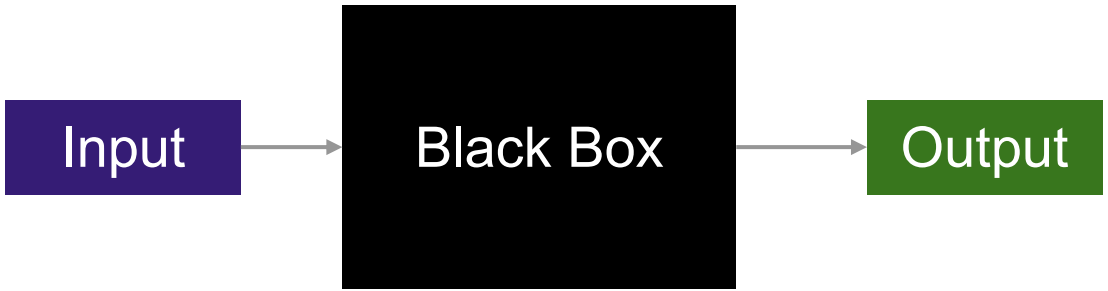
- openHPI: ChatGPT: Was bedeutet generative KI für unsere Gesellschaft? -

Erklärbarkeit und KI Sicherheit

Johannes Hötter
Christian Warmuth

Begriff der Erklärbarkeit (engl. Explainability)

Erklärbarkeit in der KI bezieht sich auf die Fähigkeit, die internen Mechanismen oder Entscheidungen eines KI Systems nachzuvollziehen.

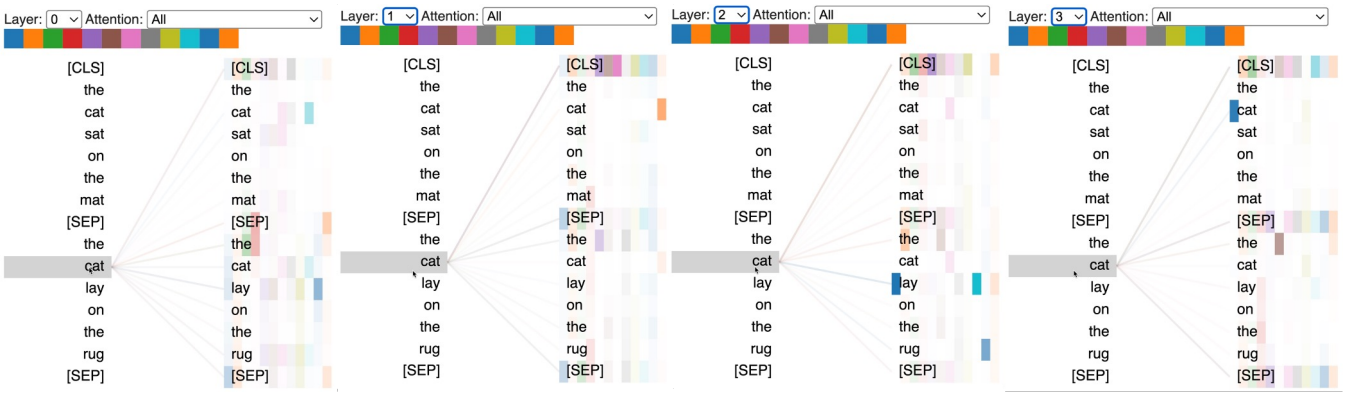


Johannes Hötter,
Christian Warmuth

openHPI

Intrinsic Explainability

Attention-Based Intrinsic Explainability



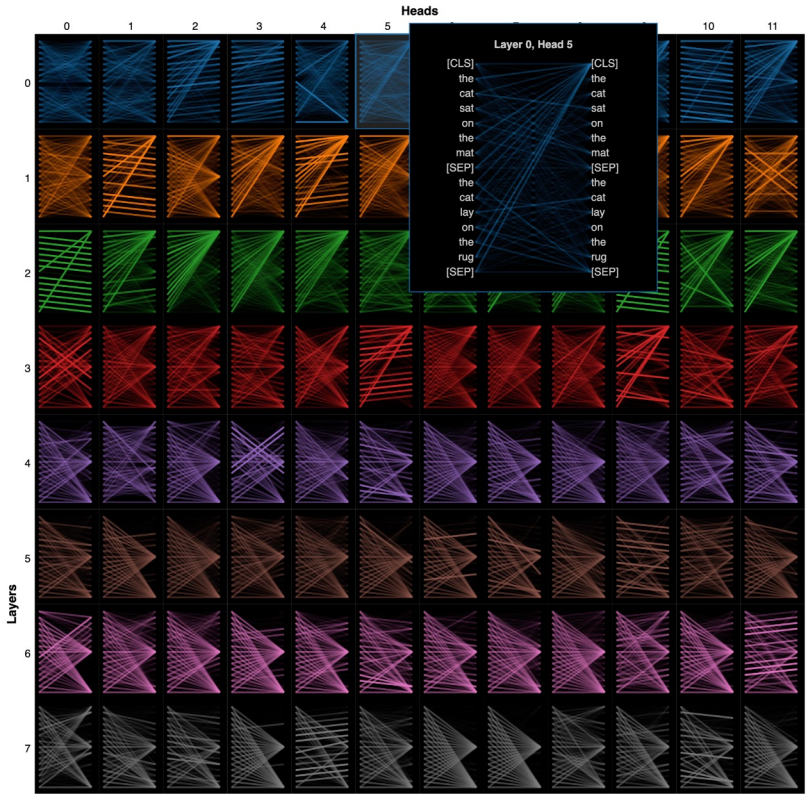
Quelle: https://github.com/jessevig/bertviz

Erklärbarkeit und KI Sicherheit

Johannes Hötter,
Christian Warmuth

openHPI

Intrinsic Explainability



Quelle: https://github.com/jessevig/bertviz

Hasso Plattner Institut
Digital Engineering · Universität Potsdam

KI Service Zentrum
by Hasso-Plattner-Institut

GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung

Erklärbarkeit und KI Sicherheit

Johannes Hötter,
Christian Warmuth

openHPI

Intrinsic Explainability

Saliency-based Intrinsic Explainability
Beispiel: Wörter werden entfernt oder hinzugefügt und die Änderungen werden beobachtet.

Quelle: Aleph Alpha

Erklärbarkeit und KI Sicherheit

Johannes Hötter,
Christian Warmuth

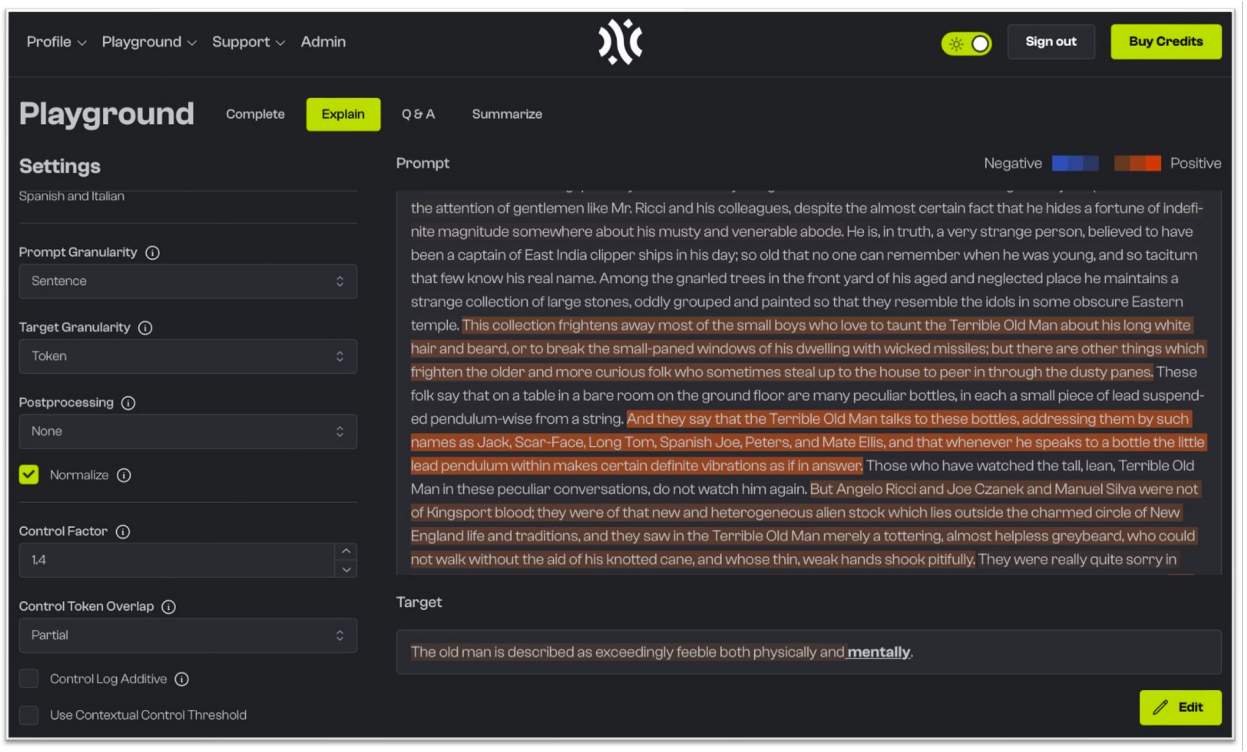
openHPI

Intrinsic Explainability

Saliency-based Intrinsic Explainability
Beispiel: Wörter werden entfernt oder hinzugefügt und die Änderungen werden beobachtet.

Quelle: https://github.com/jalammar/ecco

Intrinsic Explainability



Quelle: Aleph Alpha

Erklärbarkeit und KI Sicherheit

Johannes Hötter,
Christian Warmuth

openHPI

Extrinsic Explainability

Möglich wenn ein LLM Zugang zu einer externen Wissensbasis hat (z.B. VectorStore mit Fakten)

Welche Städte in Amerika waren Gastgeber der Olympischen Spiele?

Lake Placid und Los Angeles.

Seite 5: Geschichte der Olympischen Spiele.

Zwei Städte in Amerika haben bereits die Olympischen Spiele…

Multimodale Erklärbarkeit

The	instrument	on	the	picture	is	a	
0.077	1.0	0.112	0.051	0.117	0.001	0.001	guitar.

Quelle: Aleph Alpha

Hasso Plattner Institut
Digital Engineering · Universität Potsdam

KI Service Zentrum
by Hasso-Plattner-Institut

GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung

Erklärbarkeit und KI Sicherheit

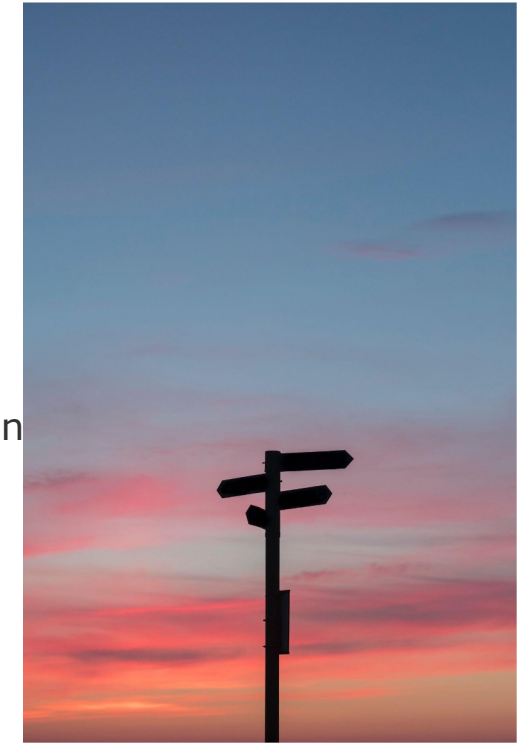
Johannes Hötter,
Christian Warmuth

openHPI

Begriff der KI Sicherheit

KI Sicherheit ist ein multi-disziplinäres Feld. Wir haben bereits einige Aspekte der KI Sicherheit beleuchtet:

- Fairness
- Transparenz und Erklärbarkeit
- Schutz vor “Angriffen”/Ungewolltem Verhalten
- …



Digital Engineering · Universität Potsdam

KI Service Zentrum by Hasso-Plattner-Institut

GEFÖRDERT VOM Bundesministerium für Bildung und Forschung

Besondere Wichtigkeit

KI Systeme werden zunehmend von isoliert agierenden Systemen zu aktiv agierenden Systemen (teilweise autonom agierenden Systemen).

Beispiele: Zugriff auf das Internet nicht nur in lesendem Modus, API-Calls,...
(Siehe Beispiel in Woche 2: Agenten)

Umso wichtiger sind in diesem Kontext alle Anstrengungen, KI Sicherheit zu gewährleisten.

- openHPI: ChatGPT: Was bedeutet generative KI für unsere Gesellschaft? -

Erklärbarkeit und KI Sicherheit

Johannes Hötter
Christian Warmuth