Künstliche Intelligenz

Wie soll die generative KI effizient eingesetzt werden ?

1

Unsere Kinder werden den Zusammenhang nie verstehen

2

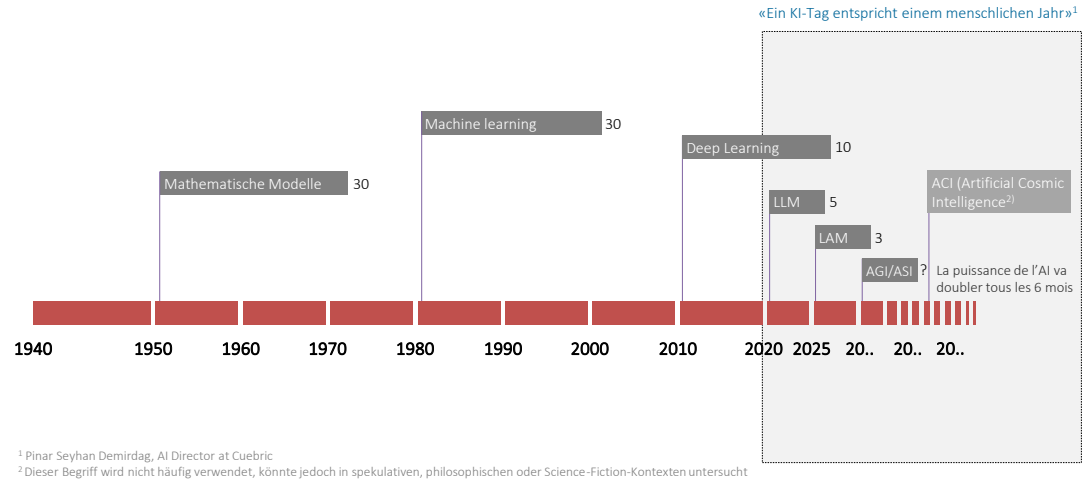
3

Agenda

1. Die Beschleunigung im Zeitalter der KI
2. Wie funktionieren künstliche Intelligenz und LLMs?
3. Acht Einschränkungen und deren Auswirkungen auf die Nutzung
4. Fazit

4

Künstliche Intelligenz beschleunigt die Entwicklung in allen Bereichen unseres Lebens in beispielloser Weise.

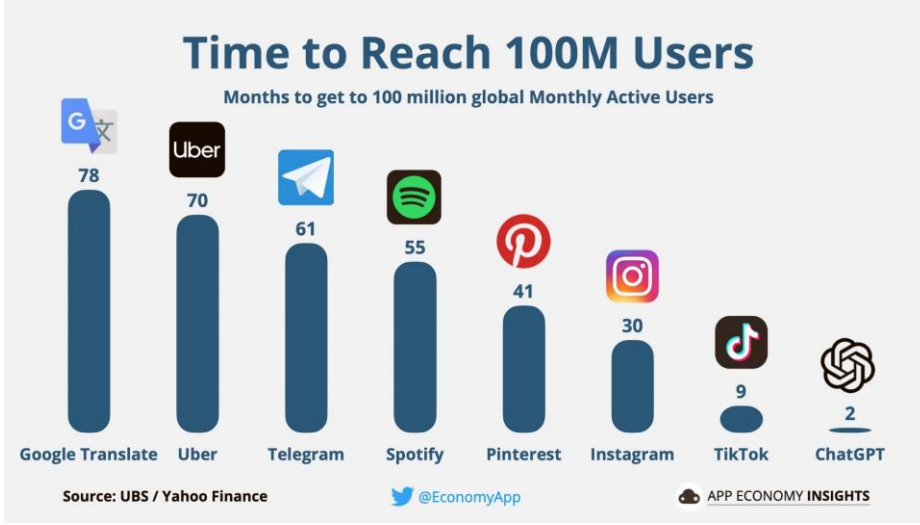


[1] Pinar Seyhan Demirdag, AI Director at Cuebric
[2] Dieser Begriff wird nicht häufig verwendet, könnte jedoch in spekulativen, philosophischen oder Science-Fiction-Kontexten untersucht werden.

5

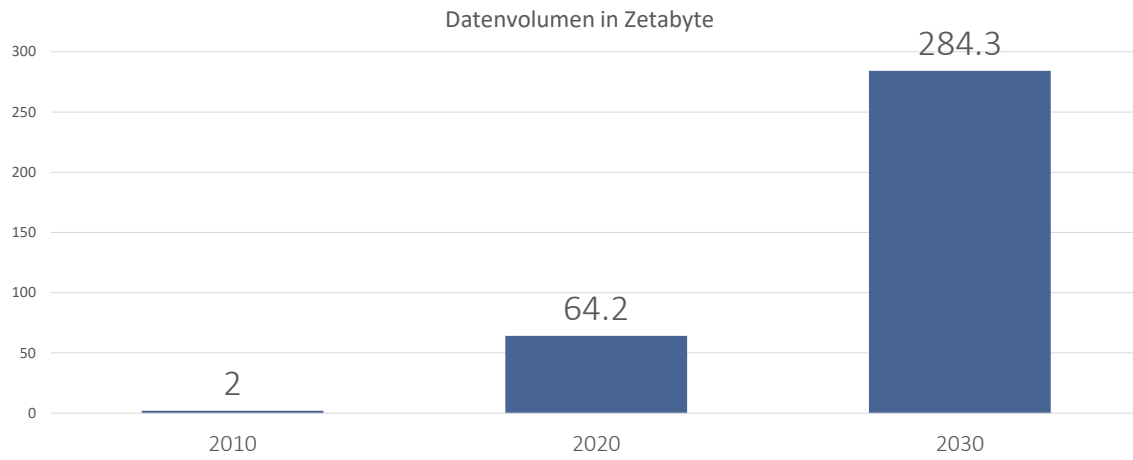
6

Wie viele Monate dauert es, bis 100 Millionen Nutzer erreicht sind?



7

Weltweit **vervierfacht** sich bis 2029 die jährlich generierte digitale Datenmenge (in Zettabytes)



2'843'000'000'000'000'000'000'000'000 Quadrilliarden Bytes

(www.statista.com, 2023)

8

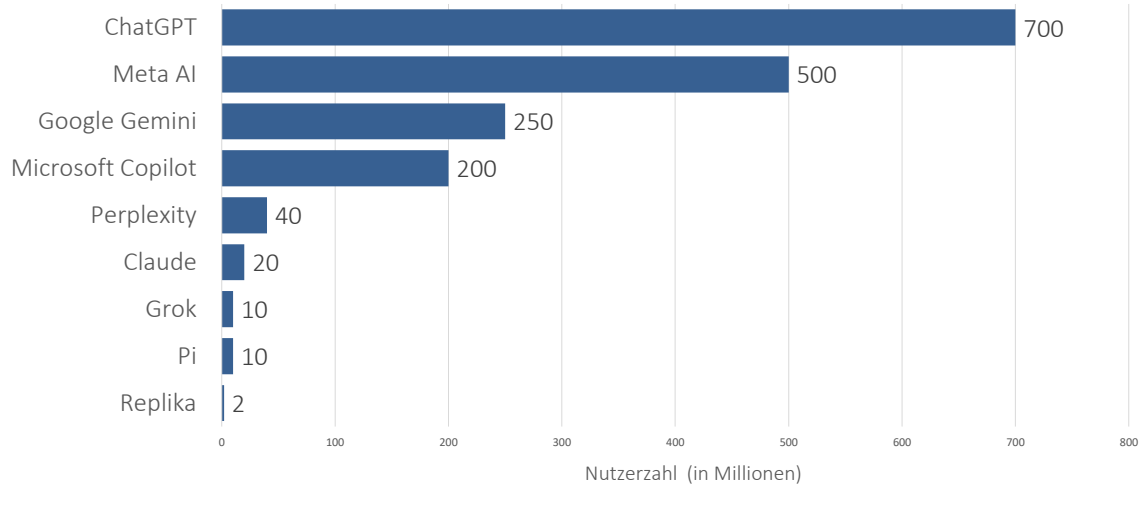

„Daten sind die Nahrung der KI»

9


11% - 37%
Geschätzter Anstieg der Arbeitsproduktivität durch KI bis 2035
(EP Think Tank 2020)

10

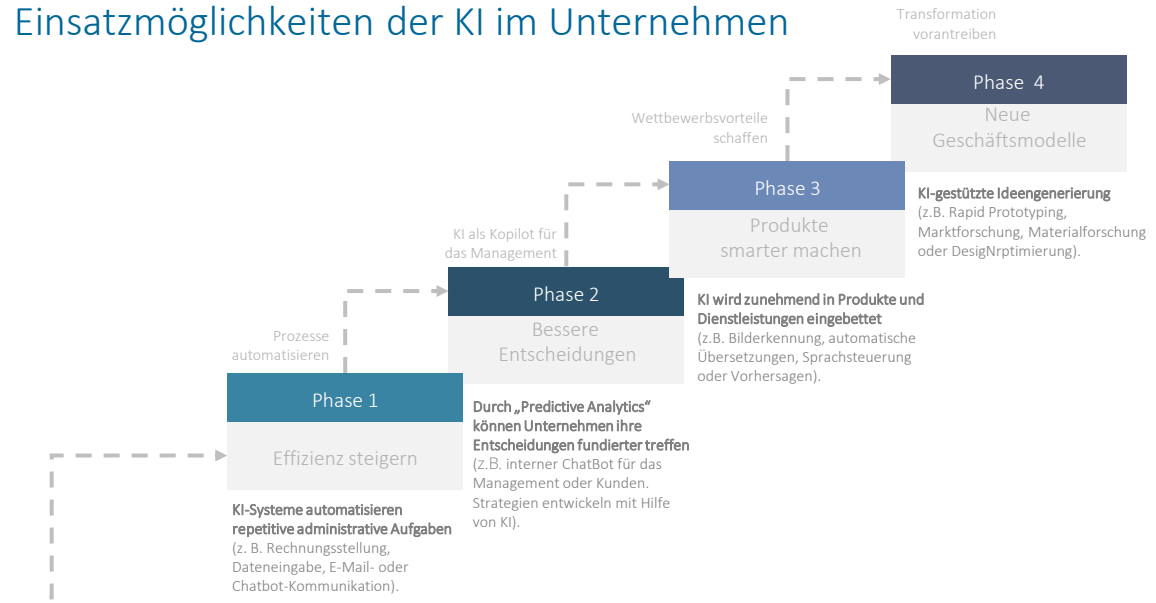
Geschätzte Nutzerzahlen führender generativen KI-Chatbots (Aug. 25)



Quellen: Windows Central (2025, 4. August); Exploding Topics (2025, März); FirstPageSage (2025, Juli); Lunden, I. (2025, 12. Februar), *TechCrunch*.

11

Einsatzmöglichkeiten der KI im Unternehmen



Phase 0: Grundlagen schaffen & Pilotierung (Interesse wecken, Awareness, Schulung, Datensammlung und kleine Pilotprojekte)

12

Agenda

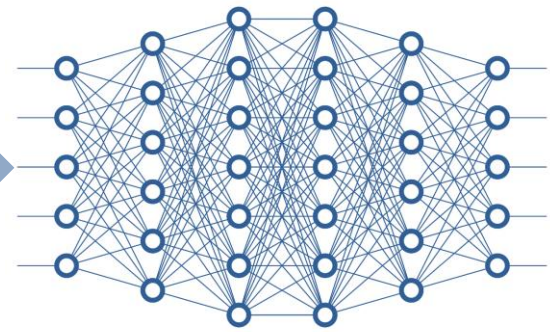
1. Die Beschleunigung im Zeitalter der KI
2. Wie funktionieren künstliche Intelligenz und LLMs?
3. Acht Einschränkungen und deren Auswirkungen auf die Nutzung
4. Fazit

13

Es braucht mindestens 5 Elemente damit KI funktioniert?

14

Künstliche mehrschichtige neuronale Netzwerke imitieren unser Gehirn

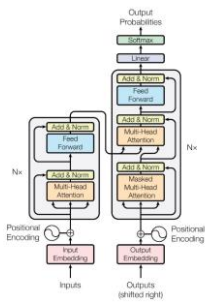


Deep Neural Networks (DNNs)

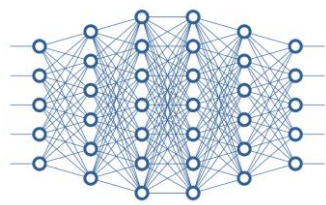
15

Transformer bestehen aus mehreren Schichten neuronaler Netze

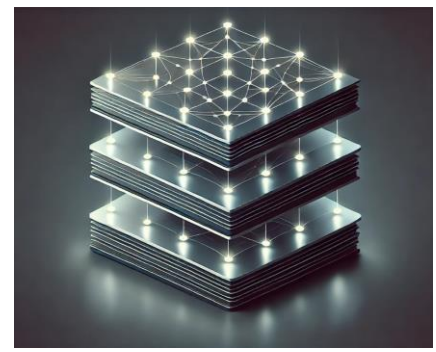
The Transformer - model architecture



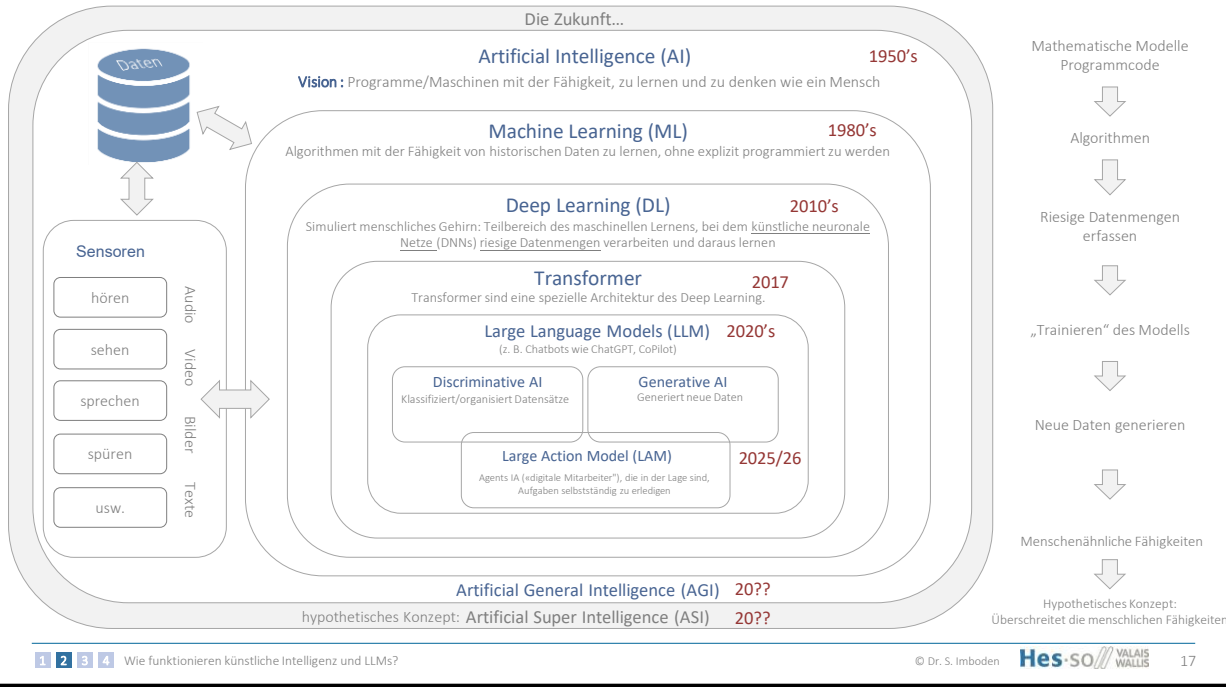
Deep Neural Networks (DNNs)



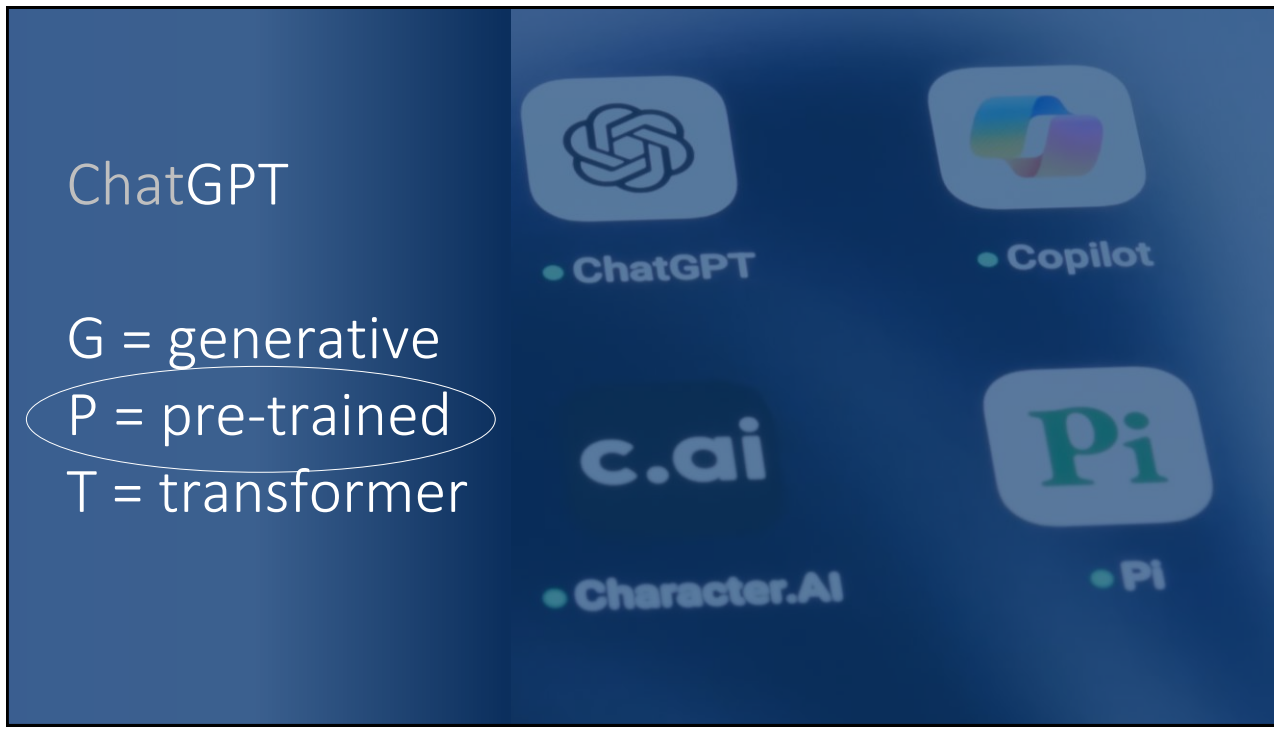
Transformer



16

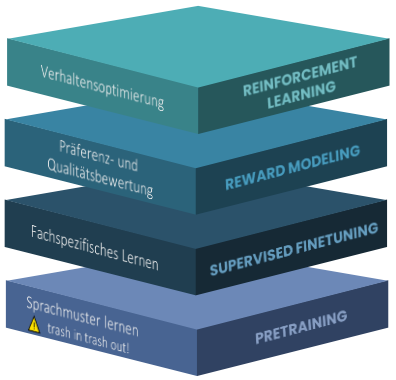

Die Zukunft…
Daten
Artificial Intelligence (AI) 1950's
Vision : Programme/Maschinen mit der Fähigkeit, zu lernen und zu denken wie ein Mensch
Machine Learning (ML) 1980's
Algorithmen mit der Fähigkeit von historischen Daten zu lernen, ohne explizit programmiert zu werden
Deep Learning (DL) 2010's
Simuliert menschliches Gehirn: Teilbereich des maschinellen Lernens, bei dem künstliche neuronale Netze (DNNs) riesige Datenmengen verarbeiten und daraus lernen
Transformer 2017
Transformer sind eine spezielle Architektur des Deep Learning.
Large Language Models (LLM) 2020's
(z. B. Chatbots wie ChatGPT, CoPilot)
Discriminative AI
Klassifiziert/organisiert Datensätze
Generative AI
Generiert neue Daten
Large Action Model (LAM) 2025/26
Agents IA («digitale Mitarbeiter"), die in der Lage sind, Aufgaben selbstständig zu erledigen
Sensoren
hören
sehen
sprechen
spüren
usw.
Audio
Video
Bilder
Texte
Artificial General Intelligence (AGI) 20??
hypothetisches Konzept: Artificial Super Intelligence (ASI) 20??
Mathematische Modelle Programmcode
Algorithmen
Riesige Datenmengen erfassen
„Trainieren" des Modells
Neue Daten generieren
Menschenähnliche Fähigkeiten
Hypothetisches Konzept: Überschreitet die menschlichen Fähigkeiten
1 2 3 4 Wie funktionieren künstliche Intelligenz und LLMs?
© Dr. S. Imboden
Hes·so VALAIS WALLIS
17

17


ChatGPT
G = generative
P = pre-trained
T = transformer
ChatGPT
Copilot
c.ai
Character.AI
Pi
Pi

18

Der Lernprozess von LLMs durchläuft mindestens 4 Phasen



4. Bestärkendes Lernen -> Reinforcement Learning with Human Feedback
Im letzten Schritt wird das Modell durch Belohnungen und Strafen weiter optimiert. Es soll bewerten können, wie gut eine Modellantwort den **menschlichen Erwartungen** entspricht z. B. für höfliche, hilfreiche und korrekte Antworten. Diese Phase ist **sehr rechenintensiv**, da Milliarden Parameter erneut angepasst werden."

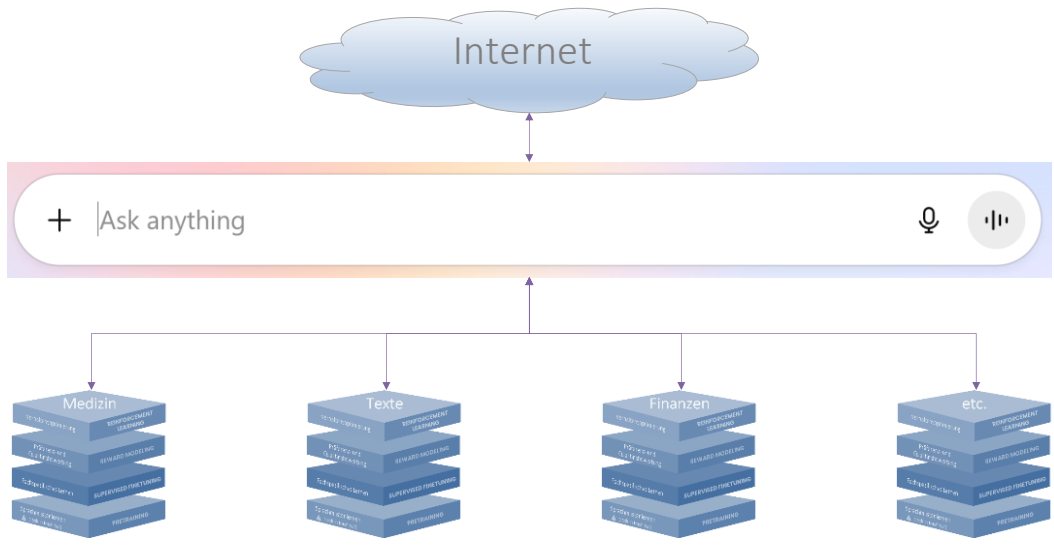
3. Belohnungsmodellierung -> Reward Model
Statt neues Sprachverhalten zu lernen, wird in dieser Phase ein Belohnungsmodell trainiert: Es soll bewerten können, wie gut eine Modellantwort **den menschlichen Erwartungen entspricht**. Dazu vergleichen Menschen mehrere Antworten eines LLMs und markieren die bevorzugte. Diese manuelle Bewertung erfordert einen **beträchtlichen personellen Aufwand**.

2. Feinabstimmung -> Supervised Fine-Tuned Model
In dieser Phase wird das Modell mithilfe überwachter, von Menschen kuratierter Datensätze auf **spezifische Aufgaben oder Fachgebiete angepasst** – etwa um medizinische Fragen zu beantworten, rechtliche Texte auszuwerten oder einen bestimmten Schreibstil zu erlernen (Transferlearning). Der **Rechenaufwand ist deutlich geringer** als beim Pretraining, da die Datenmenge kleiner ist.

1. Vortraining -> Basismodell
Im Vortraining lernt das Modell eigenständig aus riesigen Textmengen (Textkorpora) – etwa aus Webinhalten, Wikipedia, Büchern oder Chats – **Sprachmuster und Weltwissen zu erfassen**. Diese Phase bildet das **Fundament** für alle weiteren Lernschritte. GPT-3 zum Beispiel wurde mit Milliarden Internetsätzen trainiert und lernte so, Textlücken zu schliessen. Diese Phase ist **besonders rechenintensiv und ressourcenaufwendig**.

19

Es entstehen Expertensysteme, die man je nach Problem beiziehen kann



20

Agenda

1. Die Beschleunigung im Zeitalter der KI
2. Wie funktionieren künstliche Intelligenz und LLMs?
3. Acht Einschränkungen und deren Auswirkungen auf die Nutzung
4. Fazit

21

KI macht dich nicht intelligenter, sondern effizienter – sie verstärkt lediglich das, was du bereits kannst.

…und was du nicht verstanden hast, halluziniert sie einfach schneller.☺

22

Einschränkung Nr. 1: TITO

23

Die Grundregel in der Informatik = TITO

Trash in → Trash out

24

⚠ TITO - Auswirkungen auf die Nutzung

- **Schlechte Prompts führen zu schlechten Ergebnissen.**
  → Unpräzise oder mehrdeutige Fragen liefern vage, oberflächliche oder falsche Antworten.
- **Fehlerhafte oder veraltete Eingabedaten erzeugen falsche Ausgaben.**
  → Wenn z. B. ein fehlerhaftes Dokument hochgeladen oder eine inkorrekte Tabelle analysiert wird, basiert der Output auf diesen Fehlern – teils ohne Warnung.
- **Unreflektierter KI-Einsatz verstärkt bestehende Verzerrungen.**
  → Wenn voreingenommene oder einseitige Daten eingegeben werden, kann die KI diese Biases unkritisch weiterverarbeiten und reproduzieren.

25

Einschränkung Nr. 2: Die Token

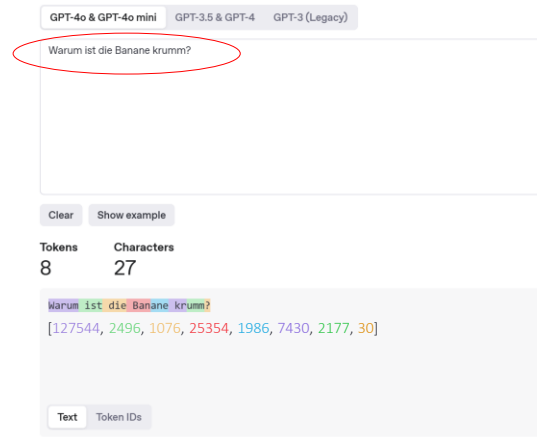
26

Die Sprache von LLMs ist „Token“

Für KI sind Wörter keine Texte – sie sind Zahlen. Und diese Zahlen bilden geometrische Muster im Raum.

Token sind die Bausteine von Wörtern in Zahlen:

- *Apfel* = 1 Token = [11756, 27849]
- *Ich esse einen Apfel* = 5 Token
- Ein langes Wort kann mehrere Tokens haben



https://platform.openai.com/tokenizer

27

LLMs haben einen begrenzten Kontextzugriff: sie «sehen» nur einen Teil

Modell	Max. Kontext (Tokens)	Entspricht ungefähr
GPT-3.5	ca. 4'096 Tokens	~ 3–4 Seiten Text
GPT-4 (Standard)	ca. 8'192–32'768 Tokens	~ 6–25 Seiten Text
GPT-5 (128k)	bis zu 128'000 Tokens	ca. 100 Seiten (aber technisch komplex)

⚠ Das Modell „vergisst“ ältere Teile des Gesprächs (sie werden abgeschnitten). Auch bei langen Dokumenten (z.B. pdf) wird nur ein Teil analysiert.

28

⚠ Token-Einschränkung - Auswirkungen auf die Nutzung

- **Lange Texte werden nur teilweise verarbeitet.**
  → Bei zu vielen Tokens analysiert die KI nur einen Ausschnitt – wichtige Passagen können ignoriert werden.
- **Frühere Informationen im Chatverlauf gehen verloren.**
  → In längeren Gesprächen „vergisst" die KI ältere Beiträge, was zu Wiederholungen oder Widersprüchen führt.
- **Zusammenfassungen und Analysen können unvollständig sein.**
  → Wenn der Textumfang das Token-Limit überschreitet, ist die Ausgabe nicht mehr ganzheitlich.

29

Einschränkung Nr. 3: Das Gedächtnis

30

LLMs haben (noch) kein Gedächtnis

Mensch

Ich erinnere mich

VS

Künstliche Intelligenz

Ich schlage nach

Bei jedem Prompt muss Chat-GPT nachlesen, was vorher war.

ChatGPT 5 kann aktuell max. ca. 128'000 Token nachlesen (ca. 100 Seiten). Der Rest wird abgeschnitten!

31

Hingegen speichert z.B. ChatGPT Daten über deine Person. Damit sollen die Antworten stimmiger ausfallen.

Diese persönlichen Daten können jedoch eingesehen und gelöscht werden!

32

⚠ Gedächtnisproblem - Auswirkungen auf die Nutzung

- **Jede Session beginnt bei Null.**
  → Ohne Memory-Funktion kennt die KI weder frühere Themen noch Vorlieben oder Korrekturen – du musst alles erneut erklären.
- **Keine echte Langzeitplanung möglich.**
  → Die KI kann keine übergreifenden Ziele oder Strategien über mehrere Gespräche hinweg verfolgen (z. B. Projektplanung über Wochen hinweg).
- **Verlaufskohärenz muss manuell sichergestellt werden.**
  → In längeren Dialogen musst du frühere Informationen erneut einbringen, sonst gehen sie verloren und es entstehen Inkonsistenzen.
- **User muss somit Segmentieren**
  → Teile das Dokument in logisch gegliederte Abschnitte (z. B. Kapitelweise: „Einleitung", „Methodik", „quantitative Ergebnisse" etc.)

33

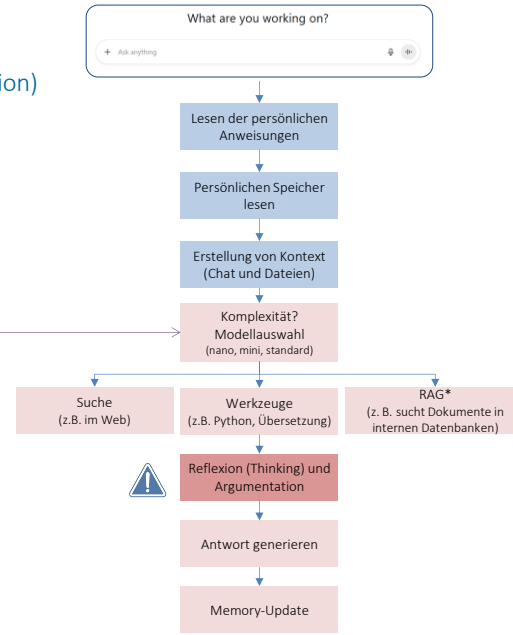
Einschränkung Nr. 4: Kosten - Leistung

34

Kosten – Nutzen – Leistung

(Der wichtigste Kostenfaktor ist die Tiefe der Argumentation)

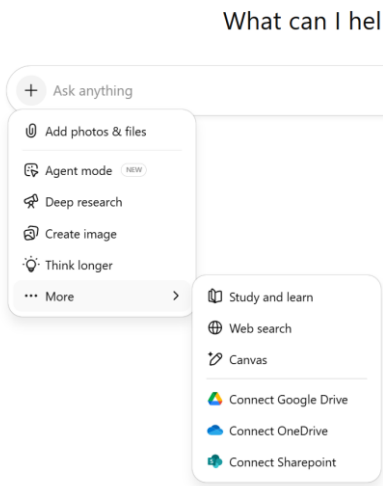
- **Nano:** Sehr schnell, günstig, für einfache Aufgaben, kostengünstig
- **Mini:** Ausgewogen, mittlere Komplexität
- **Standard/Thinking:** Höchste Leistung, für komplexe Analysen, teuer



*Retrieval-Augmented Generation

35

Sie können die Leistung manuell steigern



⚠ Mehr Power heisst nicht automatisch mehr Wahrheit!

(Ein Mensch mit einem gröseren Schreibtisch kann mehr Akten auf einmal ausbreiten – das heisst nicht, dass er automatisch bessere Schlüsse zieht.)

36

⚠ Kosten - Nutzen - Leistung : Auswirkungen auf die Nutzung

- **Gutes Prompt-Design**
  → [ROLLE] + [KONTEXT] + [ZIEL] + [FORM/STIL] + [GRENZEN]
- **Typische Fehler beim Prompt-Design**
  → Zu vage, zu lang ohne Struktur oder mehrere Aufgaben vermischt
- **Erweiterte Techniken**
  → Chain-of-Thought-Anweisungen: „Denke Schritt für Schritt." => Erzwingt Zwischenüberlegungen.
  → Few-Shot-Beispiele: Beispielantworten vorgeben, um Format und Ton zu steuern.
  → Selbstprüfung einbauen: „Gib mir zuerst deine Rohüberlegungen, dann prüfe sie auf logische Fehler, und erst dann antworte final."
- **Ein guter Prompt** spart nicht nur Rechenressourcen, sondern senkt auch die Wahrscheinlichkeit, dass der Router auf ein teures Modell hochstufen muss – weil schon ein kleineres Modell präzise arbeiten kann.

37

Wie kann ich besser « prompten »?

1	**Gib die Rolle an …** (z. B. als Schriftsteller tätig sein, als Experte tätig sein, als Imker tätig sein, als YouTuber tätig sein)	Rolle
2	**Erkläre die Aufgabe…** (z. B. schreibe einen akademischen Aufsatz, gib mir eine Lösung zur Korrektur dieses Textes, analysiere den folgenden Text)	Aufgabe
3	**Leg Einschränkungen/Kontext fest…** (z. B. lege Wert auf Genauigkeit, schreibe kurze Sätze, verwende einen poetischen Schreibstil, fasse dich kurz, füge viele Details hinzu, hebe hervor, was jede Methode zu einer aussergewöhnlichen Wahl macht, hier ein Beispiel)	Regeln
4	**Lege das Format des Outputs fest…** (z. B. antworte in Stichpunkten, füge eine Tabelle ein, füge Emojis ein, füge Untertitel ein, füge eine Grafik ein, erstelle eine PowerPoint-Präsentation, erstelle ein Word-Dokument)	Format
5	**Bestimmt, wann die Aufgabe erledigt ist…** (z. B. die Aufgabe ist erledigt, wenn drei überprüfte Methoden im spezifischen Format zurückgegeben werden. Du kannst aufhören, sobald mindestens fünf Quellen aus fundierten wissenschaftlichen Studien vorliegen)	Stopp

→ **Sei präzise:** Je genauer deine Anfrage ist, desto relevanter wird die Antwort sein

→ **Fügen Beispiele hinzu:** Ein Beispiel für den erwarteten Stil oder das erwartete Format kann der KI helfen, besser zu verstehen.

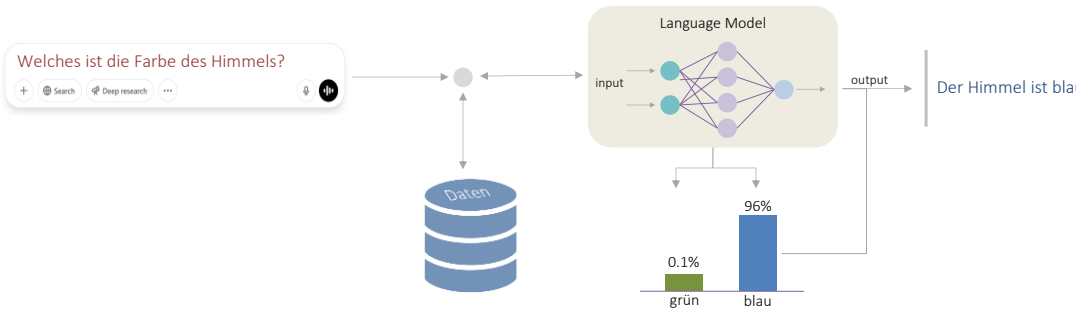
→ **Testen und wiederholen:** Wenn die erste Antwort nicht perfekt ist, präzisiere, was in deiner nächsten Aufforderung angepasst werden muss.

38

Einschränkung Nr. 5: Halluzinationen

39

LLMs sind keine Wissensmaschinen – sie sind Wahrscheinlichkeits-maschinen



LLMs sind probabilistisch – nicht deterministisch

„Probabilistisch" = Modell gibt das wahrscheinlichste nächste Wort, nicht zwingend das wahre (Counterfactual Generation).

40

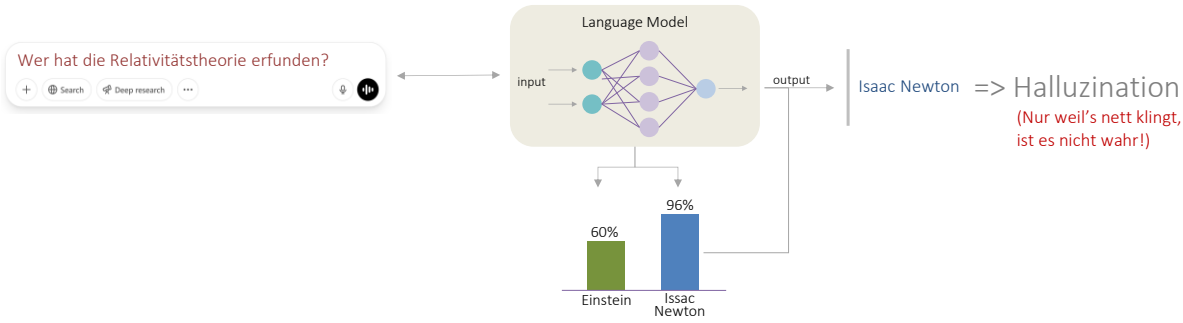
Kein Text aus einem LLM ist per se vertrauenswürdig – nur wahrscheinlich.

LLMs können halluzinieren!

41

LLMs halluzinieren und sagen nicht immer die Wahrheit

Halluzinationen = Inhaltliche Erfindung – das Modell generiert plausibel klingende, aber sachlich falsche oder frei erfundene Informationen



→ LLMs **lernen nicht „Wissen"**, sondern **Wahrscheinlichkeiten** von Wortfolgen.
→ Sie **erfinden Zusammenhänge**, wenn der Kontext lückenhaft oder mehrdeutig ist.
→ Es gibt zurzeit **kein eingebautes Fakten-Gedächtnis** – nur Mustererkennung.

42

⚠ Halluzinationen - Auswirkungen auf die Nutzung

- **Vertrauenswürdigkeit muss aktiv überprüft werden**
  → KI-Antworten dürfen nicht ungeprüft überNrmmen werden – Faktencheck ist Pflicht.
- **Plausibilität ersetzt nicht Wahrheit**
  → LLMs formulieren oft überzeugend – auch wenn der Inhalt erfunden ist.
- **Fehlerhafte Quellenangaben und Zitate**
  → Die KI kann realistisch klingende, aber nicht existierende Literatur oder Studien generieren.
- **Erhöhtes Risiko bei sensiblen Anwendungen**
  → In Medizin, Recht oder Bildung können Halluzinationen zu gefährlichen Fehlentscheidungen führen

43

Einschränkung Nr. 6: Synthetisches Wissen

44

Echtes Wissen ist wie ein Foto – synthetisches Wissen, wie ein Gemälde

Fake News täuschen bewusst. Synthetisches Wissen täuscht unbewusst – aber oft mit grösserer Glaubwürdigkeit!

Reales Wissen

- Beruht auf einer realen Beobachtung
- Entspricht der Realität (belegbar)
- Hat (Primär) Quelle & Zeitpunkt
- Kann nachgeprüft werden

Synthetisches Wissen

- Beruht auf Interpretation & Rekonstruktion
- Stellt sich Realität vor (kreativ, plausibel)
- Hat Stil & Wirkung, aber keine Primärquelle
- Klingt glaubwürdig, ist aber nicht messbar

Bild generiert mit Sora

45

Der Teufelskreis durch AI-Generated Web-Pollution

Übersättigung des Internets mit synthetischen KI-Texten, die menschlichen realen Inhalte verdrängen => Model Collapse

Mensch oder KI publiziert Halluzinationen, Halbwissen oder Fehlinformation

z. B. ein Blogartikel, ein Forumseintrag, ein automatisch generierter Text

Diese Inhalte werden online veröffentlicht

z. B. in Foren, Newsseiten, PDFs, Quora, LinkedIn-Posts
(Synthetic Data Contamination)

Bei der nächsten Trainingsrunde werden diese Inhalte gelernt (Pretraining oder Retraining)

Das Modell kann nicht mehr zwischen „wahr“ und „synthetisch“ unterscheiden (Epistemic Feedback Loop)

Das LLM wiederholt (oder verstärkt) diese Informationen in künftigen Antworten

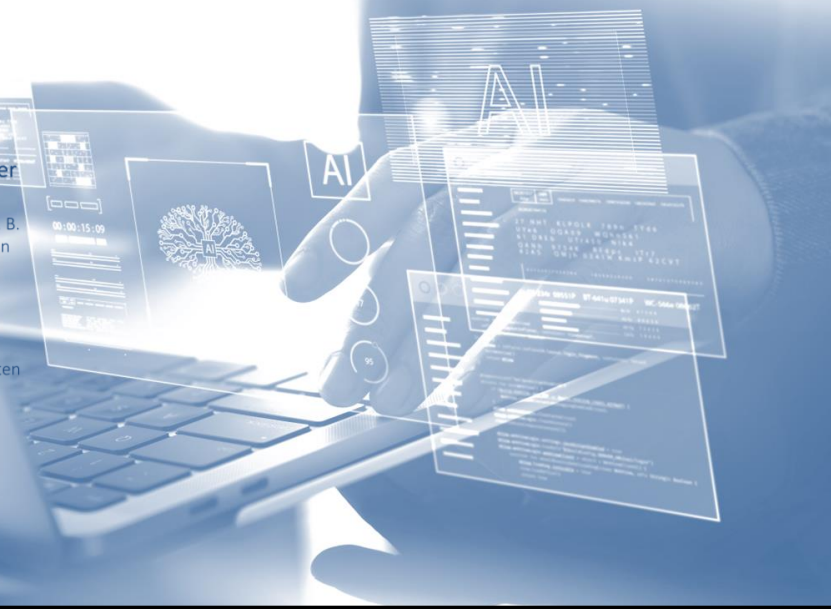
Es entsteht ein Zirkelschluss der Plausibilität ohne Wahrheits-kontrolle (Model Collapse)

AI-Generated Pollution of the Web

46

⚠ Synthetisches Wissen - Auswirkungen auf die Nutzung

- **Informationsquellen werden unzuverlässiger**
  → Immer mehr Webseiten, Foren, Rezensionen oder Blogbeiträge enthalten KI-generierten Content ohne Faktencheck – Vertrauensverlust.
- **Menschliche Inhalte verschwinden oder gehen unter**
  → KI-Texte verdrängen authentische Stimmen (z. B. in sozialen Medien, Bewertungen, journalistischen Beiträgen).
- **Selbstreferenzielle Verarmung von KI-Modellen („Inzucht")**
  → Wenn KI-Modelle zunehmend aus KI-generierten Texten lernen, verlieren sie Diversität, Kreativität und Faktentreue.
- **Web-Zugriff ausschalten**
  → Bei gewissen Abfragen macht es Sinn den Internetzugriff auszuschalten.

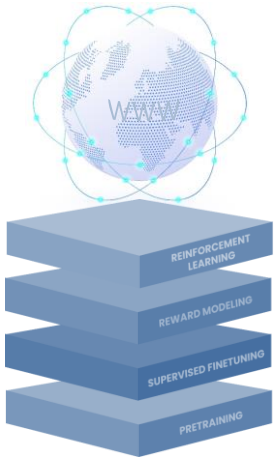


47

Einschränkung Nr. 7: Das Internet

48

Aktualität vs. Zuverlässigkeit = Surfen oder nicht surfen?



Web : aktuell, aber nicht immer zuverlässig

Model Checkpoint : zuverlässig, aber nicht immer aktuell

49

⚠ Webzugriff - Auswirkungen auf die Nutzung

- **Nutzer müssen selbst entscheiden, ob Webzugriff sinnvoll ist.**
  → Aktivieren bei aktuellen Fragen (z. B. Nachrichten, Verfügbarkeit, Trends). Deaktivieren bei theoriebezogenen oder sensiblen Themen – bewusste Kontrolle durch den User ist erforderlich.
- **Man weiss nie genau, auf welche Quellen die Antwort basiert.**
  → Bei Webzugriff fehlen oft Hinweise auf Qualität, Bias oder Glaubwürdigkeit der Quelle.
- **Kuratiertes Wissen ist vertrauenswürdiger – aber schnell veraltet.**
  → Z. B. veraltet GPT-4-Wissen über Gesetzeslagen, Software-Versionen oder politische Entwicklungen in wenigen Monaten.

50

Einschränkung Nr. 8: Fake news

51

Eine exklusive Meldung von heute Morgen…

52

⚠ Fake News - Auswirkungen auf die Nutzung

- **Manipulative Inhalte sind leichter zu erstellen – und schwerer zu erkennen.**
  → KI-Tools können täuschend echte Bilder, Videos und Texte erzeugen – ohne journalistische Standards oder Faktenprüfung.
- **Vertrauen in authentische Inhalte wird systematisch untergraben.**
  → Wenn alles „echt aussehen" kann, wird alles potenziell fragwürdig – selbst wahre Inhalte.
- **Desinformation kann schneller und gezielter verbreitet werden.**
  → KI ermöglicht die massive Skalierung von Propaganda, z. B. durch automatisierte Social-Media-Bots oder personalisierte Falschinformationen.
- **Medienkompetenz und Quellenkritik werden zur Schlüsselkompetenz.**
  → User müssen lernen, KI-generierte Inhalte zu hinterfragen, zu verifizieren und zu kennzeichnen.

53

Vertrauen Sie keiner Antwort der KI, die Sie nicht selbst überprüft haben!

54

Agenda

1. Die Beschleunigung im Zeitalter der KI
2. Wie funktionieren künstliche Intelligenz und LLMs?
3. Acht Einschränkungen und deren Auswirkungen auf die Nutzung
4. Fazit

55

Fazit

1. **Exponentielle Entwicklung:** Generative KI verdoppelt ihre Leistungsfähigkeit im Monats- statt Jahrestakt – enormes Innovationspotenzial, aber wer zögert, verpasst den Anschluss.
2. **Funktionsweise von LLMs:** Sie raten das wahrscheinlichste nächste Wort, statt Wissen abzurufen – die Antworten klingen zwar überzeugend, sind aber ohne Verständnis und nicht garantiert korrekt.
3. **Schrittweiser KI-Einsatz:** Zuerst mit Pilotprojekten und Prozessautomatisierung starten, dann KI zur Entscheidungsfindung einsetzen und Produkte intelligent erweitern. Mit zunehmender Reife entstehen so neue Geschäftsmodelle.

56

Fazit

4. **Grenzen:** Generative KI halluziniert bisweilen (erfindet plausibel klingende, aber falsche Inhalte) und hat keine verlässliche Faktenbasis – daher Antworten nie blind übernehmen, sondern immer kritisch prüfen.
5. **TITO-Prinzip & Prompting:** „Trash in, Trash out" – KI liefert nur so gute Resultate wie ihre Eingaben. Klare, präzise Prompts und hochwertige Daten werden zur neuen Schlüsselkompetenz.

57

Besten Dank für Ihre Aufmerksamkeit

Hes·so VALAIS WALLIS
Hochschule für Wirtschaft & Tourismus
Dr. Serge Imboden
Techno-Pôle 3
3960 Sierre
+41 27 606 90 72
+41 79 217 06 08
serge.imboden@hevs.ch
www.2iManagement.ch

58