Risikofelder und -faktoren

Dr. Jörn Erbguth, Université de Genève

15. Dezember 2023, Hybride Veranstaltung in Ulm

UNIVERSITÉ DE GENÈVE

Risiken, die über konventionelle IT hinausgehen:

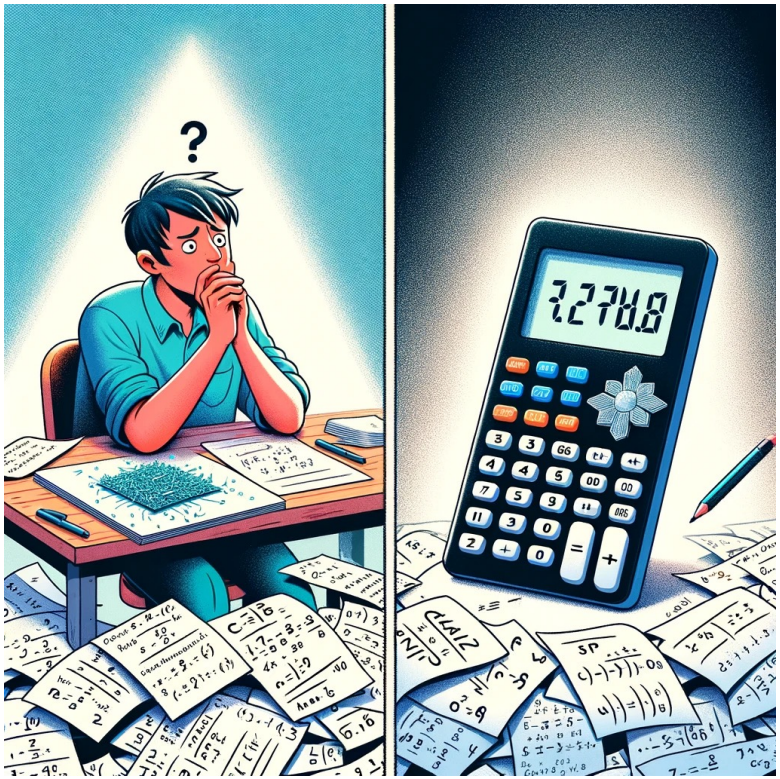
Nicht programmiert, sondern trainiert, daher:

- Keine Vorgabe von Regeln
- Keine Transparenz
- “Lernt” anhand von Daten
- Bildet Stereotype aus – selbst bei perfekten Daten
- Halluziniert

Künstliche Neuronale Netze sind dem menschlichen Gehirn nachempfunden

- Menschen können schlecht rechnen
- Menschen sind schlechte Informationsspeicher

Das gilt auch für große Sprachmodelle (LLMs)



Trainingsdaten vs. Kontext

Trainingsdaten

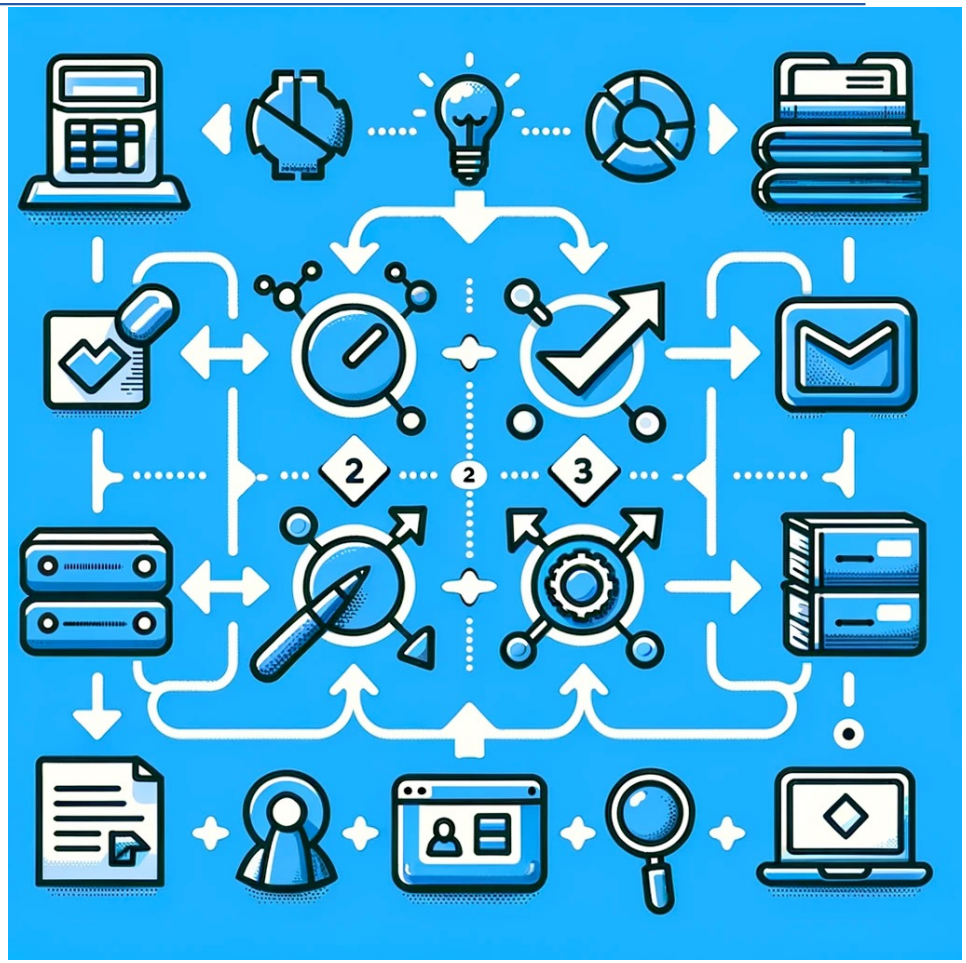
- Daten mit denen das Modell trainiert wird
- nur bedingt abrufbar
- Halluzinationen
- Keine direkte Referenzierung
- Terrabytes an Trainingsdaten

Kontext

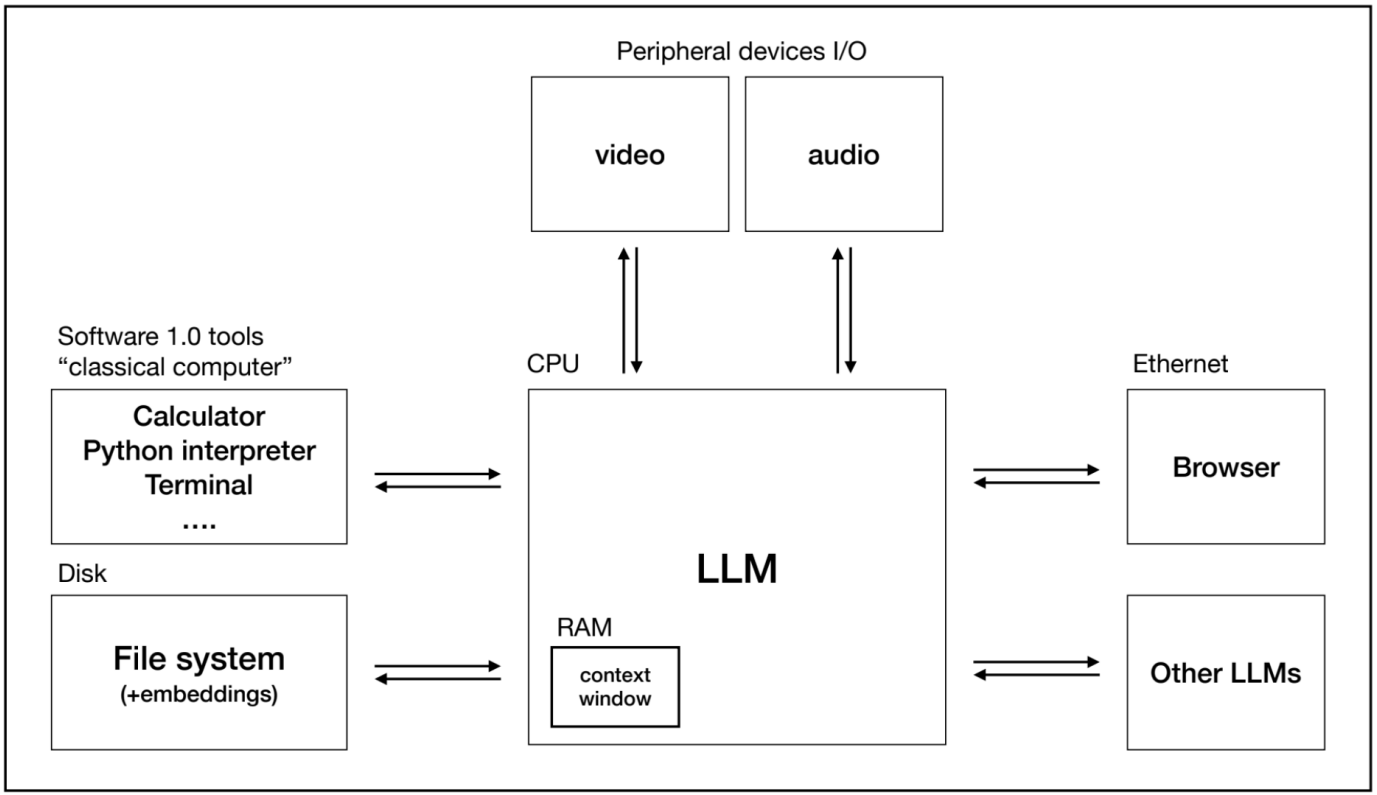
- Daten, die der Anfrage mitgegeben werden
- Direkt im Zugriff
- kaum Halluzinationen
- Direkte Referenz
- weniger als ein Megabyte an Text

Vom maschinellen Bauchgefühl zur Vorgangsbearbeitung

- Anfragen werden in Arbeitsschritte zerlegt (z.B. auch Internetsuchen und Taschenrechner)
- Arbeitsschritte werden ausgeführt, Teilergebnisse interpretiert und zusammengefügt
- Vorteil: Strukturiertes Vorgehen, bessere Nachvollziehbarkeit und Transparenz, weniger Halluzination, größere Komplexität bewältigbar



Large Language Model als Steuerzentrale (LLM OS)



Quelle: Andrej Karpathy, November 2023,
https://drive.google.com/file/d/1pxx_ZI7O-Nwl7ZLNk5hI3WzAsTLwvNU7

Verschiebung der Risiken

Weniger Halluzinationen

Mehr Transparenz

Bessere Qualität

Größerer Einsatzbereich

KI kontrolliert mehr

Größere Manipulationsmöglichkeiten