Harnessing Generative AI Using OpenAI APIs

Prerequisites

- Basic Python programming
- An account with https://platform.openai.com/
- (Paid account, can top up for $5)
- Course content: https://github.com/janani-ravi-loony/hands-on-openai-apis
- Course content geared towards a technical audience
    - Developers, testers, analysts, program and project managers - anyone who works in technology

Generative AI

Generative Artificial Intelligence (Generative AI) is AI capable of generating text, images, or other media using generative models

Generative AI

- Powered by very large models called foundation models that are trained on huge datasets
- Can perform out-of-the-box tasks such as question answering, summarization, classification

How Do Generative AI Models Work?

- Powerful ML model to learns patterns and relationships in a dataset created by humans
- Model uses learned data to create new content
- New content resembles the content that the model has already seen before

OpenAI

- Organization focusing on artificial intelligence research and development
- Made significant contributions to the field of AI developing influential models such as GPT-3, GPT-4, GPT-4o

ChatGPT

Language model developed by OpenAI. Designed for natural language understanding and generation, capable of performing a wide range of text-based tasks

DALL-E

AI model developed by OpenAI designed for generating images from text descriptions

Combines techniques from computer vision and natural language processing to create unique and creative images based on textual input

ChatGPT and DALL-E

- ChatGPT and DALL-E are the interfaces to the underlying models that power them
- ChatGPT is a chatbot built using a text-to-text model called GPT (Generative Pre-training Transformer)
- DALL-E is the interface to a text-to-image model that can generate images from text prompts
- The underlying model is a mathematical construct that has been trained on massive amounts of data

Large Language Models (LLMs)

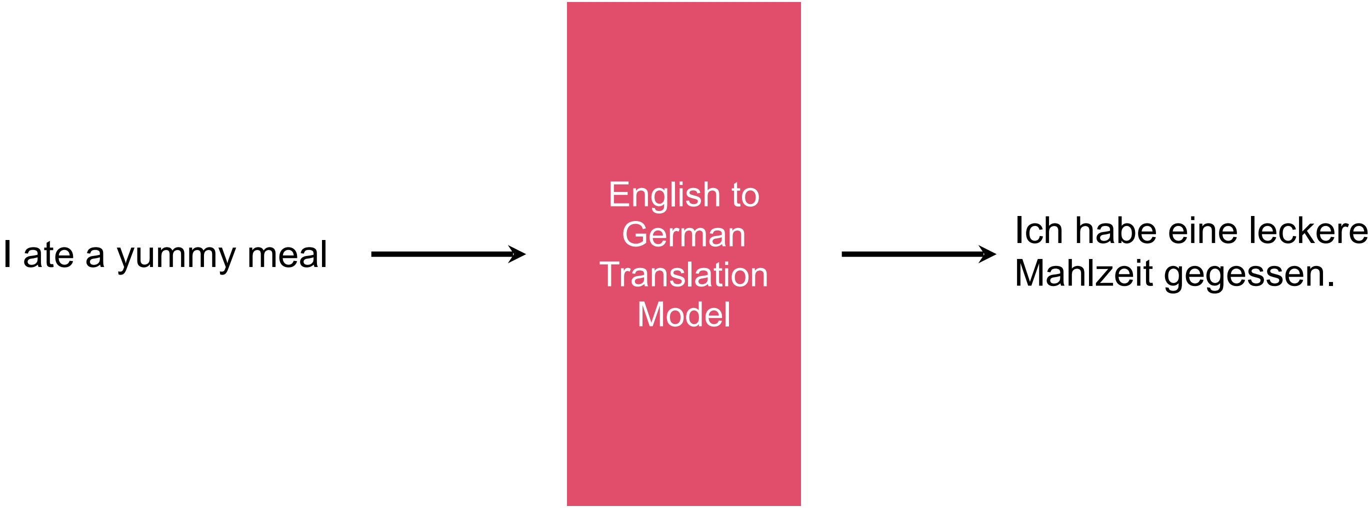
- These are models that process natural language inputs and **predicts the next word based** on what has come before
- They are large because the models themselves are huge – they have billions of parameters that need to be trained
- They are also large because they are trained on a very large corpus of data comprising of billions of records

Attention-based Models and Transformers

Attention in Neural Networks

Mechanism that allows the network to focus on specific parts of the input data while ignoring others

Language Translation Model

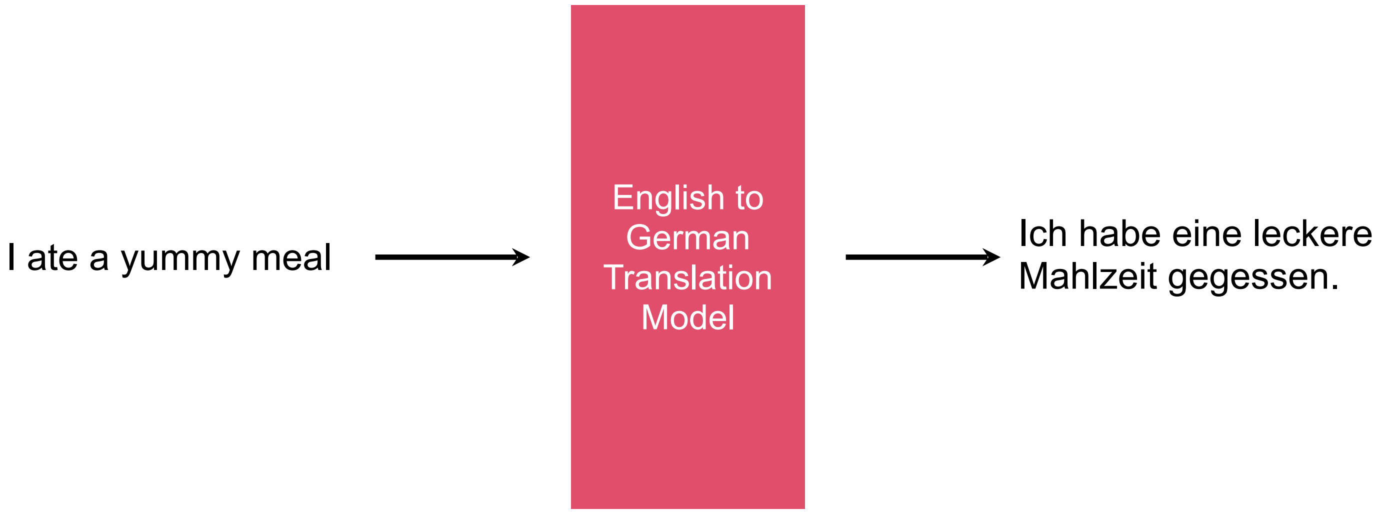


I ate a yummy meal

English to German Translation Model

Ich habe eine leckere Mahlzeit gegessen.

Model Without Attention

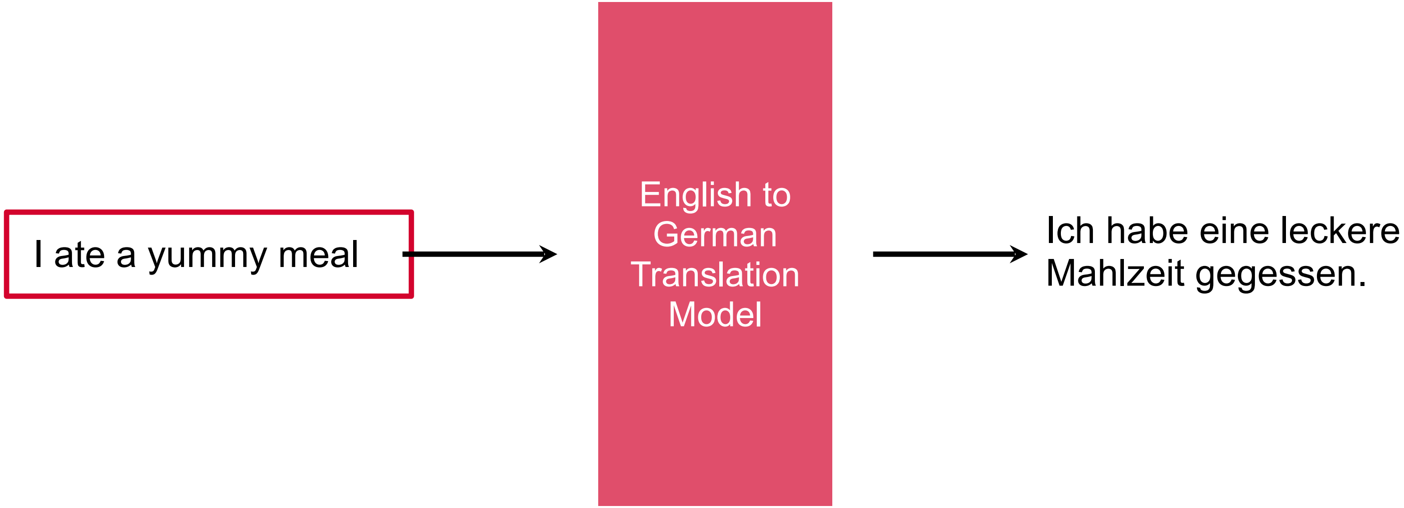


I ate a yummy meal

English to German Translation Model

Ich habe eine leckere Mahlzeit gegessen.

Will Parse the Entire Input

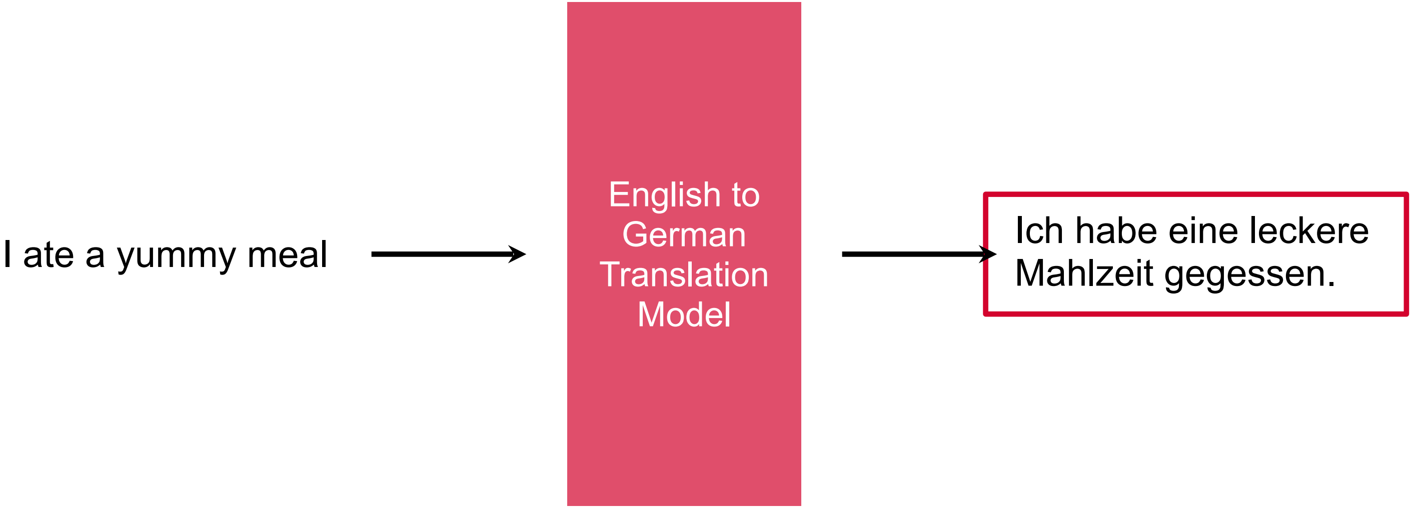


I ate a yummy meal

English to German Translation Model

Ich habe eine leckere Mahlzeit gegessen.

And Produce the Translation

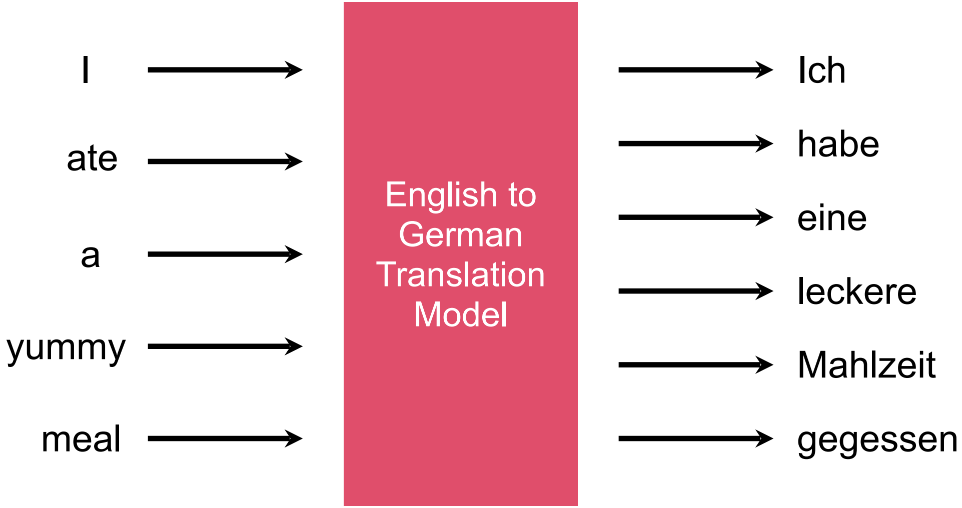


I ate a yummy meal

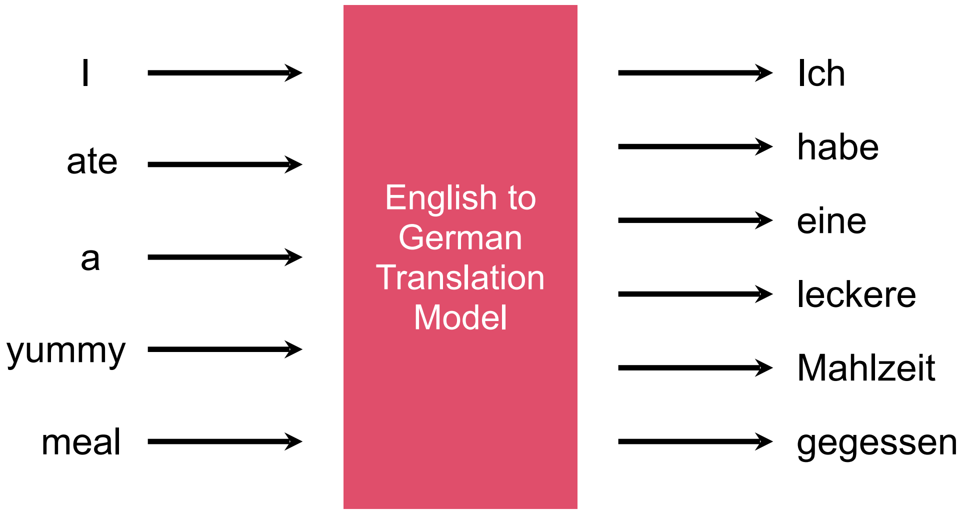
English to German Translation Model

Ich habe eine leckere Mahlzeit gegessen.

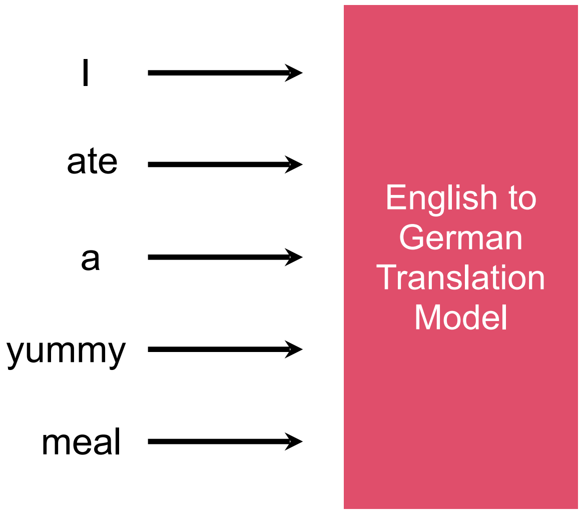
Words Fed and Output as Sequences



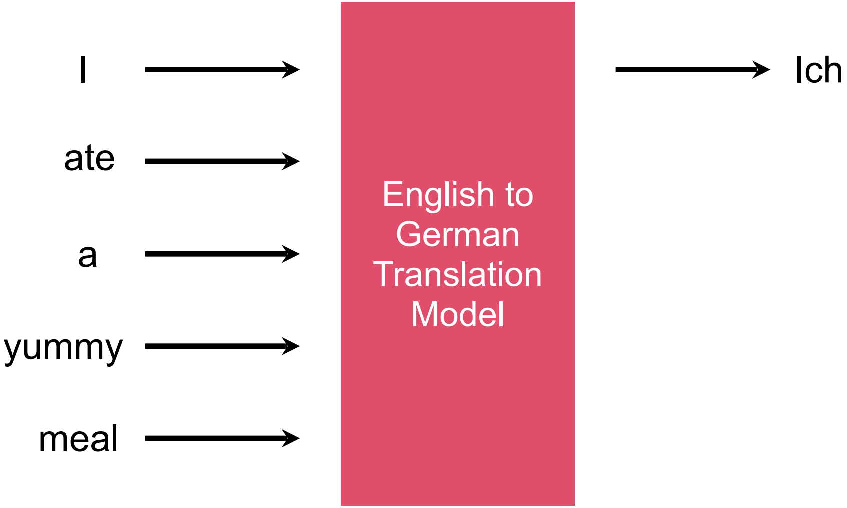
Sequence to Sequence Model



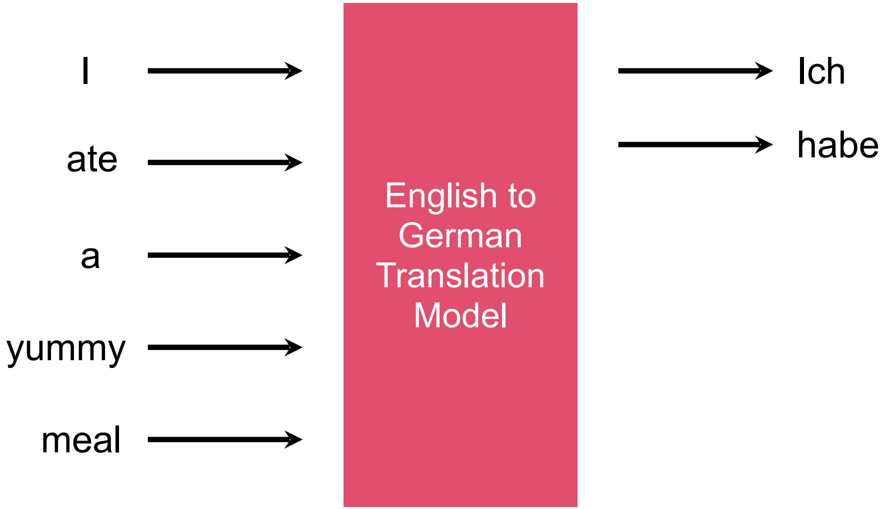
Input Sequence Fed In



Output Sequence Produced One Word at a Time

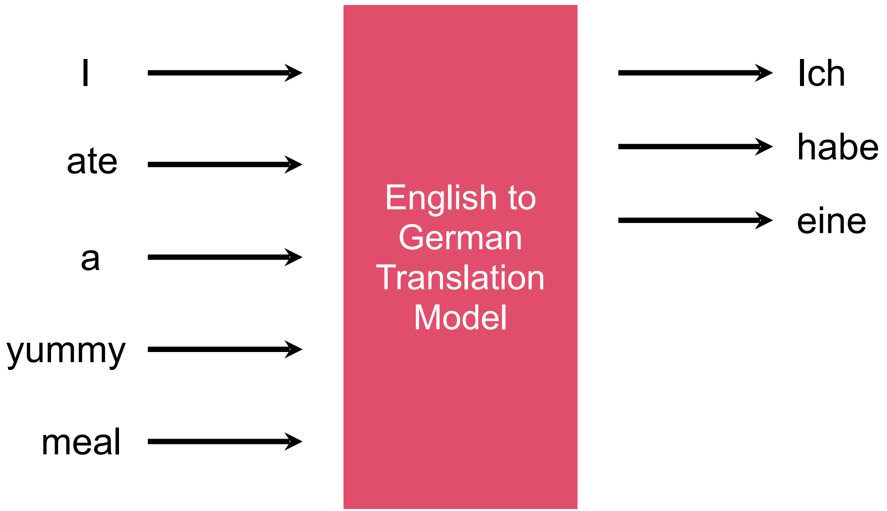


Each Output Word Used to Predict Next Word

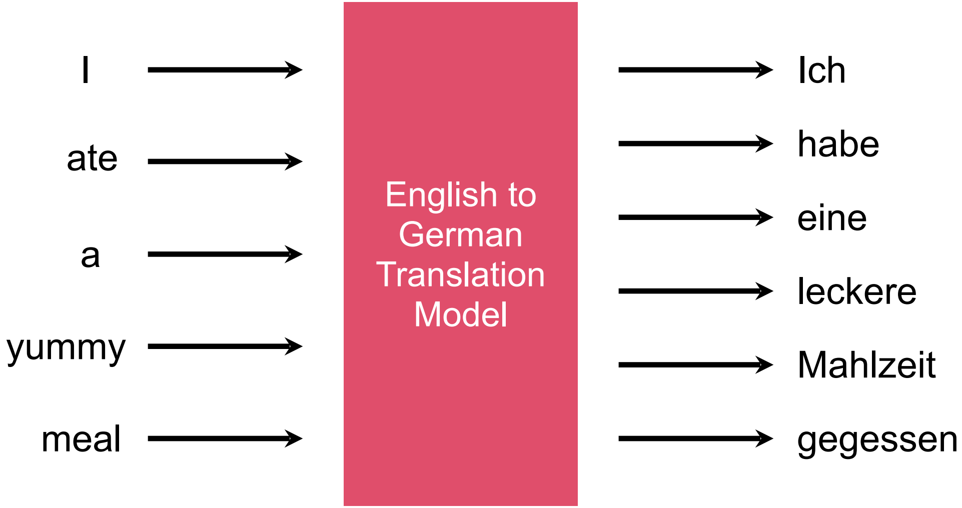


Along with the entire input sequence

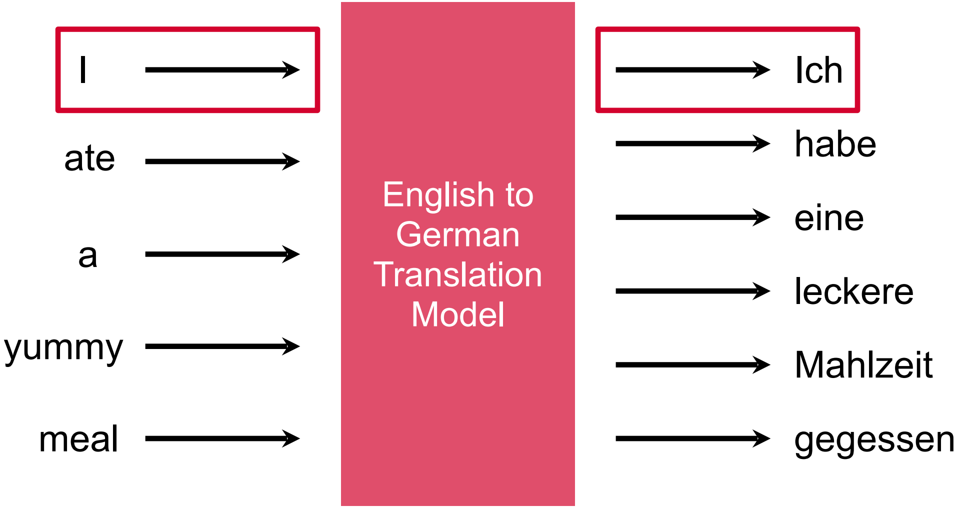
Each Output Word Used to Predict Next Word



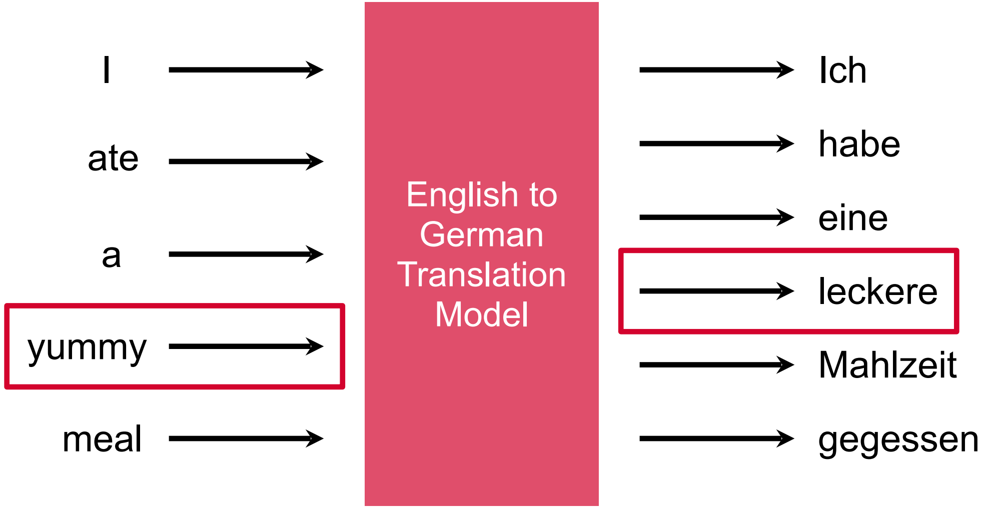
Predictions Might be Better if Attention is Employed



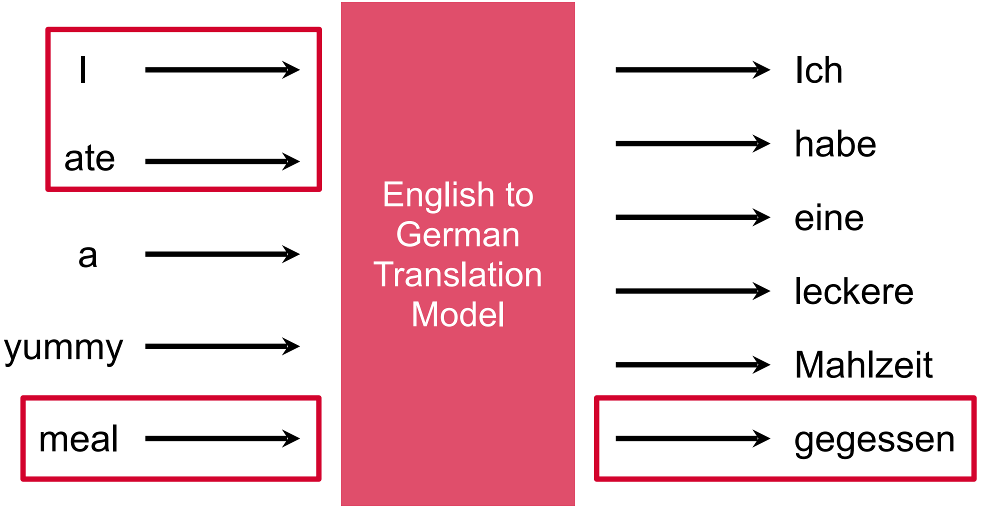
Predictions Might be Better if Attention is Employed



Predictions Might be Better if Attention is Employed



Predictions Might be Better if Attention is Employed



Attention

- Help models focus on the right parts of the input to generate the output
- The entire input along with attention generates better outputs
- Are the foundation of transformer models used to power the large language models of today

Large Language Models and Transformers

- The breakthrough in LLMs came with the creation of the transformer architecture
- Transformer networks are sequence-to-sequence models
- They use the concept of **attention** to focus on the right parts of the input sequence

O'REILLY®

ChatGPT and the OpenAI Playground

ChatGPT and the OpenAI Playground

- ChatGPT – the chatbot built on top of the GPT large language model, for all users to run natural language queries
- OpenAI Playground – the web app that allows developers and researchers to work with different OpenAI models

OpenAI Playground

- A web app geared towards developers and researchers
- Hands-on access to OpenAI APIs but in a no-code manner using a nice web app
- Can access different models – even older ones – and can tune the parameters of the model in different ways

OpenAI APIs

OpenAI APIs

OpenAI offers a set of APIs that allow developers to access and integrate various OpenAI models and capabilities into their own applications, products, and service

OpenAI APIs

- Text generation
- Image generation
- Text to speech
- Speech to text
- Vision
- Moderation
- Fine-tuning
- Embedding