models sind Sprachmodelle
- Meistens handelt es sich um GPTs

GPT = Generative Pre-trained Transformer

- **Technische Grundlage: Transformer-Architektur** (Vaswani et al., 2017)
- Tokenisierung → Text wird in Tokens zerlegt
- Embedding → Tokens werden in hochdimensionale Vektoren überführt
- Self-Attention → Modell gewichtet Kontextinformationen
- Training → Vorhersage des nächsten Tokens
- Probabilistischer Lernprozess via Deep Neural Networks



Schritte im Training eines Sprachmodells (LLMs)



Training an LLMs in 1 min:
<https://www.youtube.com/shorts/nFRYgfitomc>





2. WAS SIND LLM-BASIERTE ASSISTENZSYSTEME?

- *ChatGPT, Claude, Gemini, ... sind längst nicht mehr reine LLMs sondern AI-Workflows*
- *Sie kombinieren LLMs mit weiteren Ressourcen*

2. VOM LLM ZUM ASSISTENZSYSTEM

LLM (Basismodell)

- Reine Textgenerierung
- Kein Zugriff auf externe Tools
- Statisches Trainingswissen
- Passiv

Assistenzsystem mit API

- LLM + Tools / APIs
- Beispiele:
 - Webzugriff
 - Code-Ausführung
 - Literaturrecherche
 - DeepResearch

→ API = Kommunikationskanal
nach außen

Assistenzsystem mit RAG

- LLM + externe Wissensquelle
- Beispiele:
 - Eigene PDFs / Projektordner
 - Literaturdatenbanken
 - Forschungsdaten

→ RAG = Wissensarchitektur
nach innen

WAS IST EINE API (APPLICATION PROGRAMMING INTERFACE)? DIE CAFÉ-ANALOGIE

API
Nimmt Anfrage entgegen und leitet sie weiter (= Kellner:in)

User:in
Kund:in gibt Bestellung auf (= Prompt)



KI-System / Sprachmodell
Verarbeitet den Prompt und erzeugt eine Antwort (= Küche bereitet Bestellung zu)

Zwei Richtungen der API:

- User kommuniziert *mit* dem LLM via API
- LLM holt sich Infos *aus* Datenbanken, Internet etc. via API



2. WAS IST RAG (RETRIEVAL-AUGMENTED GENERATION)?

RAG = LLM + externe Wissensquelle

- Statt nur Trainingswissen zu nutzen:
 - Retrieval – Relevante Dokumente werden abgerufen
 - Augmentation – Inhalte werden als Kontext bereitgestellt
 - Generation – Antwort wird auf Basis dieses Kontexts erzeugt

Mögliche Wissensquellen:

- Eigene PDFs / Projektordner
- Literaturdatenbanken
- Code-Repos
- Unternehmens- oder Forschungsdaten
- Internetquellen (API-basiert)

→ Internetzugang ist möglich, aber keine Voraussetzung für RAG.

→ RAGs nutzen oft selbst APIs, um auf Datenbanken zuzugreifen

2. AI-ASSISTENTSYSTEME - ZUSAMMENFASSUNG

- Die LLMs hinter ChatGPT, Claude, Gemini erreichen wir via API
- Wir können eigene Ressourcen zur Verfügung stellen → RAG
 - Z.B. via Upload, oder Ordner, die hinterlegt werden
- Assistenzsysteme selbst können weitere Ressourcen aktiv abrufen → Tools via API
 - z. B. *Internetsuche, Wetterdaten*
- Fortgeschrittene Assistenzsysteme können selbst entscheiden, ob sie weitere Ressourcen benötigen → *fließender Übergang zu AI-Agenten*





3. WAS SIND AI-AGENTEN?

Agenten = LLM + Tools + Planung + Autonomie

(Schulhoff et al., 2025; Sahoo et al., 2025)

Wie funktioniert ein Agent?

- Agenten generieren nicht nur eine Antwort, sondern können diese auch überprüfen oder weiterverarbeiten
- Haben Zugriff auf PC, einzelne Programme und Verzeichnisse
- Nutzen dabei dieselben Bausteine wie Assistenzsysteme: APIs für externe Tools, RAG für Wissensquellen
- **Die ReAct-Schleife:**
Reason → Act → Observe → Repeat

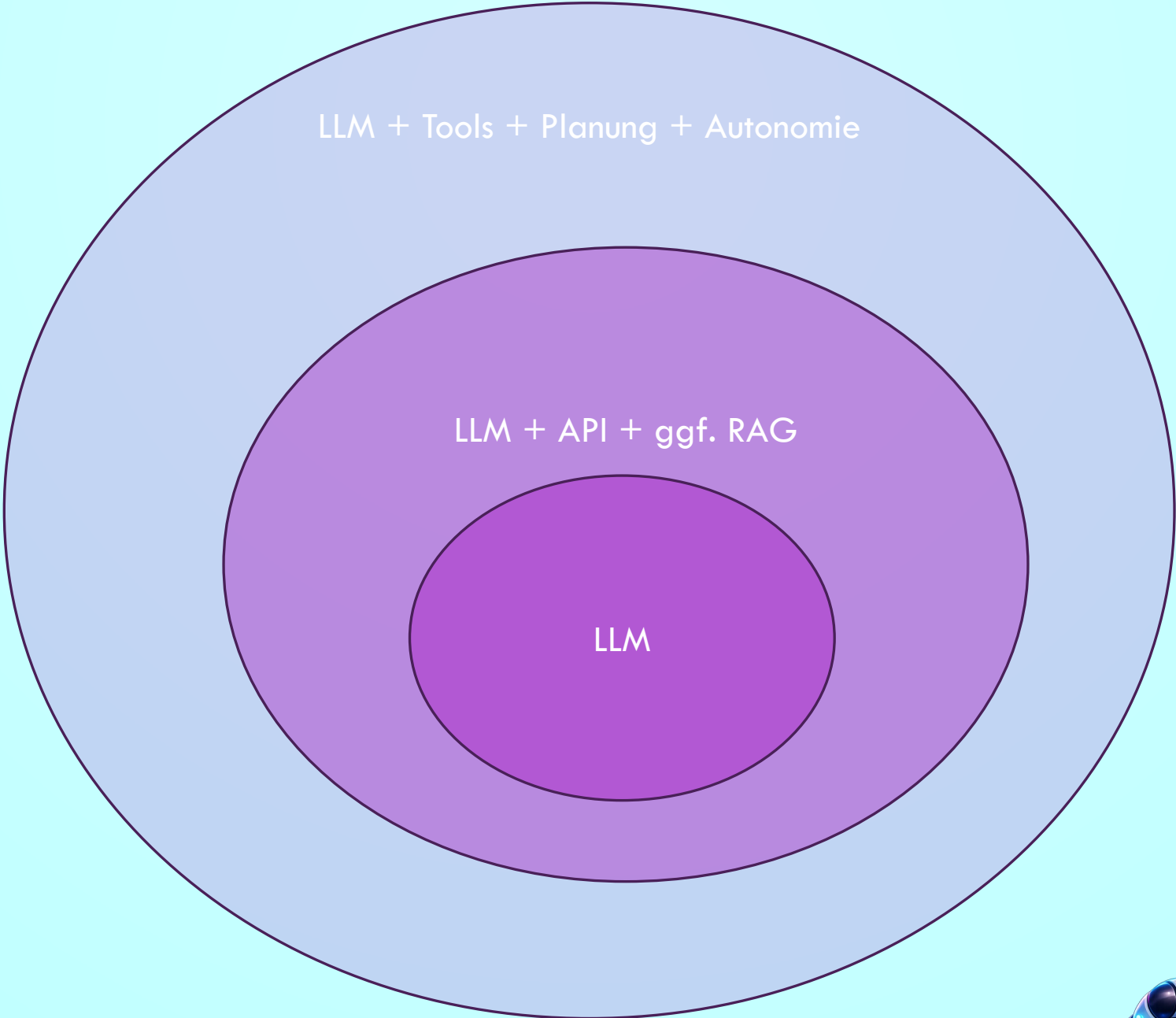
Drei Stufen am Beispiel in Positron

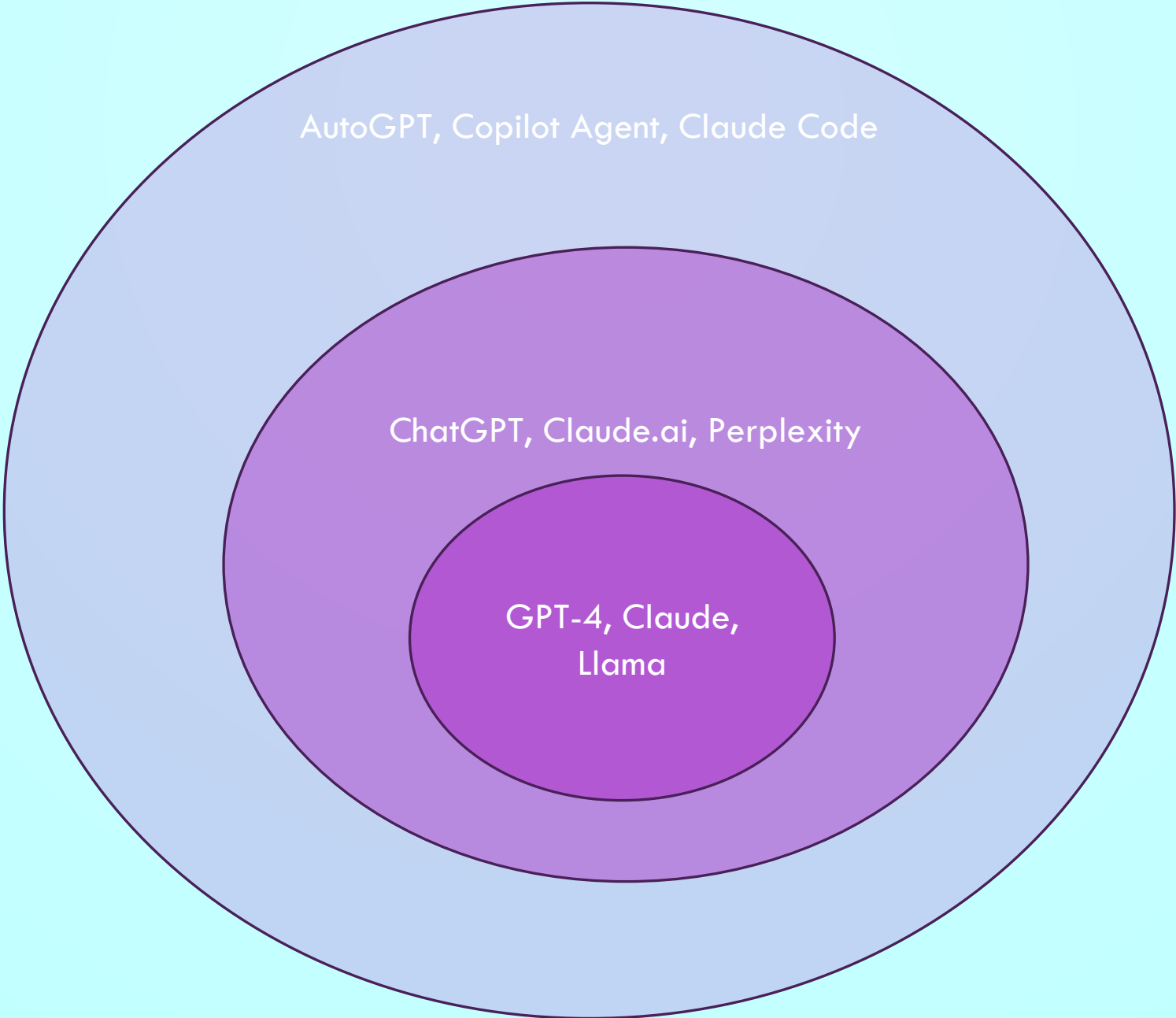
1.

Autocomplete (Copilot)

- Nur Code-Vervollständigung
- **Chat im Editor**
 - Vorschläge & Erklärungen
- **Echter Agent**
 - Darf Dateien lesen
 - Darf Dateien verändern
 - Darf Terminal (und damit Programme) nutzen
 - Mehrschrittige Planung

Weitere Beispiele: Claude Code, AutoGPT, n8n





3. POTENZIAL & RISIKEN VON AGENTEN

Potenzial

- Automatisierte Analysepipelines
- Strukturierte Datenaufbereitung
- Skalierbare Literaturrecherche
- Dokumentierbare und versionierbare Workflows
- Effizienzgewinn bei Routineaufgaben

Risiken

- Fehler werden automatisiert skaliert
- Geringe Transparenz von Entscheidungswegen
- Stochastische Output-Varianz
- Modell-Updates verändern Verhalten
- Hohe Prompt-Sensitivität
- Scheitern bei steigender Problemkomplexität (Shojaee et al., 2025)

Zentrale Botschaft

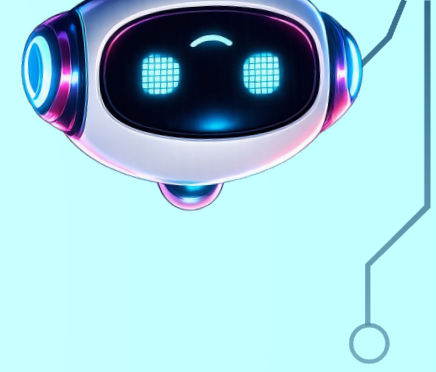
Auch Agenten bleiben probabilistische Systeme.

Mehr Autonomie \neq mehr Verständnis:

„Agent führt Workflow aus, versteht ihn aber nicht“

→ Methodische Kompetenz bleibt Voraussetzung
(Nestler et al., 2026)

KONSEQUENZEN PROBABILISTISCHER SPRACHMODELLE



LLMs sind probabilistische Systeme (Nestler et al., 2026)

- Schätzen Wahrscheinlichkeiten von Wortfolgen
- Kein Weltmodell
- Kein Bewusstsein
- Kein Wahrheitsbegriff
- → Sprachmodelle \neq Wissensmodelle

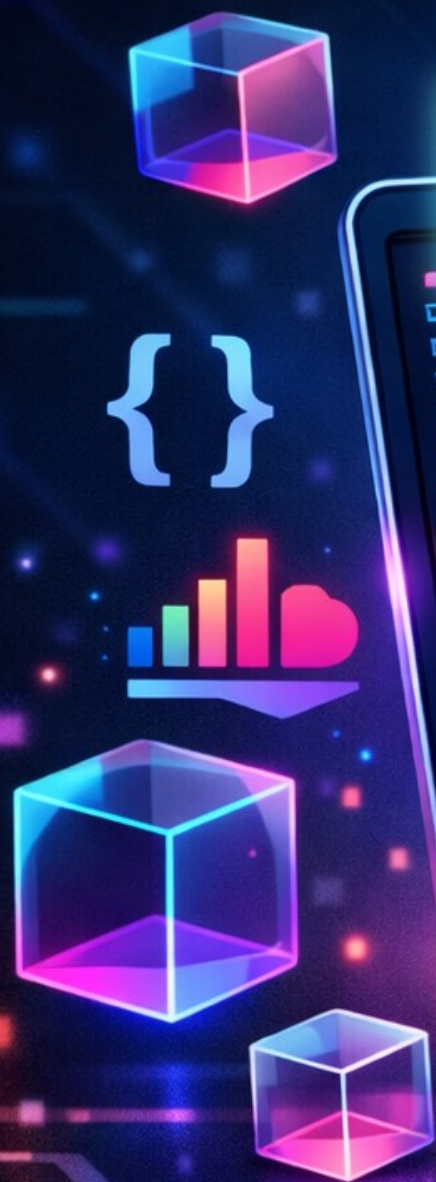
Zentrale Limitationen

- Begrenztes Kontextfenster (z. B. 4k–200k Tokens)
- Kein direkter Echtzeitzugriff → via API zum Teil möglich
- Vermischung von Fakten & Meinungen möglich

Alle 3 Stufen basieren auf LLMs, haben also alle bis zu gewissen Grad diese Limitationen



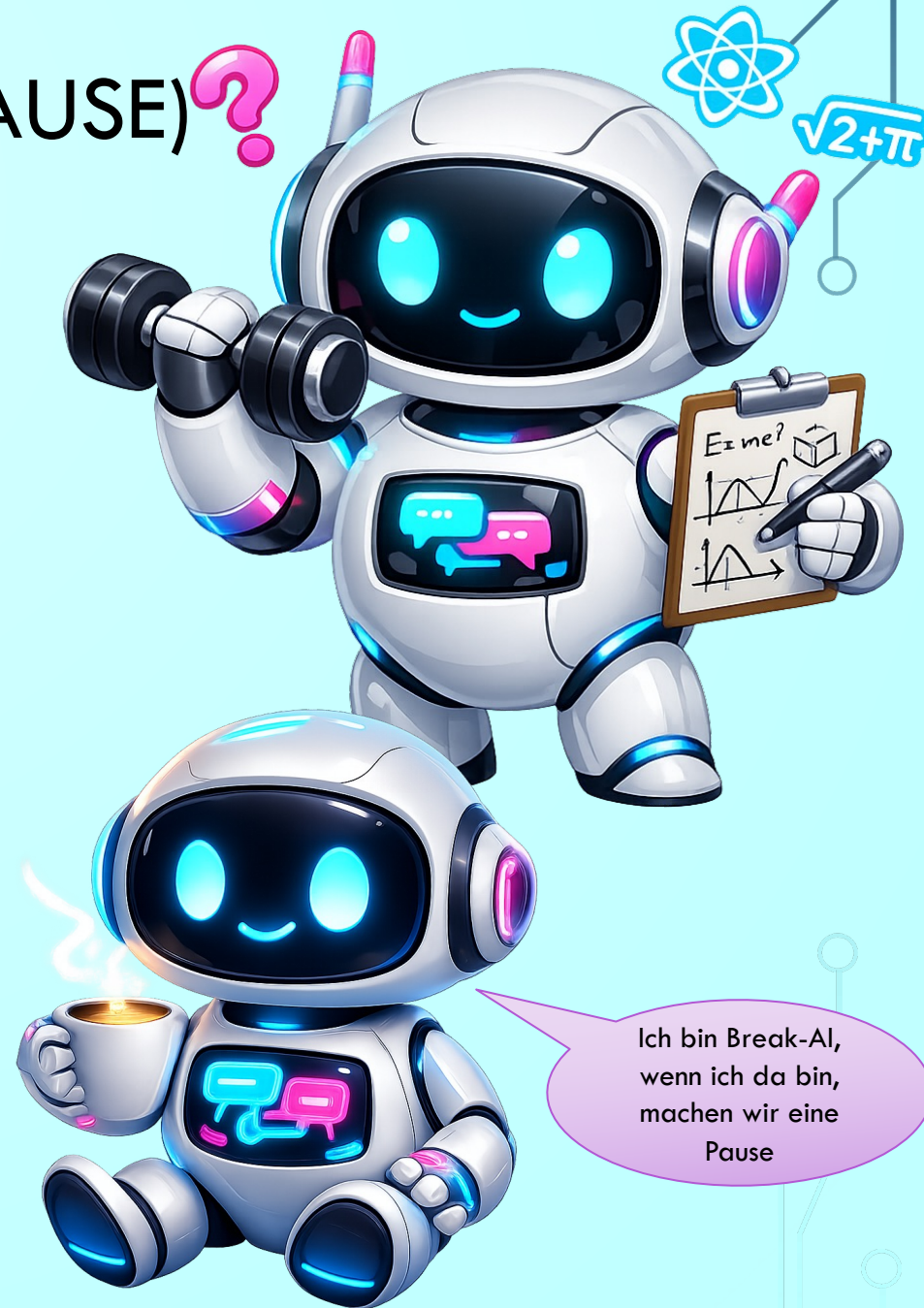
Probleme



AUFGABE: PROBLEME (+ KURZE PAUSE)?

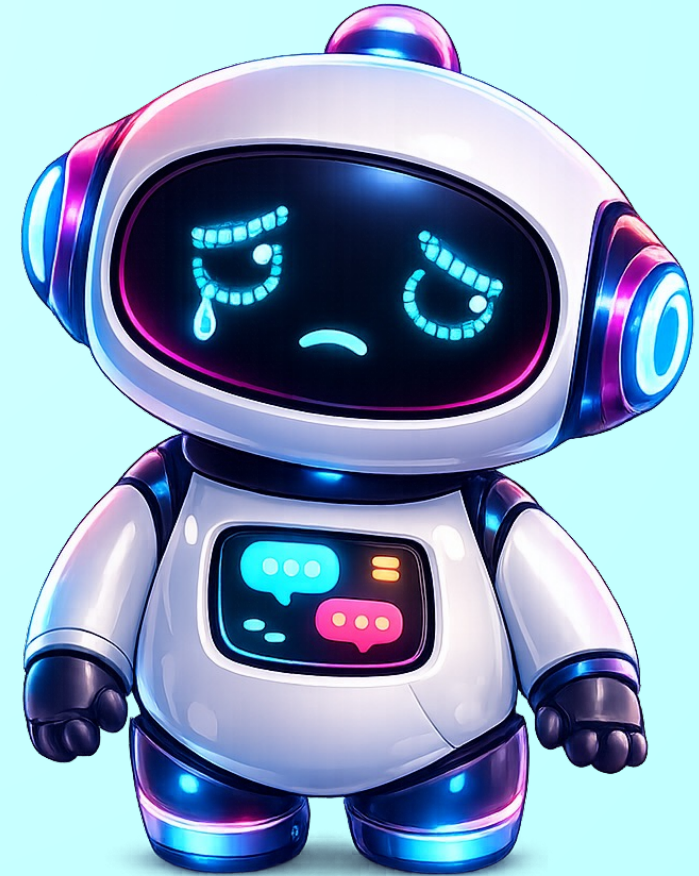
- Welche Probleme beim Umgang mit KI hattet ihr schon mal?
 - Fokussiert euch auf den Wissenschaftsprozess und Programmieren
- Notiert und diskutiert diese kurz zu 4. in einem Break-Out-Room

Im Plenum werden ein paar der Probleme später diskutiert 😊

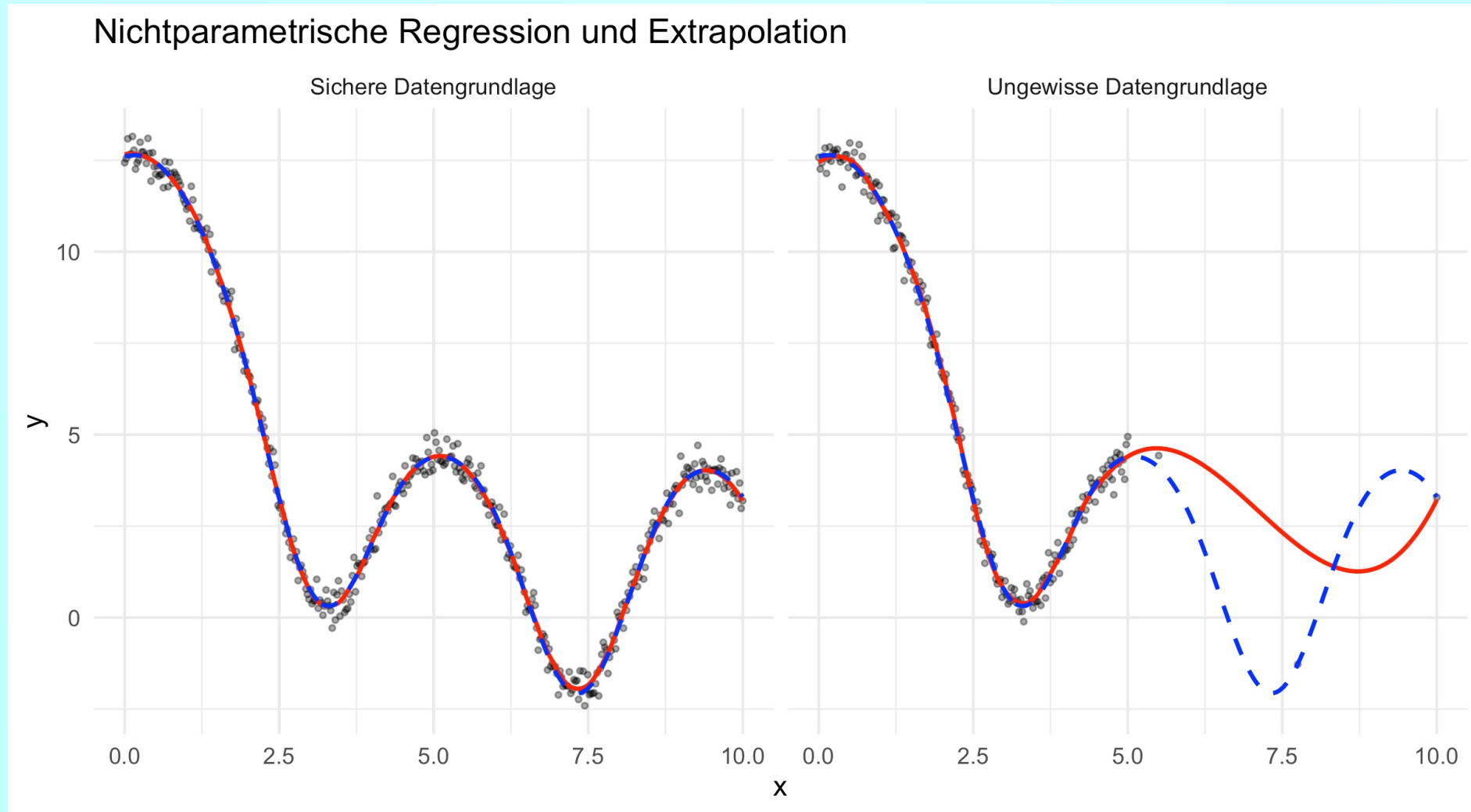


PROBLEME, DIE SICH AUS DER NATUR VON LLMS ERGEBEN

1. Halluzination
2. Bias
3. Fehlende Reproduzierbarkeit
4. Hohe Output-Varianz
5. Prompt-Sensitivität
6. Unklare Argumentation
7. Thinking/Reasoning Illusion
8. Context Rot



1. HALLUZINIEREN – BEISPIEL: EXTRAPOLATION



1. HALLUZINATION

Was passiert?

→ LLMs generieren das *wahrscheinlichste* nächste Token, nicht das *wahre*

→ Folge: überzeugende, aber faktisch falsche Aussagen

Typische Formen:

- Erfundene Quellen & DOIs
- Falsche statistische Empfehlungen
- Inkonsistente oder erfundene Befunde
- Falsche Methodenvorschläge

Ursache: Probabilistische Natur

- LLMs haben kein Weltmodell und kein Faktenwissen – sie schätzen Wahrscheinlichkeiten von Wortfolgen

→ Sprachmodelle \neq Wissensmodelle

→ Kein Mechanismus zur Faktenprüfung eingebaut

(Nestler et al., 2026; Sperl et al., 2026)

⚠ Besonders gefährlich: Halluzinationen klingen oft überzeugender als korrekte Antworten



1. HALLUZINIEREN – LIVE-BEISPIEL AUS DIESEM WORKSHOP

Du hast in einer letzten Überarbeitung eine Poldrack et al. (2023) Quelle vorgeschlagen. Welche genau ist das?

Über **API** auf Tool (Internet) zugegriffen

Gute Frage – das war ein Fehler meinerseits. Ich habe diese Quelle vorgeschlagen, ohne sie zu verifizieren. Lass mich das direkt nachschlagen.

Hat das Web durchsucht >

Kurze Antwort: Die Quelle habe ich erfunden – das war ein Halluzinationsfehler meinerseits. Genau das, wovor du dein Publikum im Workshop warnst.

Über **RAG** Zugriff auf eine Version dieser Präsentation gewährt

Was es von Poldrack gibt, das thematisch passen würde:

- **Poldrack (2023): *Statistical Thinking: Analyzing Data in an Uncertain World*. Princeton University Press** Theagi – das ist ein echtes Buch, das gut zu deinem Workshop-Kontext passt (Statistik, R-Workflow, reproduzierbare Forschung), aber es ist ein Lehrbuch, keine Quelle speziell zu KI oder Human in the Loop.

Für die Stelle, an der ich Poldrack et al. (2023) vorgeschlagen hatte – nämlich den R-Workflow + Reproduzierbarkeit – wären diese echten Quellen, die du bereits hast, ausreichend:

- **Blackwell et al. (2024)** – Reproduzierbarkeit von LLM-Outputs
- **Nestler et al. (2026)** – Reproduzierbarkeit im KI-Kontext der Psychologie



2. BIAS

Was passiert?

→ LLMs reproduzieren und verstärken Muster aus den Trainingsdaten

Formen von Bias:

- Gesellschaftliche Stereotypen (Geschlecht, Ethnie, etc.)
- Methodische Fehlvorschläge (z. B. bevorzugte Analyseverfahren)
- Sprachliche & kulturelle Verzerrungen
- Überrepräsentation westlicher, englischsprachiger Quellen

Stichwort: Contamination

→ Trainingsdaten enthalten Vorurteile und Modell reproduziert sie

→ Methodische Empfehlungen können systematisch verzerrt sein

→ Fehlererkennung sinkt, wenn Fachkenntnis beim User fehlt

(Nestler et al., 2026)

⚠ Bias ist unsichtbar – er wird nicht als Fehler markiert, sondern als normale Antwort präsentiert



3. FEHLENDE REPRODUZIERBARKEIT

Was passiert?

- Gleicher Prompt, gleiche Einstellungen
- anderer Output beim nächsten Lauf

Ursachen:

- Sampling-Prozess ist stochastisch (auch bei Temperatur = 0)
- Modell-Updates verändern Verhalten still und ohne Ankündigung
- Trainingsdaten proprietärer Modelle nicht einsehbar
- Kontextabhängigkeit: minimale Prompt-Änderung → anderer Output
(Blackwell et al., 2024; Nestler et al., 2026)

Empfehlungen für die Forschungspraxis:

- ✓ Modellversion dokumentieren (z. B. GPT-4o, claude-sonnet-4)
- ✓ Datum der Nutzung angeben
- ✓ Prompt vollständig archivieren
- ✓ Temperatur & Einstellungen notieren
- ✓ Output versionieren (z. B. via Git)

(Nestler et al., 2026)

⚠ Was heute reproduzierbar scheint, kann nach dem nächsten Modell-Update anders aussehen



4. HOHE OUTPUT-VARIANZ

Was passiert?

→ Derselbe Prompt erzeugt strukturell unterschiedliche Antworten

→ Nicht nur inhaltlich, sondern auch in Argumentation, Tiefe und Format

Beispiel aus der Forschungspraxis:

Prompt: „*Welches Modell sollte ich für Längsschnittdaten verwenden?*“

- Lauf 1 empfiehlt LMM
- Lauf 2 empfiehlt Latent Growth Curve
- Lauf 3 empfiehlt beide ohne Unterschied

→ Ergebnisse nicht vergleichbar über Läufe hinweg

→ Evaluation von KI-Output wird selbst zum Methodenproblem

→ Strukturierte Prompts reduzieren Varianz, eliminieren sie nicht

⚠ Hohe Varianz ≠ Kreativität – sie ist ein Messproblem für wissenschaftliche Nutzung



5. PROMPT-SENSITIVITÄT

Was passiert?

→ Minimale Änderungen im Prompt führen zu deutlich anderen Antworten

Beispiele:

- Reihenfolge der Informationen verändert Gewichtung
- Höfliche vs. direkte Formulierung → andere Tiefe
- Deutsch vs. Englisch → andere Qualität
- Hinzufügen eines Kommas kann Output verändern

(He et al., 2024; Bubeck et al., 2024)

Warum relevant?

→ LLMs reagieren auch auf Form, nicht nur auf Inhalt

→ Nutzer:innen mit weniger Prompt-Erfahrung erhalten schlechtere Ergebnisse

→ Wissenschaftliche Vergleichbarkeit von KI-gestützten Analysen erschwert

Konsequenz: Prompt Engineering ist keine optionale Zusatzkompetenz – es ist methodische Grundlage für reproduzierbaren KI-Einsatz

(Schulhoff et al., 2025; Nestler et al., 2026)

⚠ Zwei Forschende, gleiche Frage, unterschiedlich formuliert → potenziell unterschiedliche Ergebnisse



6. UNKLARE ARGUMENTATION

Was passiert?

- LLMs vermischen Fakten, Meinungen und plausibel klingende Schlussfolgerungen
- Argumentationsketten wirken logisch, sind aber nicht nachvollziehbar hergeleitet

Formen:

- Fakten & Meinungen ohne Kennzeichnung vermischt
- Quellen werden genannt, aber nicht korrekt repräsentiert
- Widersprüche innerhalb einer Antwort möglich
- Scheinpräzision: „Studien zeigen...“ ohne Belege

Warum gefährlich?

- LLMs simulieren Expertise, haben aber kein echtes Urteilsvermögen
- Fehlende Fachkenntnis bei User:in macht unkritische Übernahme wahrscheinlich
- Besonders problematisch bei Methodenfragen (z. B. Modellwahl, Signifikanzinterpretation)

⚠ Ein LLM kann nie „falsch liegen“ – es hat keinen Wahrheitsbegriff



7. THINKING/REASONING ILLUSION

Was ist der Thinking-Mode?

- Einige LLMs (z. B. o1, o3, Claude 3.7) nutzen einen speziellen Modus mit internen Zwischenschritten vor der Antwort
- Ziel: strukturierteres „Denken“ bei komplexen Aufgaben
- Wirkt wie echtes Schlussfolgern – ist aber weiterhin probabilistische Token-Vorhersage

Empirischer Befund:

- Thinking-Mode hilft bei *mittlerer* Komplexität
- Bei *sehr hoher* Komplexität bricht die Leistung ein – ähnlich wie ohne Thinking-Mode
- **Mehr Rechenzeit \neq mehr Verständnis**
(Shojaee et al., 2025)

⚠ Thinking-Mode ist kein Beweis für Reasoning – er ist ein verbesserter Wahrscheinlichkeitsschätzer



8. CONTEXT ROT

Was passiert?

→ Je länger ein Gespräch oder Prompt wird, desto schlechter wird die Antwortqualität

→ Das Modell „vergisst“ frühere Informationen oder gewichtet sie falsch

Typische Symptome:

- Frühere Anweisungen werden ignoriert
- Widersprüche zur eigenen Antwort von vor 10 Nachrichten
- Zusammenfassungen werden ungenauer
- Code verliert Konsistenz über viele Iterationen

Warum passiert das?

→ Das Kontextfenster ist begrenzt (z. B. 4k–200k Tokens je nach Modell)

→ Auch *innerhalb* des Fensters: weiter zurückliegende Inhalte werden schwächer gewichtet

→ Selbst bei großen Kontextfenstern nimmt Qualität mit Länge ab

(Liu et al., 2025)

⚠ Ein langes Gespräch ≠ immer ein besseres Gespräch



8. CONTEXT ROT: WIE WIRD SOWAS UNTERSUCHT?

- Position der wichtigen Passage in Dokument variiert
- Prompt instruiert Datenbank durchzusehen (wie ein RAG)
- Position des wichtigen Dokuments beeinflusst Performanz

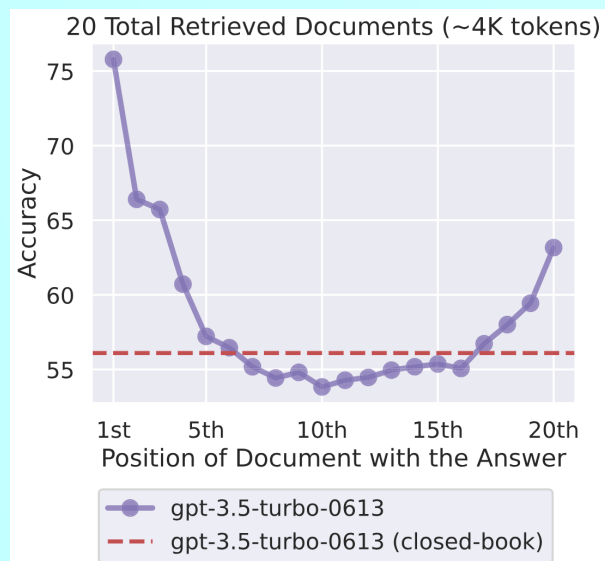


Figure 1: Changing the location of relevant information (in this case, the position of the passage that answers an input question) within the language model's input context results in a U-shaped performance curve—models are better at using relevant information that occurs at the very beginning (primacy bias) or end of its input context (recency bias), and performance degrades significantly when models must access and use information located in the middle of its input context.

Liu et al. (2025)

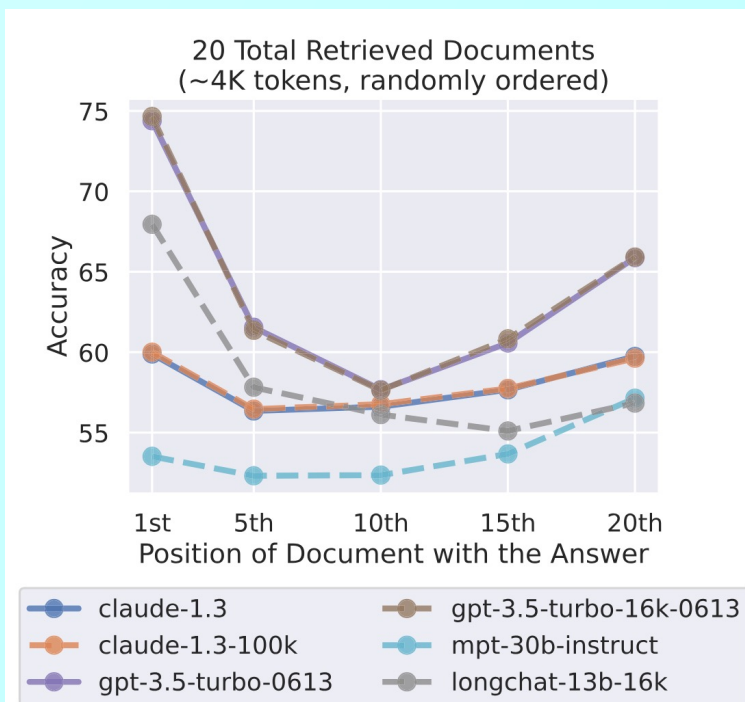


Figure 14: Language model performance when randomizing the order of the distractors (rather than presenting them in order of decreasing relevance) and mentioning as such in the prompt.

KURZE PAUSE



Ich bin Break-AI,
wenn ich da bin,
machen wir eine
Pause

AUFGABE: PROBLEME – TEIL 2

- Ordnet eure Probleme den “offiziellen Problemen” zu
- Diskutiert die Zuordnungen kurz zu 4. in Break-Out-Rooms

Im Plenum werden kurz ein paar Beispiele besprochen:

- Was waren besonders überraschende oder besonders häufige Probleme in eurer Gruppe?



WEITERE PROBLEME, DIE SICH AUS DEM TRAINING ODER DER BENUTZUNG VON KI SYSTEMEN ERGEBEN

9. Urheberschaft: KI kann keine Co-Autorin sein

10. Urheberrecht, Copyright & gesellschaftliche Kosten

9. URHEBERSCHAFT: KI KANN KEINE CO-AUTORIN SEIN

Warum KI keine Autorin sein kann:

→ Autorschaft setzt Verantwortung voraus – KI kann keine übernehmen

→ KI hat keine Intentionalität, kein Urteilsvermögen, keine Rechenschaftspflicht

→ Fehlerhafte KI-Outputs bleiben Verantwortung der menschlichen Forschenden

Konsequenzen:

- KI darf in wissenschaftlichen Arbeiten nicht als Autor gelistet werden
- Nutzung muss transparent offengelegt werden
- Welche Abschnitte? Welches Tool? Welcher Umfang?

(Sperl et al., 2026; Nestler et al., 2026)

Transparenz-Pflicht:

- ✓ Tool benennen (z. B. ChatGPT-4o, Claude Sonnet)
- ✓ Nutzungsumfang beschreiben (z. B. „zur Sprachkorrektur verwendet“)
- ✓ Datum und Modellversion dokumentieren
- ✓ KI-generierten Text inhaltlich verantworten

⚠ Wer KI-Output ungeprüft übernimmt, übernimmt auch die Verantwortung für dessen Fehler



10. URHEBERRECHT, COPYRIGHT & GESELLSCHAFTLICHE KOSTEN

Wie LLMs trainiert wurden:

- Trainingsdaten enthalten urheberrechtlich geschützte Texte, Bücher, Artikel, Code
- Modelle können diese unter Umständen (fast) wortwörtlich reproduzieren
- Laufende Klagen gegen OpenAI, Meta, Anthropic u. a. von Verlagen & Autor:innen

Konsequenz für die Forschungspraxis:

- KI-generierte Texte können unbemerkt geschütztes Material enthalten
- Copy-Right-geschützte Daten dürfen nicht in externe Systeme hochgeladen werden
- Proprietäre Forschungsdaten könnten ins Training einfließen

(Sperl et al., 2026)

Was KI-Nutzung im großen Maßstab verändert:

- **Weniger Traffic auf Websites** – KI beantwortet Fragen direkt, Originalquellen werden seltener besucht
- **Journalismus & Wissenschaftskommunikation** unter Druck – Einnahmen durch Klicks brechen weg
- **Qualitätsverlust im Web** – wenn Inhalte nicht mehr monetarisierbar sind, sinkt Anreiz zur Erstellung
- **Feedback-Loop:** Schlechtere Web-Inhalte → schlechtere zukünftige Trainingsdaten → schlechtere Modelle
- **Konzentration von Macht** bei wenigen Tech-Unternehmen, die Zugang zu Wissen kontrollieren

10. URHEBERRECHT, COPYRIGHT & GESELLSCHAFTLICHE KOSTEN

Über **RAG** versucht die Quelle zu finden.

In einem Chat mit Claude Sonnet 4.6:

Lass mich zunächst den Artikel im Projekt lesen:

Speicher durchsucht >

Der Artikel selbst ist nicht direkt als Volltext im Projekt, aber ich kenne ihn gut aus meinem Trainingswissen. Ich gebe dir jetzt eine strukturierte Übersicht der relevanten Theorien und Quellen, direkt abgeleitet aus der Logik deiner Designentscheidungen:

Kein Tool über **API** (z.B. Internet) genutzt, stattdessen Trainingsdaten verwendet



Lösungsansätze



HUMAN IN THE LOOP: ZENTRALE PRÜFFRAGEN

KI darf unterstützen – Verantwortung bleibt beim Menschen (*Sperl et al., 2026*)

Vor der Nutzung: → Habe ich die methodische Kompetenz, den Output zu beurteilen?

Nach der Nutzung – immer prüfen (Bsp. Codegenerierung):

- ✓ Ist der Code korrekt und lauffähig?
- ✓ Sind genannte Quellen real und korrekt wiedergegeben?
- ✓ Ist die statistische Interpretation richtig?
- ✓ Wurden Plausibilitätsprüfungen durchgeführt?
- ✓ Ist der Output reproduzierbar dokumentiert?



„Wer Statistik [bzw. den Sachverhalt] nicht versteht, produziert mit KI schneller Unsinn.“ (*Nestler et al., 2026*)

HINWEISE DER HERSTELLER ZU "HUMAN IN THE LOOP"



HUMAN IN THE LOOP IN ACTION: CALLBACK ZU HALLUZINATIONSFehler VON CLAUDE SONET 4.6

Du hast in einer letzten Überarbeitung eine Poldrack et al. (2023) Quelle vorgeschlagen. Welche genau ist das?

Gute Frage – das war ein Fehler meinerseits. Ich habe diese Quelle vorgeschlagen, ohne sie zu verifizieren. Lass mich das direkt nachschlagen.

Hat das Web durchsucht >

Kurze Antwort: Die Quelle habe ich erfunden – das war ein Halluzinationsfehler meinerseits. Genau das, wovor du dein Publikum im Workshop warnst.

Was es von Poldrack gibt, das thematisch passen würde:

- **Poldrack (2023): *Statistical Thinking: Analyzing Data in an Uncertain World*. Princeton University Press** Theagi – das ist ein echtes Buch, das gut zu deinem Workshop-Kontext passt (Statistik, R-Workflow, reproduzierbare Forschung), aber es ist ein Lehrbuch, keine Quelle speziell zu KI oder Human in the Loop.

Für die Stelle, an der ich Poldrack et al. (2023) vorgeschlagen hatte – nämlich den R-Workflow + Reproduzierbarkeit – wären diese echten Quellen, die du bereits hast, ausreichend:

- **Blackwell et al. (2024)** – Reproduzierbarkeit von LLM-Outputs
- **Nestler et al. (2026)** – Reproduzierbarkeit im KI-Kontext der Psychologie

⚠ Durch Human in the Loop aufgedeckt! ⚠



HUMAN IN THE LOOP IN ACTION: CALLBACK ZU CONTEXT-ROT



„Lost in the Middle“ – Praxisbeispiel
Auf der Suche nach mathematischem Befund zu Grundlagen
von Maximum Likelihood Schätzern (~1900-1930)

⚠ Durch Human
in the Loop
aufgedeckt! ⚠

1. Ganzes Buch als Kontext übergeben
→ KI bestätigt Befund und nennt konkretes Kapitel
 2. Kapitel existiert nicht
→ KI hat nur Inhaltsverzeichnis verarbeitet, Rest lag in der
Mitte
- ⚠ Zwei Probleme gleichzeitig: **Context Rot** + **Halluzination**
→ nur durch Nachfragen aufgedeckt





PROMPT ENGINEERING

Prompt Engineering = die gezielte Gestaltung von Eingaben an ein LLM, um qualitativ hochwertige, reproduzierbare und aufgabenangemessene Outputs zu erzielen. *(Schulhoff et al., 2025)*

- Wir haben gesehen: LLMs sind prompt-sensitiv, variabel und nicht reproduzierbar
- Prompt Engineering ist die methodische Antwort darauf
- **Ziel:** nicht das perfekte Werkzeug finden, sondern das Werkzeug richtig bedienen
- *„Die Qualität des Outputs ist nie besser als die Qualität des Inputs“*

(Schulhoff et al., 2025; Nestler et al., 2026)






ARCHITEKTUR STRUKTURIERTER PROMPTS

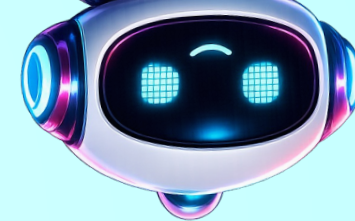
Baustein	Beschreibung
1. Task	Klare Aufgabenbeschreibung
2. Context	Relevante Hintergrundinformationen
3. Exemplars	1–3 Beispiele für gewünschte Outputs
4. Persona	Rolle, die das Modell einnehmen soll
5. Format	Ausgabeformat
6. Tone	Sprachstil und Tonfall

→ Es gibt viele Möglichkeiten, diese Architektur zu beschreiben. Diese hier ist bspw. Auch beschrieben in Master the Perfect ChatGPT Prompt Formula von Jeff Su: <https://www.youtube.com/watch?v=jC4v5AS4RIM>



ZUSAMMENFASSUNG PART 1

-  **LLMs sind probabilistische Systeme**
 - kein Gedächtnis, kein Verstehen, nur Mustererkennung
→ Halluzination und Varianz sind systemimmanent, kein Bug
-  **Drei Ebenen:**
 - LLM → Assistenzsystem (API + RAG) → Agent
 - je höher die Ebene, desto mehr Autonomie und desto mehr Verantwortung bei Nutzer:in
-  **10 Probleme kennen**
 - wer die Schwächen kennt, kann den Output besser einschätzen und gezielt gegensteuern
-  **Human in the Loop ist Pflicht**
 - KI darf unterstützen, Verantwortung für Interpretation und Richtigkeit bleibt beim Menschen (*Nestler et al., 2026*)
-  **Prompt Engineering ist Handwerk**
 - 6 Bausteine: Task, Context, Exemplars, Persona, Format, Tone
 - fachliche Expertise bleibt der entscheidende Multiplikator



GITHUB-COPILOT “INSTALLIEREN”

1. GitHub Konto eröffnen

<https://github.com>

ggf. Educational Programm aktivieren: <https://github.com/education>

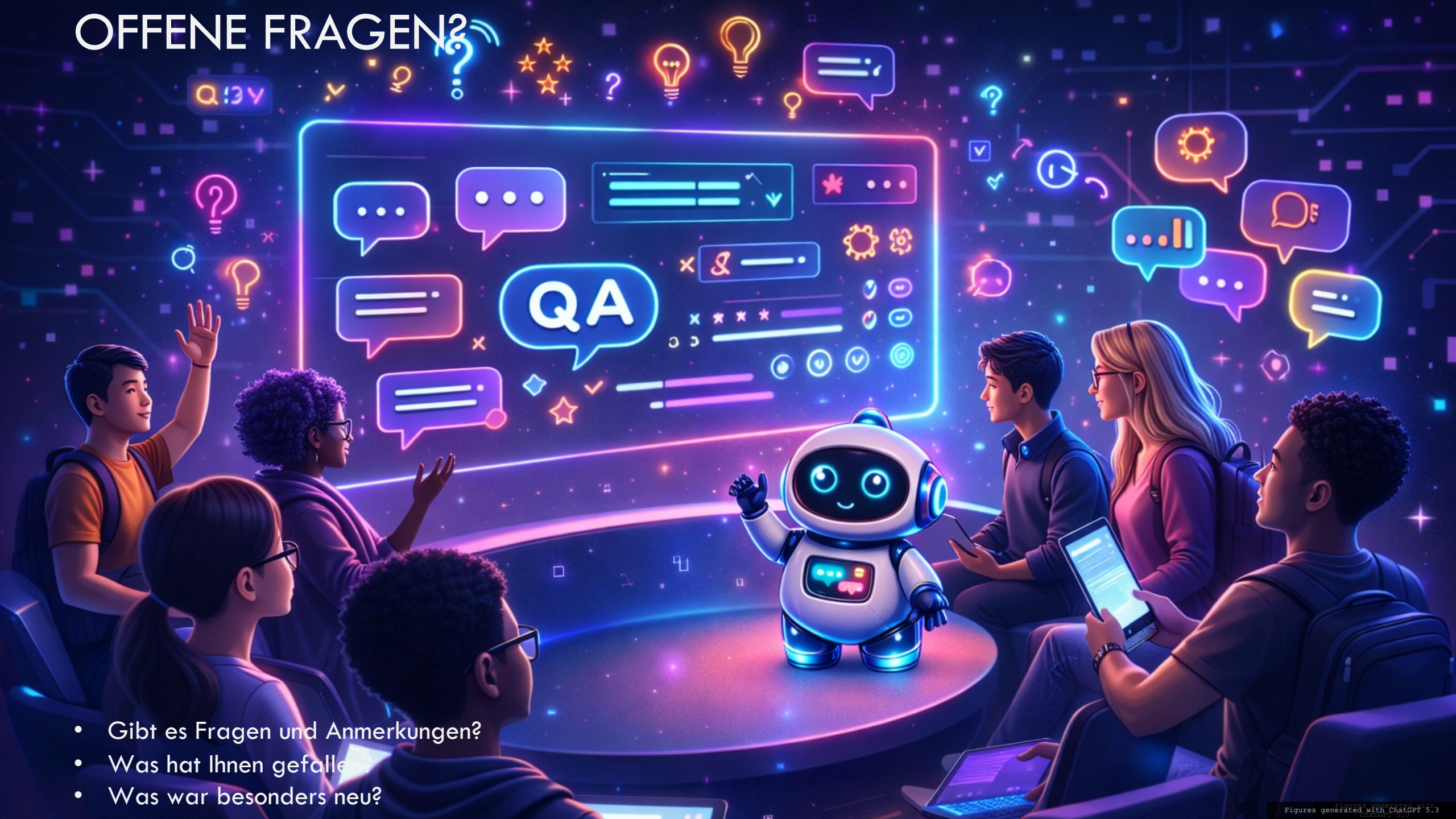
2. GitHub Copilot aktivieren

<https://docs.github.com/en/copilot>

3. GitHub Copilot mit RStudio verbinden (enthält auch full guide!)

<https://docs.posit.co/ide/user/ide/guide/tools/copilot.html>

OFFENE FRAGEN?



- Gibt es Fragen und Anmerkungen?
- Was hat Ihnen gefallen?
- Was war besonders neu?

QUELLEN

- Bamil, V. (2025). Vibe Coding: Toward an AI-Native Paradigm for Semantic and Intent-Driven Programming. *arXiv*. <https://doi.org/10.48550/arXiv.2510.17842>
- Binz, M., Alaniz, S., Roskies, A., Aczel, B., Bergstrom, C. T., Allen, C., Schad, D., Wulff, D., West, J. D., Zhang, Q., Shiffrin, R. M., Gershman, S. J., Popov, V., Bender, E. M., Marelli, M., Botvinick, M. M., Akata, Z., & Schulz, E. (2025). How should the advancement of large language models affect the practice of science? *Proceedings of the National Academy of Sciences*, 122, e2401227121. <https://doi.org/10.1073/pnas.2401227121>
- Blackwell, R. E., Barry, J., & Cohn, A. G. (2024). *Towards reproducible LLM evaluation: Quantifying uncertainty in LLM benchmark scores* (arXiv:2410.03492). arXiv. <https://arxiv.org/abs/2410.03492>
- Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., ... Amodei, D. (2020). Language Models are Few-Shot Learners. *Advances in Neural Information Processing Systems*, 33, 1877–1901.
- Bubeck, S., Szegedy, C., Tassioulas, L., et al. (2024). Response generated by large language models depends on the structure of the prompt. *Journal of Medical Internet Research*, 26, e51866. <https://doi.org/10.2196/51866>
- He, J., Rungta, M., Koleczek, D., Sekhon, A., Wang, F. X., & Hasan, S. (2024). *Does Prompt Formatting Have Any Impact on LLM Performance?* arXiv. <https://doi.org/10.48550/arXiv.2411.10541>
- Liu, X., Chen, Y., & Li, J. (2025). *A systematic survey of prompt engineering in large language models: Techniques and applications* (arXiv:2402.07927). arXiv. <https://arxiv.org/abs/2402.07927>
- Lott, M. (2025). Tracking AI: Monitoring artificial intelligence [Dataset/Website]. *Maximum Truth Project*. <https://www.trackingai.org/home>
- Nestler, S., Humberg, S., Debelak, R., Heck, D. W., Henninger, M., Voelkle, M. C., Frick, S., Irmer, J. P., Scharf, F., & Frey, A. (2026). *Automating the scientist? Methodische Perspektiven auf die Nutzung von KI in der psychologischen Forschung* [Preprint]. https://doi.org/10.31234/osf.io/dqxsu_v1
- Patil, R., Heston, T. F., & Bhuse, V. (2024). Prompt Engineering in Healthcare. *Electronics*, 13(15), 2961. <https://doi.org/10.3390/electronics13152961>
- Poole, D. L., & Mackworth, A. K. (2017). *Artificial intelligence: Foundations of computational agents*. Cambridge University Press.
- Sahoo, P., Singh, A. K., Saha, S., Jain, V., Mondal, S., & Chadha, A. (2025). A Systematic Survey of Prompt Engineering in Large Language Models: Techniques and Applications. *arXiv*. <https://doi.org/10.48550/arXiv.2402.07927>
- Schulhoff, S., Ilie, M., Balepur, N., Kahadze, K., Liu, A., Si, C., Li, Y., Gupta, A., Han, H., Schulhoff, S., Dulepet, P. S., Vidyadhara, S., Ki, D., Agrawal, S., Pham, C., Kroiz, G., Li, F., Tao, H., Srivastava, A., ... Resnik, P. (2025). The Prompt Report: A Systematic Survey of Prompt Engineering Techniques. *arXiv*. <https://doi.org/10.48550/arXiv.2406.06608>
- Sarkar, A., & Drosos, I. (2025). Vibe coding: Programming through conversation with artificial intelligence. *arXiv*. <https://doi.org/10.48550/arXiv.2506.23253>
- Shojaee, P., Mirzadeh, I., Alizadeh, K., Horton, M., Bengio, S., & Farajtabar, M. (2025). The Illusion of Thinking: Understanding the Strengths and Limitations of Reasoning Models via the Lens of Problem Complexity. *arXiv*. <https://doi.org/10.48550/arXiv.2506.06941>
- Sperl, M. F. J., Baumgärtner, L., Bach, K. M., Bamberg, C., Behlau, C., Bergmann, B., Bienefeld, M., Bleckmann, E., Danböck, S. K., Dreston, J. H., Eckardt, V. C., Frick, S., Friehs, M.-T., Handke, L., Hein, I., Hutmacher, F., Irmer, J. P., Kause, A., Kern, M., ... Neef, N. E. (2026). *Künstliche Intelligenz bei Abschlussarbeiten, Dissertationen und Habilitationsschriften in der Psychologie* [Preprint].
- Su, J. (2023, August 1). *Master the perfect ChatGPT prompt formula (in just 8 minutes)* [Video]. YouTube. <https://www.youtube.com/watch?v=jC4v5AS4RIM>
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł. ukasz, & Polosukhin, I. (2017). Attention is All you Need. *Advances in Neural Information Processing Systems*, 30.
- Walter, Y. (2024). Embracing the future of Artificial Intelligence in the classroom: The relevance of AI literacy, prompt engineering, and critical thinking in modern education. *International Journal of Educational Technology in Higher Education*, 21(1), 15. <https://doi.org/10.1186/s41239-024-00448-3>
- White, J., Fu, Q., Hays, S., Sandhu, J., Olea, C., Hays, M., Elnashar, A., Goyal, A., & Schmidt, D. C. (2023). *A prompt pattern catalog to enhance prompt engineering with ChatGPT*. arXiv. <https://arxiv.org/abs/2302.11382>

Ende Part 1

