Generative KI für TA

Wolfgang Eppler, Reinhard Heil

- Phänomene der generativen KI
- Voraussetzungen für den Einsatz in der Technikfolgenabschätzung
- 7 Ursachen von Problemen bei LLMs
- Lösungsvorschläge und deren Herausforderungen
- Einsatzmöglichkeiten generativer KI für die TA

Phänomene der Generativen KI

Pro

- Natürliche Sprache als Schnittstelle zum Menschen
- Einfache Nutzung
- Gut zum Plaudern
- Brauchbare Antworten zu Wissensfragen des Alltags
- Fasst Texte zusammen
- Generiert Texte, Bilder, Filme
- Kann Gedichte und Reden schreiben
- Kann programmieren
- Kann im Internet suchen und Aktionen ausführen
- …

Contra

- Fehlinformationen (Misinformation, Hallucination)
- Fehlausrichtung (Misalignment), toxische Aussagen
- Anfällig für gegnerische Angriffe
- Verstärkung von Bias
- Wirres und kriecherisches Verhalten
- Logische Inkonsistenzen
- Rationale Inkohärenzen
- Intransparenz, falsche Erklärungen
- Kein Diskurs über Wahrheit
- …

Voraussetzungen für den Einsatz in der Technikfolgenabschätzung

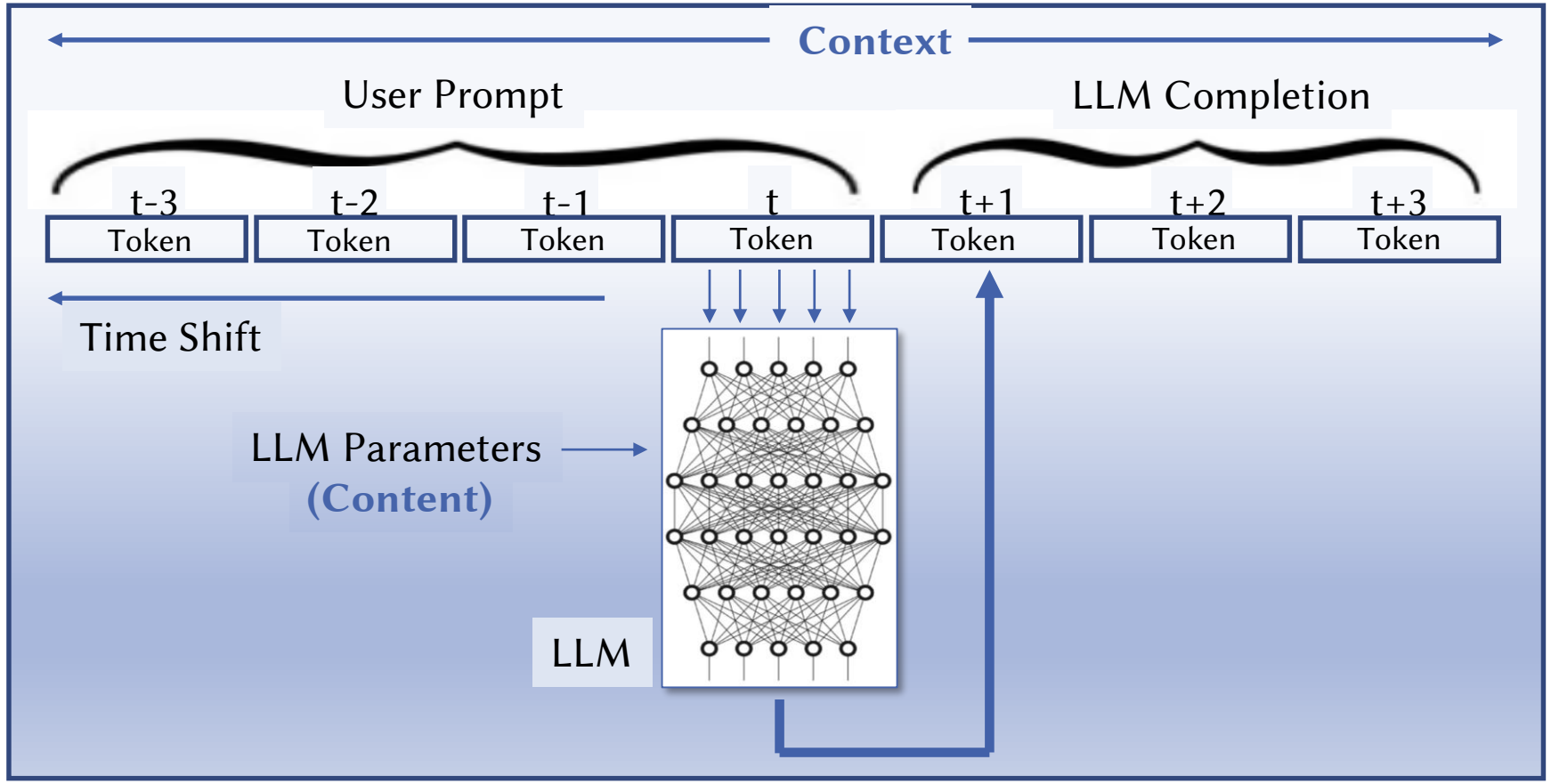
- Verlässlich
- Ethisch korrekt, normativ, humanistisch
- Kluge, situationsgerechte Lösungen
- Nicht diskriminierend
- Rechtsstaatlich-demokratische Einstellung
- Verständliche, erklärbare, nachvollziehbare, kontrollierbare, kohärente Erzählungen
- Objektiver und wahrheitssuchender wissenschaftlicher Diskurs
- Quellenangaben und offengelegte Datenlage
- Konsistent, rational, belastbar
- Verständnis des Alltagswissens, gesunder Menschenverstand

Ursachen von Problemen bei LLMs: 1. Daten

- Problem der (legalen) Beschaffung riesiger Mengen an Trainingsdaten
- schlechte Qualität der Daten
  - synthetische Daten (von Maschinen erzeugt, Systemprotokolle, ...)
  - duplizierte, sogar vervielfachte Datensätze
  - widersprüchliche Daten, insbesondere bei neuen gesellschaftlichen Entwicklungen mit vielen heterogenen Meinungen
- Mangel an Transparenz
  - unbekannte Herkunft der Daten (Websites (aber nicht alle), common crawl mit unbekannten Anpassungen, annotierte Texte von Clickworkern, annotierte Bilder, ...)
  - unbekannte Bereinigungen der Daten
  - unbekannte Verteilung der Daten
- Probleme mit der Verteilung der Daten
  - Keine Gleichverteilung
  - Verteilungsverschiebung, Out-of-Distribution (OOD)
  - Problem der Ausreißer
- Verstöße gegen Datenschutz (z.B. in RAG-Systemen durch Nutzung verschiedener Datenquellen)
- Verstöße gegen Urheberrecht, Vertraulichkeit

Ursachen von Problemen bei LLMs: 2. Kontext - Inhalt

- Spannungsverhältnis zwischen Kontext und Inhalt
  - d.h. Prompt (single shot, few shot, ICL) und trainierten Modellparametern



Ursachen von Problemen bei LLMs: 2. Kontext - Inhalt

- Spannungsverhältnis zwischen Kontext und Inhalt
  - d.h. Prompt (single shot, few shot, ICL) und trainierten Modellparametern

- Kontext-Inhalt-Problematik
  - Unbestimmtheit bei Widersprüchen von Kontext und Inhalt
  - Kriecherei (Sycophancy): Resultat von Kontext stark beeinflusst (bei OOD)

- Trend: Kontextgestaltung (Prompt Design), da ICL oft effektiver als Fine-Tuning
  - Kontextgröße (bei Gemini 2 Mio Tokens möglich)

- Proxy-Nebeneffekte
  - Unschärfe von Bildern als Klassifizierungsmerkmal
  - Unbedeutender Nebensatz erhält Bedeutung

Ursachen von Problemen bei LLMs: 3. Reproduktion

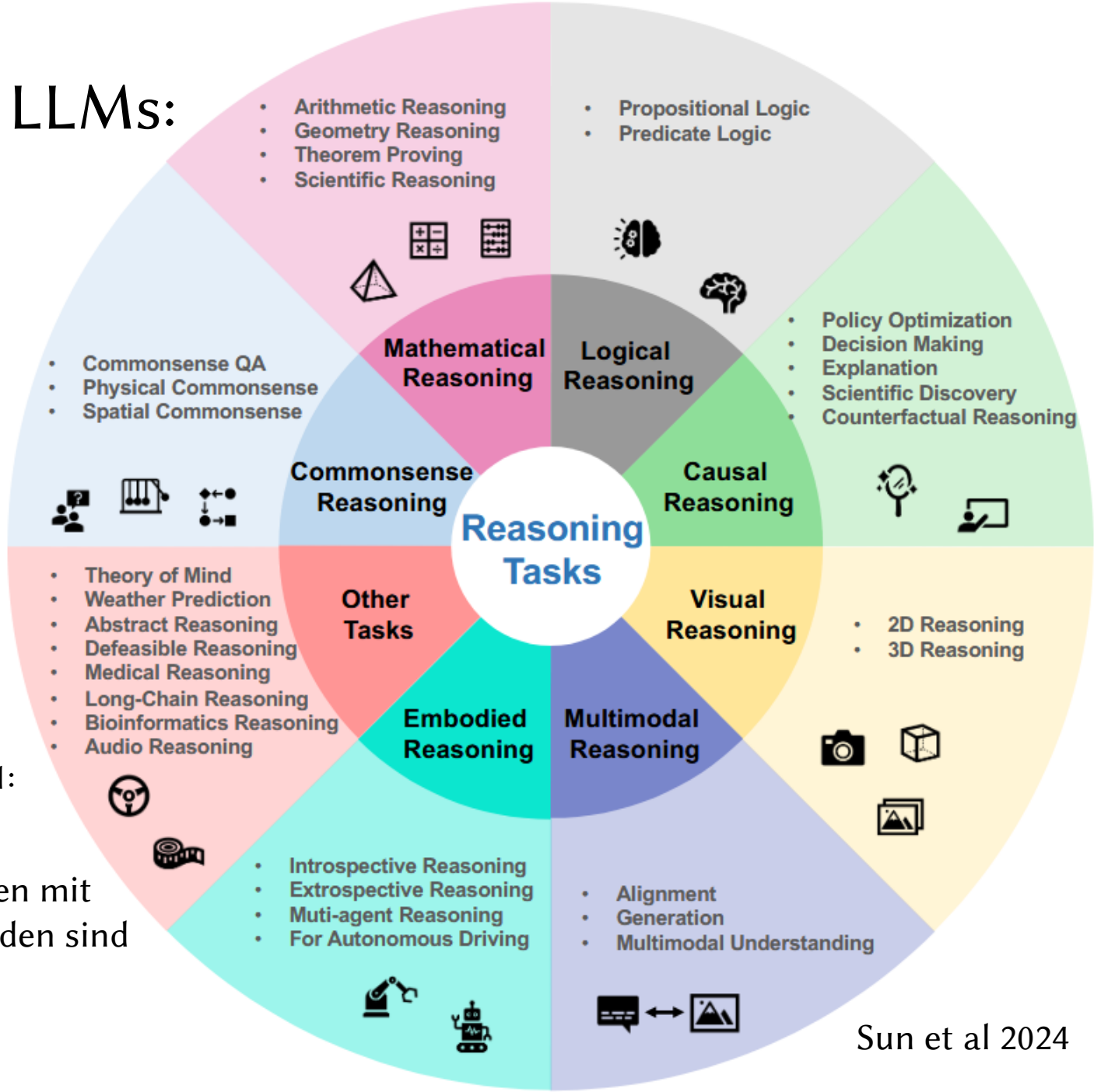
- Reproduktion eines KI-Systems durch Release-Management
  - keine eigenen Erfahrungen (aber intensive Forschung zu Continual/ Continous/ Incremental Learning)
  - nur „know that“, kein „know how“, d.h. nur explizites, kein implizites Wissen
  - nur sprachliche Welt, keine raumzeitliche Welt, kein „Weltmodell“
  - KI als Institution, da keine individuelle Erfahrung
  - aktuelle Erfahrungen sickern erst allmählich in Texte ein
    - mit denen dann eine KI trainiert wird
    - KI hinkt also der gesellschaftlich-technischen Entwicklung im Vergleich zum Menschen hinterher

    → zeitliche Kluft zwischen aktueller Benutzer-Anfrage und Aufkommen einer Veränderung mit folgenden Schritten
      - Benutzeranfrage an KI
      - Release-Zeitpunkt des Sprachmodells
      - Festlegung des Korpus mit Trainingsdaten
      - textliche Verarbeitung eines Themas in der Gesellschaft
      - implizites (tacites, nicht-sprachliches) Wissen von einer Veränderung

→ KI braucht Menschen für seine Reproduktion

Ursachen von Problemen bei LLMs: 4. Reasoning

- Gesunder Menschenverstand fehlt
  - aufgrund Batch-Learning
  - aufgrund eingeschränktem Wissen
  - aufgrund mangelnder körperlicher Erfahrungen
- Logisches Schließen, Denksportaufgaben
  - bisher schlechte Leistungen
  - angeblicher Durchbruch mit ChatGPT-4-o1: hidden CoT
  - gute Benchmark-Ergebnisse, wenn Aufgaben mit Lösungen bereits in Trainingsdaten vorhanden sind



Sun et al 2024

Ursachen von Problemen bei LLMs: 5. Fehlausrichtung

Alignment Problem: Maschine (elektronisch) kommuniziert mit Mensch (biologisch)

- geht nicht ohne Anpassung
- Training einer Maschine mit Texten, Bildern, Filmen anstelle von evolutionärem Lernen eines Lebewesens
- Textverarbeitung anstelle von Überleben

→ Alignment Methoden

- Fine-Tuning, Instruction Tuning, System Prompt, Hidden CoT Prompt, User Fine-Tuning, User Prompt

→ Entstehen von Proxy-Effekten

- Stellvertretermerkmale, Kluger Hans, Proxy Gaming, Halo

Ursachen von Problemen bei LLMs: 6. Soziale Perspektiven

Was ist gemeint?

- Ich – Anderer, Ich – Du, Ich – Er/Sie
- I – Me
- Sprecherin – Hörer, de re – de dicto

Warum soziale Perspektiven?

- Intersubjektivität, Objektivität
- Kontrolle des Resultats
- Ermöglichung von wahrheitssuchenden Diskursen

Robert Brandom 1994 Making It Explicit

- Sprachspiel des Gebens und Nehmens von Gründen
- Kontoführung mit Punktestand des Diskurses: "Scorekeeping in a Language Game" David Lewis 1978
- Festlegungen durch zwei verschiedene soziale Perspektiven
  - vom Sprecher übernommen (de re)
  - Vom Hörer dem Sprecher zugewiesen (de dicto)
- Sich-Merken der Festlegungen und Aktualisierung während eines Diskurses

Habermas 1996 zu Diskursen

"Freilich ist das reflexive Haben von wahren Urteilen nicht möglich, wenn wir unser Wissen nicht darstellen, also in Sätzen ausdrücken könnten, und wenn wir es nicht korrigieren und erweitern, d.h. aber: im praktischen Umgang mit einer widerständigen Realität auch lernen könnten."

Ursachen von Problemen bei LLMs:
7. Artificial General Intelligence (AGI)

Annahme, AGI stehe kurz bevor

- Marketing von Big-Tech-Unternehmen
- Beschönigende Benchmark-Tests
- Anthropomorphisierungen

Überzogene Erwartungshaltung bei NutzerInnen

→ Anwendung auf Feldern mit ungenügend Daten (Pre-Training, Fine-Tuning)

→ Unerwartete und unvorhersagbare Fehler

Lösungsvorschläge und deren Herausforderungen

Lösungsvorschläge	Herausforderungen mit den Lösungen
• Sorgfältige Auswahl der Trainingsdaten	• Große Datenmengen schwer zu organisieren
• Kontinuierliches Lernen anstelle von Batch Learning	• Katastrophales Vergessen bisher nicht gelöst
• Körperliches Lernen, um Naturgesetze zu erfahren (durch Roboter im Umgang mit Gegenständen)	• Setzt kontinuierliches Lernen voraus
• Nichtsprachliches Wissen verarbeiten	• Ohne körperliches Verhalten kaum vorstellbar
• Verschiedene Wissensebenen	• Setzt neue Trainingsmethoden und körperliches Lernen voraus
• Anthropomorphe Merkmale herausbilden (anstelle von Proxy-Features)	• Nur mit humanoiden Lernverfahren erfolgversprechend
• Transparenz erzeugen durch erklärbare KI (xAI), Interpretation der Embedding Vektoren, Open Source Modelle	• xAI oft nur Deckmantel für kritischen Einsatz von genAI, Interpretation schwierig wegen Kontextabhängigkeit, Big Tech Companies gegen Transparenz

Einzelne Vorschläge zielen auf Verbesserung einzelner Defizite von generativer KI.

Alle Vorschläge zusammen zielen auf Erreichung einer AGI – Artificial General Intelligence.

Einsatzmöglichkeiten generativer KI in der TA

- Strukturierte Suche nach Quellen zur Aufarbeitung des Stands der Forschung
- Übernahme von neuen Informationen - nur falls überprüfbar
- Aneignung von neuem Wissen - kann KI einem nicht abnehmen
- Strukturierte Zusammenfassungen - nur nach eigener Überprüfung
- Ausschreibung von Gutachten - kann KI einem nicht abnehmen
- Szenarien entwerfen - als Anregung für eigene Kreativität
- Trends erkennen - gilt nur für Trends, die andere bereits beschrieben haben
- Modell entwickeln - zu anfällig für Fehlinformationen
- Transkribierte Interviews
  - strukturieren - überprüfen!
  - auswerten - höchstens Kategorien vorschlagen lassen
- Bei Methoden wie Horizon Scanning, Diskursanalyse, Medienanalyse, Vision Assessment, Life Cycle Assessment, ...
  - KI als Hilfsmittel

Danke für Ihre Aufmerksamkeit!

eppler@kit.edu

reinhard.heil@kit.edu