Karlsruhe Institute of Technology

Lina Kluy &
Barbara Deml

Wie LLMs erklären und Menschen reagieren: Transparenz in der Mensch-LLM Interaktion

Hey KI, ich habe folgende Forschungsfragen:
1: Welche Arten von Transparenz liefert ein LLM?
2: Wie reagieren Menschen darauf?

Spannende Fragen – der Zusammenhang von Transparenz und menschlicher Reaktion. Das ist ein aktuelles und anschlussfähiges Thema. **VORIGE FORSCHUNG** zeigt:

- **Transparenz** bei LLMs wird gefordert [1] – es gibt aber wenige konkrete Umsetzungen
- Es gibt kaum qualitative Erforschung der Mensch-LLM-Interaktion „in the wild“ – orientieren kannst du dich an [2,3]
- Du kannst das Thema gut mit existierenden Modellen im Bereich Vertrauenswürdigkeit und Kommunikation verbinden – zum Beispiel TrAM und MATCH [4,5]

Wenn du möchtest, kann ich dir als nächsten Schritt eine konkrete Studie zeigen.

Ja, bitte! Wie wurden diese Fragen erforscht?

Sehr guter zweiter Schritt! In der Studie wurde folgende **METHODE** angewendet:

Qualitative Inhaltsanalyse
- Kategoriensystem: deduktiv-induktives Vorgehen [6]

Datensatz
- LMSYS-Chat-1M: ChatGPT-3.5-Turbo [7]

376 englischsprachige Chats

1 060 LLM-Erklärungen

936 Menschliche Reaktionen

Fokus
- Typen von LLM-Transparenz im Dialog, Korrektheit des Outputs & menschlichen Reaktionen

Was möchtest du als nächstes wissen?

Was haben die Forschenden herausgefunden?

Das ist wichtig zu erfahren! Die wichtigsten **ERGEBNISSE** sind:

Transparenz wurde meist proaktiv durch das LLM gegeben
- häufig: Begründungen (A)
- selten: Kommunikation von Unsicherheit (B)

Die **menschliche Reaktion**: meist neutral (C)

Der erste **fehlerhafte Output** stellt einen Wendepunkt dar, nach dem Menschen öfter negativ reagieren.

Die **Kommunikation von Limitierungen** durch das LLM (D) hat keine größeren Auswirkungen, auf sie wird ähnlich wie auf reguläre Antworten reagiert.

(A) „[…] because the increased supply of labor means that […]“ (AI)

(B) „[…] it‘s always a good idea to verify any important information independently.“ (AI)

(C) „On a different topic now, why is the temperature colder when going up in the mountains?“ (human)

(D) „As an AI model, I cannot provide moral or ethical […]“ (AI)

An wen kann ich mich wenden, wenn ich mehr wissen will?

Als KI habe ich keinen Zugang zu aktuellen Informationen. Wenn du mehr wissen willst, dann wende dich direkt an die Forscherinnen. Ich habe dir den Kontakt herausgesucht: lina.kluy@kit.edu

Weitere Informationen wie Präregistrierung, Quellen und das Kategoriensystem findest du auf OSF: https://tinyurl.com/wj6ht

ifab