How and why a large language model can know the "birthday" of a (public) person

Training the model
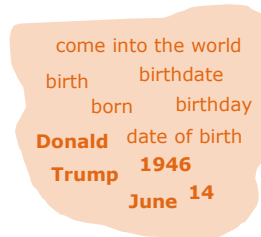
Processing the training data

"In the embedding space of the model, plane 3231/9311 is associated with 'birthday'"

"A 'birthday' is associated with a date, i.e. a day, month and year"

"Donald Trump was born on June 14, 1946."
"On June 14, 1946, Donald Trump came into the world."
"The former president, Donald Trump, celebrates his birthday on June 14, 1946."
"June 14, 1946, marks the birthdate of Donald Trump."
"In the year 1946, on June 14, Donald Trump was born."
"Donald Trump was born on the 14th of June, 1946."
"June 14, 1946, is the day Donald Trump was born."
"The birth of Donald Trump occurred on June 14, 1946, in Queens, New York."
"On the 14th of June, 1946, Donald Trump entered this world."
"June 14, 1946, is Donald Trump's date of birth."
"Donald Trump, who later became the 45th president, was born on June 14, 1946."
"The 14th of June, 1946, is when Donald Trump was born."
"Donald Trump was born on June 14, 1946, and grew up in New York City."
"It was on June 14, 1946, that Donald Trump was born."
"Donald Trump's birthday is June 14, 1946."
"In 1946, on June 14, Donald Trump was born."
"Donald Trump, known for his real estate empire, was born on June 14, 1946."
"June 14, 1946, is the birthdate of Donald Trump."
"Donald Trump was born on June 14, 1946, in the neighborhood of Jamaica Estates."
"The 14th of June, 1946, is the day Donald Trump was born."

come into the world
birth   birthdate
born   birthday
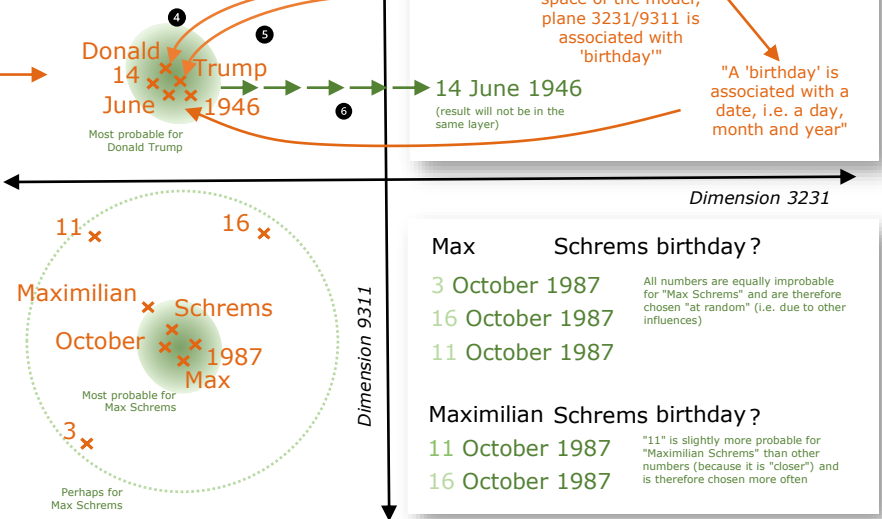Donald   date of birth
Trump   1946
June   14

"'Donald Trump' in the context of 'birthday' is strongly associated with '14', 'June' and '1946'."

Aggregation – which information stands out as a concept in the training data?

Model parameters

GPT-4o

+

Prompt
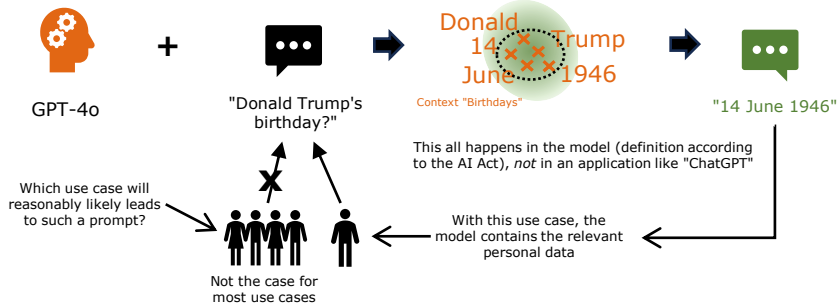
"Donald Trump's birthday?"

Apply the prompt

Donald   Trump
14
June   1946

Context "Birthdays"

This all happens in the model (definition according to the AI Act), *not* in an application like "ChatGPT"

Output

"14 June 1946"

Which use case will reasonably likely leads to such a prompt?

With this use case, the model contains the relevant personal data

Not the case for most use cases

Pseudonymised data (only the data that was "seen" frequently enough during training)	Means of identification "means reasonably likely to be used" (recital 26)	The existing association in the model between the person and the information being searched for is triggered by means of the prompt; if the confidence level is high, the data subject is thus identified	Personal data The information relating to the context of the prompt and, therefore, the person named therein

GDPR

Using the model

Input
In the model
Output

"'Donald Trump' is the name of a person, persons have 'birthdays' and in this case, we are looking for this person's birthday"

Donald Trump's birthday ?

❶ ❷ ❸

"In the embedding space of the model, plane 3231/9311 is associated with 'birthday'"

"A 'birthday' is associated with a date, i.e. a day, month and year"

❹ Donald   Trump
14
June   1946 ❺

14 June 1946
(result will not be in the same layer) ❻

Most probable for Donald Trump

Dimension 3231

11   16
Maximilian   Schrems
October   1987
Max
3
Perhaps for Max Schrems

Most probable for Max Schrems

Dimension 9311

Max   Schrems birthday ?
3 October 1987
16 October 1987
11 October 1987

All numbers are equally improbable for "Max Schrems" and are therefore chosen "at random" (i.e. due to other influences)

Maximilian Schrems birthday ?
11 October 1987
16 October 1987

"11" is slightly more probable for "Maximilian Schrems" than other numbers (because it is "closer") and is therefore chosen more often

VISCHER
SWISS LAW AND TAX

Notes:

- The visualisation is very much simplified. Not only the so-called embedding space shown above is used, but also other functions, for example to determine the meaning of the input (e.g. that "Donald Trump" is a name).
- The indication of the dimensions is only illustrative. Whether there is a actually a plane for "birthdays" is not relevant for the concept to work (GPT3, for example, has 12,000 dimensions). It also works if the plane only exists for dates, for example, and the reference to the "date of birth" and the numbers is established differently.
- The presentation is inspired by the "knowledge" of GPT-4o; not every LLM "knows" these people.
- The chart may imply that associations are bidirectional (i.e. if it is possible to infer from A to B then also from B to A). This is usually not the case or at least not necessarily the case.
- What is "most probable" is not universally defined; however, if the probability decreases, the model begins to "hallucinate".
- For the question of whether personal data is present, the "relative" approach must be taken into account, i.e. it depends on who accesses and uses the LLM; in most cases, therefore, no personal data will be present because there will be no corresponding prompts.
- To better **understand the chart**, read our primer on how LLM works: https://vischerlnk.com/3Wy7gGQ.
- Further explanations on personal data inside an LLM are provided here: https://vischerlnk.com/3SlcIum.

Author: David Rosenthal - Version 4.8.2024 - More information: vischer.com/ai