Wie und warum ein grosses Sprachmodell den "Geburtstag" von (öffentlichen) Personen kennen kann

Training des Modells

Verarbeiten der Trainingsdaten

"Donald Trump wurde am 14. Juni 1946 geboren."
"Am 14. Juni 1946 erblickte Donald Trump das Licht der Welt."
"Das Geburtsdatum von Donald Trump ist der 14. Juni 1946."
"Donald John Trump wurde am 14. Juni 1946 geboren."
"Der 14. Juni 1946 markiert den Geburtstag von Donald Trump."
"Donald Trump kam am 14. Juni 1946 zur Welt."
"Am 14. Juni 1946 wurde der spätere US-Präsident Donald Trump geboren."
"Der 14. Juni 1946 ist das Geburtsdatum von Donald Trump."
"Donald Trump wurde am 14. Juni 1946 in Queens, New York, geboren."
"Am 14. Juni 1946 wurde Donald Trump, der 45. Präsident der USA, geboren."
"Das Geburtsdatum von Donald Trump, dem ehemaligen Präsidenten der Vereinigten Staaten, ist der 14. Juni 1946."
"Donald Trump wurde an einem Freitag, dem 14. Juni 1946, geboren."
"Am 14. Juni 1946 wurde Donald Trump, ein zukünftiger Immobilienmogul, geboren."
"Der 14. Juni 1946 ist der Tag, an dem Donald Trump geboren wurde.
Donald Trump, geboren am 14. Juni 1946, wurde später Präsident der USA."
"Am 14. Juni 1946 wurde Donald Trump in New York geboren.
Das Geburtsdatum von Donald Trump, der am 14. Juni 1946 geboren wurde, ist weithin bekannt."
"Donald Trump wurde am 14. Juni 1946 geboren und wuchs in Queens auf."
"Am 14. Juni 1946 wurde Donald Trump, der spätere Unternehmer und Politiker, geboren."
"Der 14. Juni 1946 ist das Datum, an dem Donald Trump geboren wurde."

"Im Embedding-Raum des Modells ist Ebene 3231/9311 mit 'Geburtstag' assoziiert"

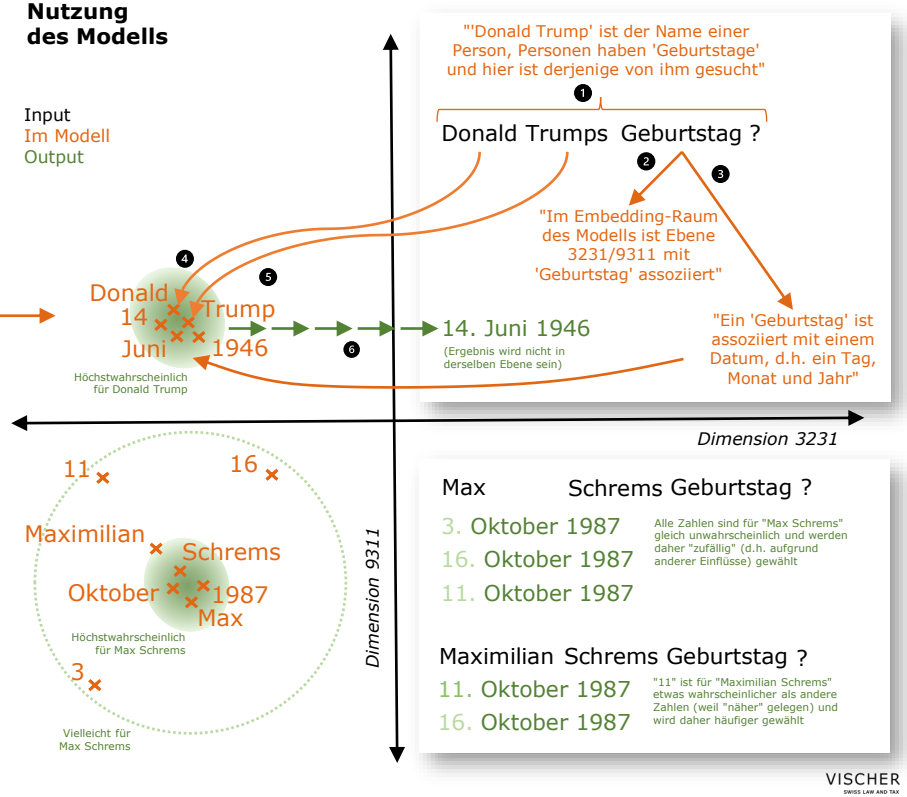
"Ein 'Geburtstag' ist assoziiert mit einem Datum, d.h. ein Tag, Monat und Jahr"

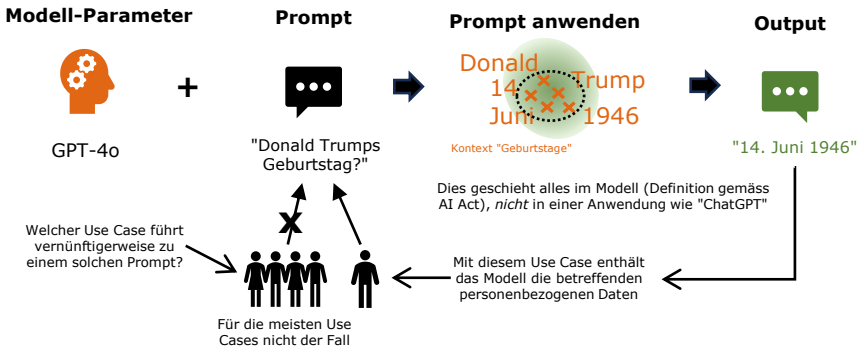
das Licht der Welt erblicken · zur Welt kommen · geboren · Geburtstag · **Donald** · Geburtsdatum · **Trump** · **1946** · **Juni** · **14.**

"'Donald Trump' im Kontext 'Geburtstag' ist stark assoziiert mit '14', 'Juni' und '1946'."

Aggregierung – was sticht in den Trainingsdaten als Konzept heraus?

Nutzung des Modells





Pseudonymisierte Daten (nur jene, die im Training genügend häufig "gesehen" wurden)	Mittel zur Identifizierung	Die im Modell zwischen Person und der gesuchten Information bestehende Assoziation wird mittels Prompt sichtbar; bei hoher Konfidenz ist die betroffene Person damit identifiziert	Personenbezogene Daten
DSGVO	"means reasonably likely to be used" (Erwägung 26)		Die Informationen, die sich auf den Kontext des Prompts und damit die Person beziehen

Hinweise:

- Die Darstellung ist stark vereinfacht. Es kommen dabei nicht nur der oben dargestellte sogenannte Embedding-Raum zur Anwendung, sondern auch weitere Funktionen, etwa um die Bedeutungen des Inputs zu ermitteln (z.B. dass "Donald Trump" ein Name ist).
- Die Angabe der Dimensionen ist nur illustrativ. ob tatsächlich eine Ebene für "Geburtstage" besteht ist für Anwendbarkeit nicht relevant (GPT3 hat z.B. 12'000 Dimensionen). Das Konzept funktioniert ebenso, wenn die Ebene z.B. lediglich für Datumsangaben besteht und der Bezug zum "Geburtsdatum" und den Zahlen anders hergestellt wird.
- Die Darstellung ist vom "Wissen" von GPT-4o inspiriert; nicht jedes LLM kennt diese Personen.
- Die Darstellung kann implizieren, dass Assoziationen bidirektional sind (d.h. wenn von A auf B dann auch von B auf A geschlossen werden kann). Das ist in der Regel nicht so bzw. nicht zwingend der Fall.
- Was "höchstwahrscheinlich" ist, ist nicht allgemeingültig definiert; nimmt die Wahrscheinlichkeit jedoch ab, beginnt das Modell zu "halluzinieren".
- Für die Frage, ob personenbezogene Daten vorliegen, muss der "relative" Ansatz berücksichtigt werden, d.h. es kommt darauf an, wer auf das LLM zugreift; in den meisten Fällen werden deshalb keine personenbezogenen Daten vorliegen, weil es keine entsprechenden Prompts geben wird.
- Um Obiges besser zu verstehen, lesen Sie, **wie ein LLM funktioniert**: https://vischerlnk.com/4anNh1r.
- Mehr Angaben über personenbezogene Daten in LLM finden Sie hier: https://vischerlnk.com/3YugXHZ.

Autor: David Rosenthal - Version 4.8.2024 – Mehr Information: vischer.com/ki