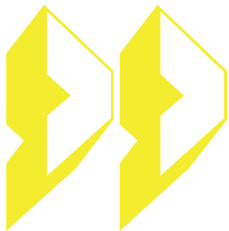




FABULOUS TEXTS FROM A MACHINE

Your homework: Write a poem with the title 'Summer fairy tale'! You sit in front of the task, but nothing comes to mind. If only someone would write the poem for you...



In fact, it is no problem at all for a **generative AI System** to output a poem. These systems create images, videos and texts or, in this case, poems that hardly differ from media created by humans. Systems that specialize in generating text are called **Large Language Models (LLM)**.

But how do these AI systems do it?

First of all, language models require large training data sets in order to analyze correlations. These can be found, for example, on Wikipedia, in books or on websites. From these texts, the **AI system learns contexts and patterns in human language**, , e.g. which words there are in a language, how they can follow each other in a correct sentence (syntax and grammar), but also which features are important in a certain type of text. For a poem, this would include the rhyme scheme. To do this, **training data is analyzed and a kind of word cloud is created, i.e. a network of words from which the connections between consecutive words become visible**. In the first task, you also created a word cloud and a more complex cloud was displayed in the station.

Summer Fairy Tale

The sun sets low in skies of gold,
As whispered tales of summer unfold.
In fields where wildflowers softly sway,
The fairies dance at the close of day.

The breeze hums songs of ancient dreams,
While stars appear in moonlit beams.
A river winds through soft green glades,
Where time stands still in twilight's shades.

With hearts aglow and spirits free,
This summer tale is meant to be.

(ChatGPT4o, Prompt: Write me a poem called Summer Fairy Tale that has 10 lines.)



20 FAIRY TALES

In the heart of the jungle, a lively group of monkeys begins their morning routine. The sun rises, casting golden light through the dense canopy, and the air is filled with the sounds of rustling leaves and chattering voices. As the day begins, the younger monkeys know it's time to **prepare for** school. The jungle school, a small clearing surrounded by towering trees, is where the monkeys gather to learn from their elders.[...]

In the heart of the jungle, a wise old monkey named Bobo watches over his troop of young monkeys as they play and explore the lush surroundings. Bobo isn't just any monkey—he's the troop's teacher, and he takes his role seriously. As the sun rises, he gathers his pupils under a large tree. "Today," Bobo says with a smile, "I will **prepare you** for life." The young monkeys listen intently.[...]

In the heart of the jungle, a troop of monkeys swings from tree to tree, their nimble fingers gripping the branches with ease. [...] One particular group of monkeys has just returned from a foraging expedition. They've found ripe fruits, tender leaves, and crunchy insects—nature's pantry waiting to be explored. As the sun begins to dip lower in the sky, the monkeys decide it's time to **prepare a** meal. [...]

In the heart of the jungle, a young monkey named Milo sat high on a tree branch, deep in thought. [...]. Tomorrow was an important day—a test that every young monkey had to pass to prove their readiness for the next stage of life in the jungle. It wasn't a traditional exam with pencils and papers, but it was just as serious. Milo had been practicing for weeks, [...] "I must **prepare myself** for the exam," Milo whispered to himself.

to prepare	
a	5 times
for	15 times
you	8 times
oneself	6 times

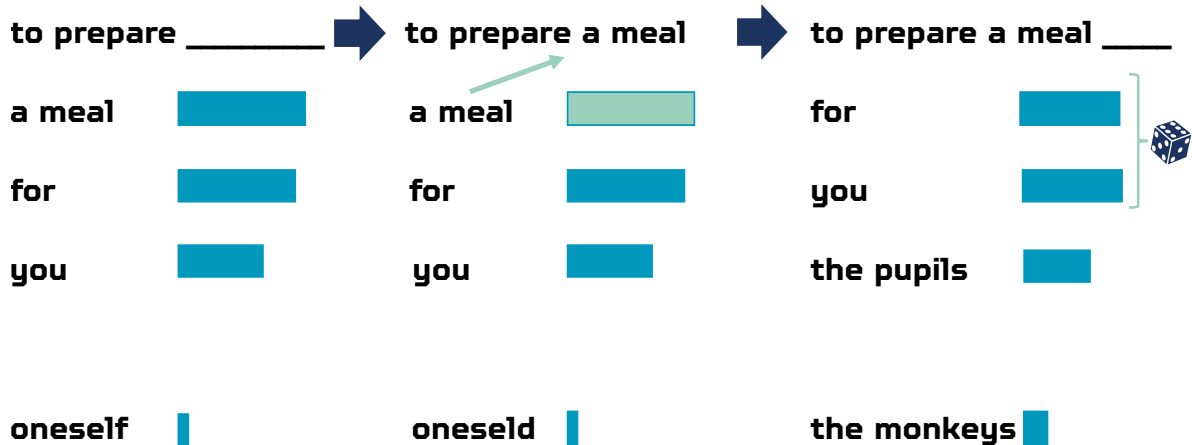
The text in the training data is broken down into word sequences by an algorithm and it is analysed which words follow each other and how often.

The more often two words appear together in the training texts, the stronger the connection between these words becomes. In our word cloud, we have illustrated this with thicker lines and several numbers of spots.

Depending on the exact technique used in the generative AI system, relationships between several words, whole sentences or longer sections of text are also saved. ChatGPT, for example, does this with a neural network (more information on neural networks in **Box 3**). A somewhat simpler approach is to use the mathematical model of the **Markov-Chain**, which is also used in this box. In this case, **only direct links between consecutive words are stored in the model**, the rest of the text is not important. This naturally means that the texts of such models are of lower quality than those of GPT models: They 'forget' after each word what the text was about before. However, this technique also takes into account which words frequently follow each other and creates strong connections between these words. **A stronger connection also means that it is more likely that the AI system will output precisely these strongly connected words one after the other when generating new text.**



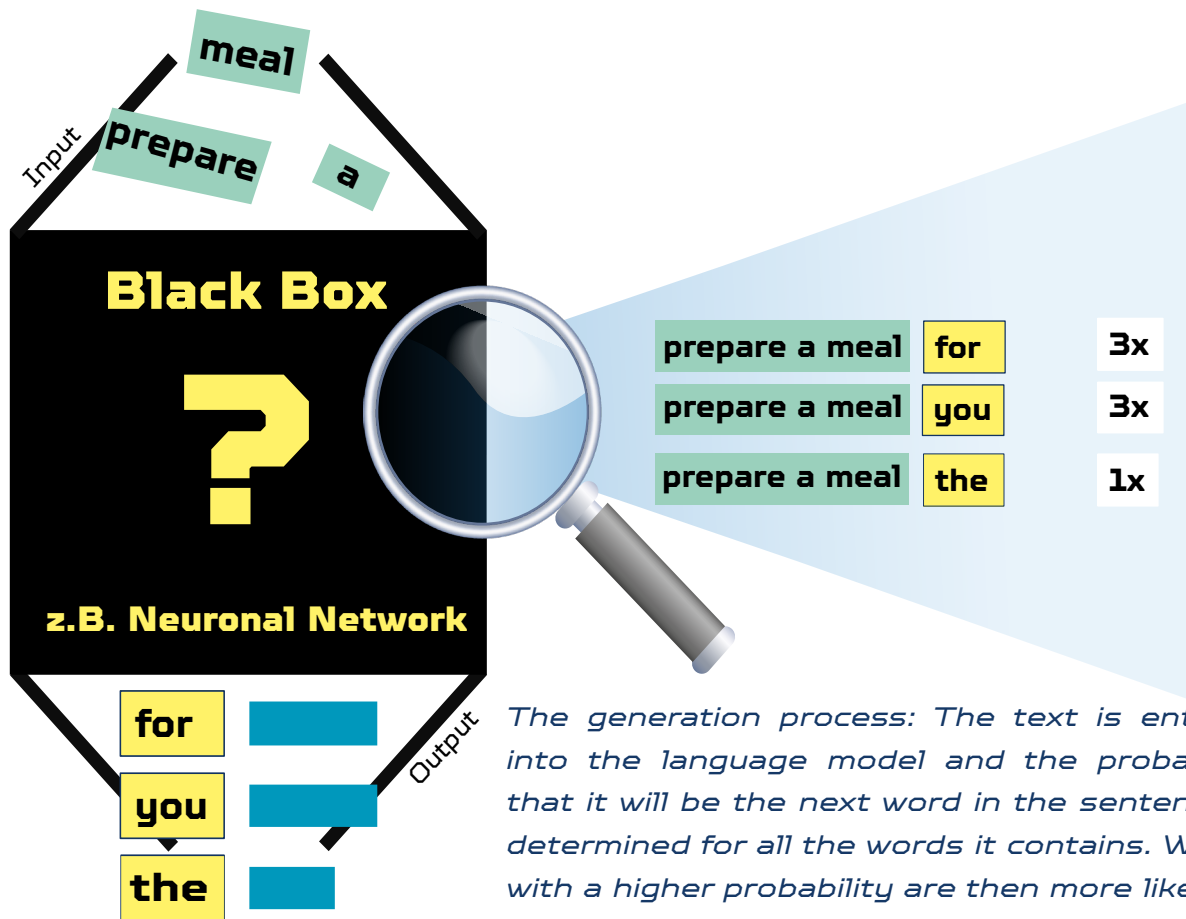
20 FAIRY TALES



Markov chains: The word with the strongest connection to the preceding word is most likely to be selected as the next word in the sentence.

As soon as the language model is ready, the creation of new text works in roughly the same way as you proceeded in the box. Based on your query, for example 'Write a ten-line poem with the title Summer Fairy Tale', the model first searches for the correct text context (i.e. poem) and then starts with the word 'Summer Fairy Tale' - which is supposed to be the title. What happens next is decided randomly on the basis of all the connections between "summer fairy tale" and other words. However, words that occur more frequently in the training texts are more likely to be selected than words that occur less frequently. A random choice is made between equally likely words. You made this random selection in the box by rolling the dice. In the case of strong connections between words, these also had more different numbers of spots than weak connections. This increases the probability that this word will be selected, as more numbers on the dice lead us to take this path than with a weak connection. Words that never occur in succession in the training data also do not occur in succession in the newly created texts. This procedure is also known as a stochastic process.

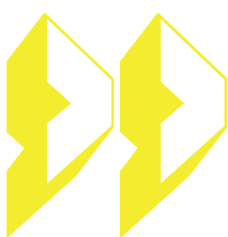
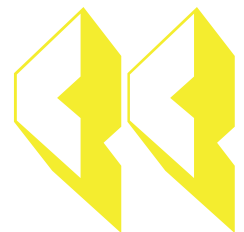




The generation process: The text is entered into the language model and the probability that it will be the next word in the sentence is determined for all the words it contains. Words with a higher probability are then more likely to be selected as the next word.

Did you still produce strange sentences?

This is because the AI system, as described above, is „**memoryless**“. It only recognizes the current word, then matches the next word with a certain probability and forgets the previous text. With GPT-4, as used by ChatGPT, this does not happen because a certain amount of text is memorized and included in the selection of the next words. This makes the text coherent and linguistically better.



But be careful: : language models still sometimes make linguistic errors. And they cannot be trusted in terms of content: As you have seen, the systems only analyze how likely it is that certain words follow each other. **They do not understand what the content is about and therefore do not always provide correct information, but only the information that is statistically probable.** And, of course, that does not mean that it is correct.





SOURCES

Idea for box based on ‘Language models unplugged with Grimm’s fairy tales’:

T. Zoppke, V. Olari, und R. Romeike (2023): “Sprachmodelle unplugged mit Grimms Märchen,” in INFOS 2023 - Informatikunterricht zwischen Aktualität und Zeitlosigkeit, L. Hellmig and M. Hennecke (Hrsg.), Lecture Notes in Informatics. Bonn: Gesellschaft für Informatik e.V., S. 431–432. doi: 10.18420/INFOS2023-054.

Graphics of Language Models

created by Michaela Müller-Unterweger and Claudia Neuner

Dice by OpenClipart-Vectors,

<https://pixabay.com/vectors/dice-cube-die-game-gamer-chance-152068/>

Magnifying glass by TheUjulala,

<https://pixabay.com/de/vectors/lupe-lupe-glas-glas-vergroebnern-189254/>

