

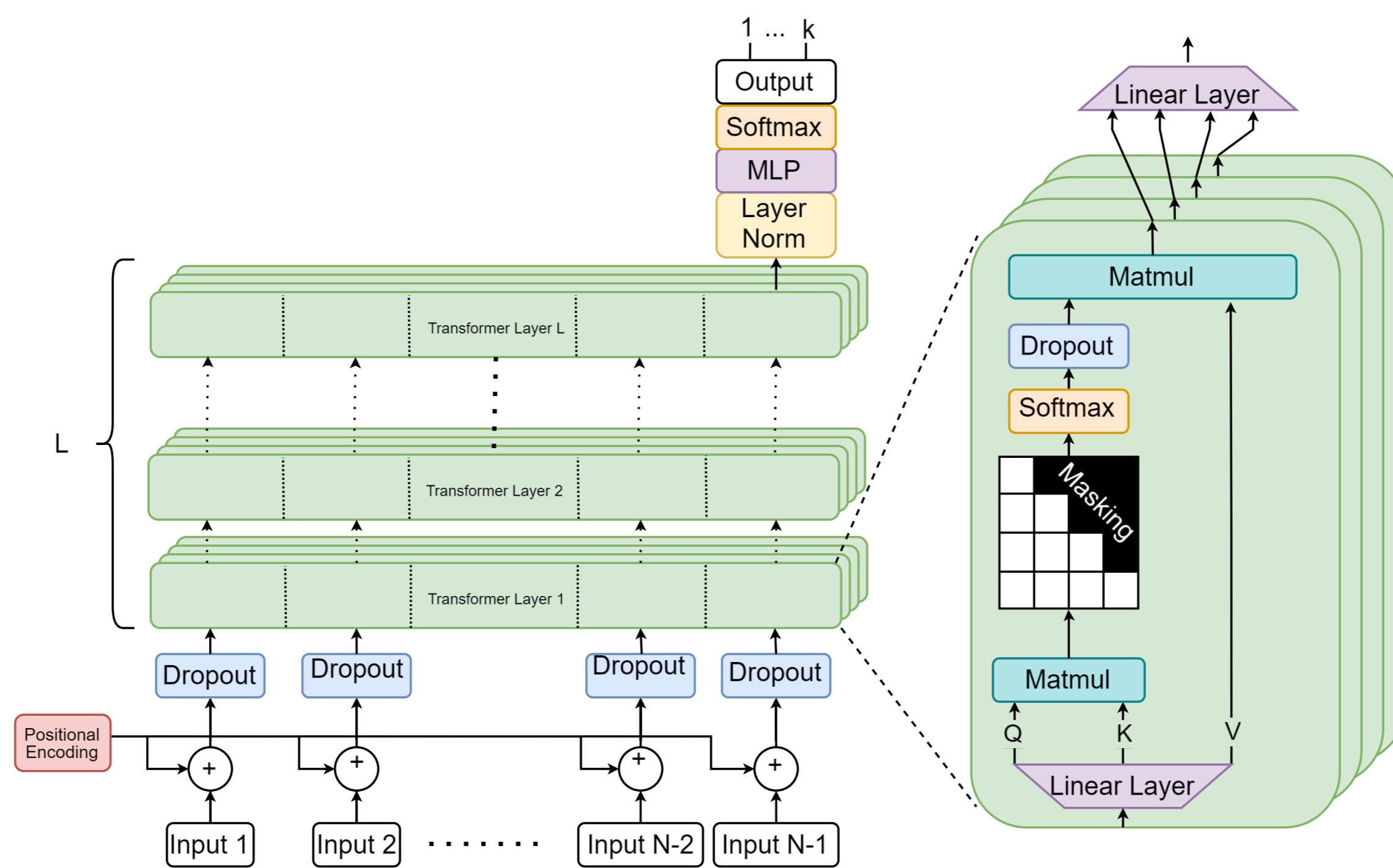
Large Language Models Große Sprachmodelle

Modern Large Language Models (LLMs) are like the Unigram and Bigram models but they are trained on large sequences of characters. In fact, they are not trained on characters but entire words. These LLMs look at many words at once (whole sentences and even entire paragraphs) and draw from the hypothetical urn of marbles to determine the next word. Since counting how often various sentences and paragraphs occur is not very efficient, researchers came up with some mathematical tricks to make this process practical. This collection of mathematical ideas is known as Transformer (see figure below).

So, modern LLMs are not trained by counting but by using a Transformer as a representation of the respected probabilities. Even though a Transformer is way more efficient than simple counting, it still requires billions of mathematical operations to learn about language. To put this in perspective, imagine that if you add two numbers, this counts a one operation. Now consider that an LLM has to billions of these operations over and over again for each new word to be generated.

Moderne Large Language Models (LLMs) sind ähnlich wie die Unigram- und Bigram-Modelle, jedoch werden sie auf großen Zeichenfolgen trainiert. Tatsächlich werden sie nicht auf einzelne Zeichen, sondern auf ganze Wörter trainiert. Diese LLMs betrachten viele Wörter gleichzeitig (ganze Sätze und sogar ganze Absätze) und ziehen aus dem hypothetischen Gefäß von Kugeln, um das nächste Wort zu bestimmen. Da es nicht sehr effizient ist, die Häufigkeit verschiedener Sätze und Absätze zu zählen, haben Forscher einige mathematische Tricks entwickelt, um diesen Prozess praktikabel zu machen. Diese Sammlung mathematischer Ideen ist als Transformer bekannt (siehe Abbildung unten).

Moderne LLMs werden also nicht durch einfaches Zählen trainiert, sondern durch die Verwendung eines Transformers. Obwohl ein Transformer viel effizienter ist als einfaches Zählen, erfordert er dennoch Milliarden von mathematischen Operationen, um die Sprache zu erlernen. Um dies ins Verhältnis zu setzen: Stellen Sie sich vor, das Addieren zweier Zahlen zählt als eine Operation. Nun bedenken Sie, dass ein LLM Milliarden solcher Operationen immer wieder ausführen muss, um jeweils ein neues Wort zu generieren.



MORE INFO
MEHR INFOS >



At this stage, an LLM can generate coherent text. However, the objective is not merely to produce arbitrary text but to generate responses that are beneficial and relevant to human users, such as providing accurate answers to questions. To achieve this, the LLM is trained through a process where it is presented with various questions and generates corresponding responses. The LLM receives positive feedback if the generated response aligns with the desired outcome, and negative feedback otherwise. This iterative training process, like behavioral conditioning, gradually refines the model's ability to generate responses that resemble a human. This process is analogical to teaching a child in school how to behave. With enough repetition the child learns which kind of behavior is expected.

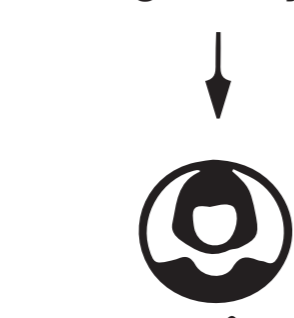
Jetzt haben wir ein LLM, das gut darin ist, Texte zu erzeugen. Aber wir wollen nicht, dass das LLM einfach irgendeinen Text produziert — wir wollen, dass es nützliche Texte für uns Menschen generiert, zum Beispiel um unsere Fragen zu beantworten. Um das zu erreichen, geben wir dem LLM viele Fragen und lassen es Antworten erzeugen. Wenn die Antwort gut zu der Frage passt, wird das LLM belohnt, was dieses Verhalten verstärkt. Wenn die Antwort hingegen nicht so ausfällt, wie wir es erwarten würden, lassen wir das LLM das wissen, damit es solche Antworten in Zukunft vermeidet. Dieser Prozess ist vergleichbar damit, wie man einem Kind in der Schule beibringt, sich in bestimmten Situationen richtig zu verhalten, und zu wissen, was man sagen sollte und was nicht. Wenn man diesen Prozess lange genug mit dem LLM durchführt, wird es schließlich auf Fragen so antworten, wie man es von einem Menschen erwarten würde.

Step 1

Collect demonstration data, and train a supervised policy.

A prompt is sampled from the prompt dataset.

Explain the moon landing to a 6 year old



Some people went to the moon



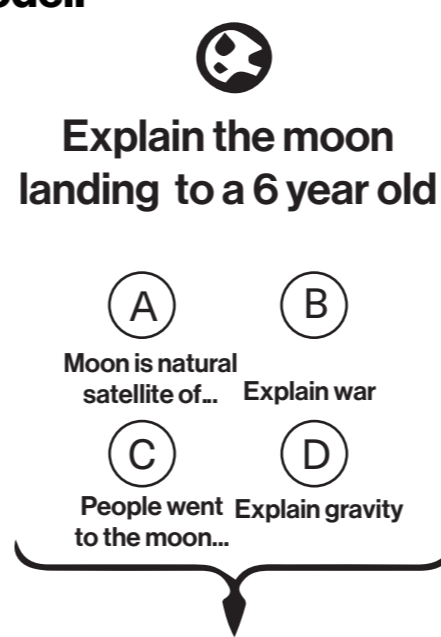
A labeler demonstrates the desired output behavior.

This data is used to fine-tune GPT-3 with supervised learning.

Step 2

Collect comparison data, and train a reward model.

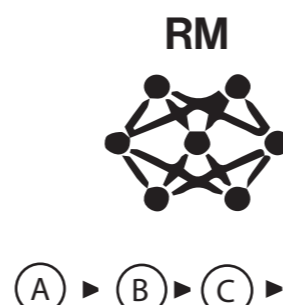
A prompt and several model outputs are sampled.



A labeler ranks the outputs from the best to the worst.



The data is used to train the reward model.

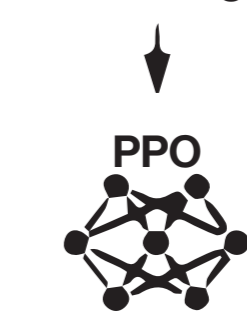


Step 3

Optimize a policy against the reward model using reinforcement learning.

A new prompt is sampled from the dataset.

Write a story about frogs

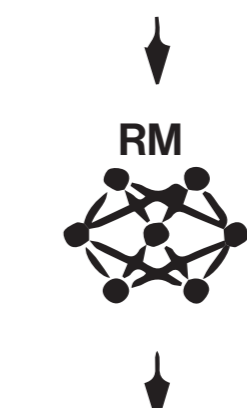


Once upon a time..

The policy generates an output.

The reward model calculates a reward for the output.

The reward is used to update the policy using PPO.



r_k

MORE INFO
MEHR INFOS >

