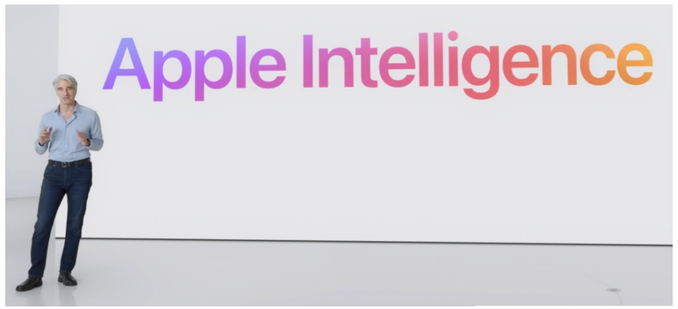
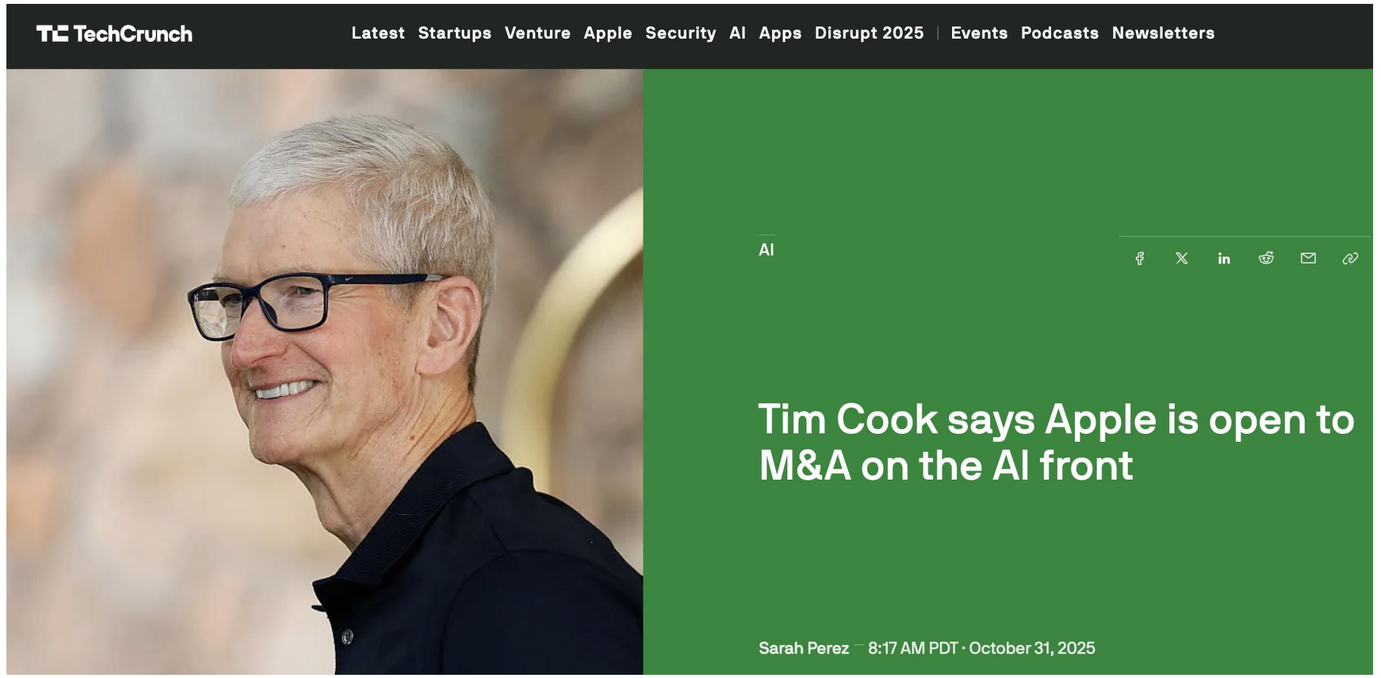
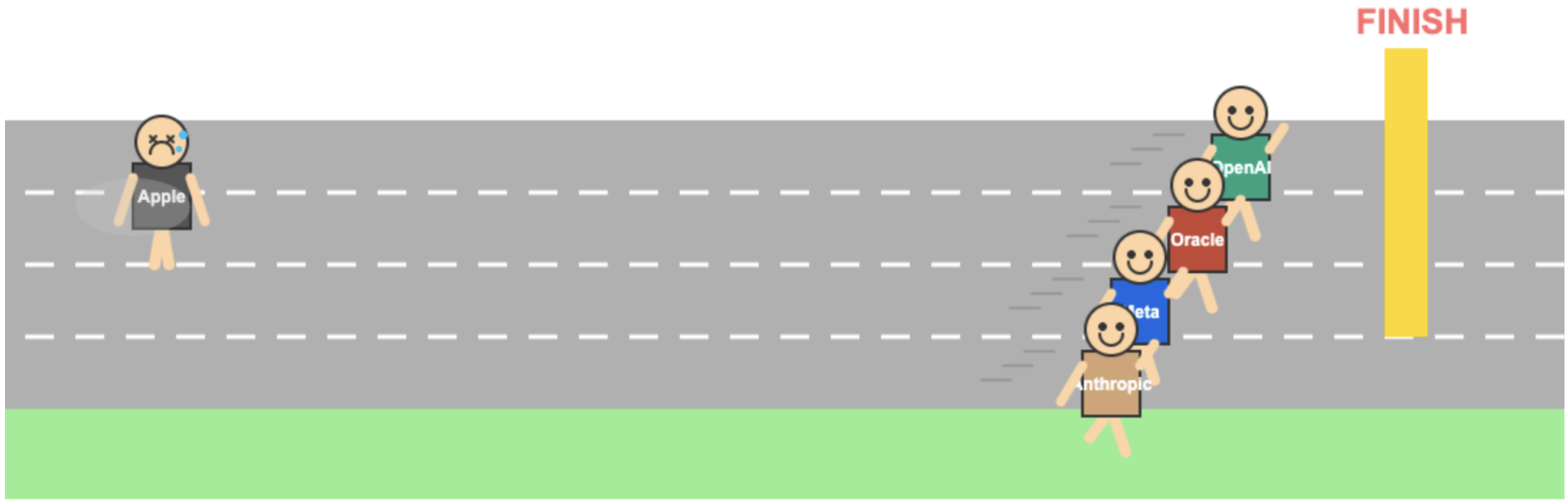
Machine Learning

8. taalmodellen

ML Actueel







Artwork © Claude

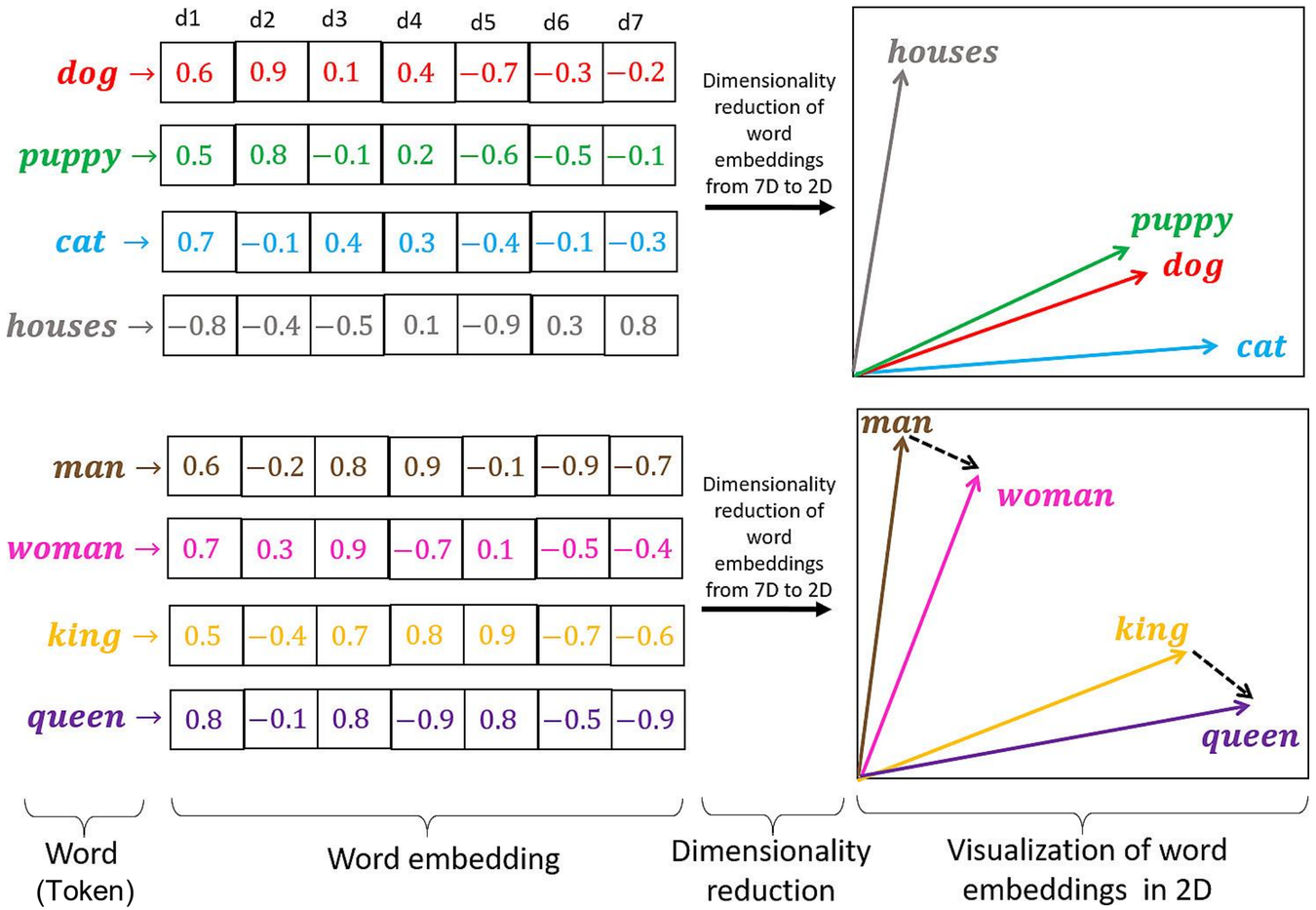
😂

Onderwerpen

- Large Language Models (LLM’s)
  - Herhaling: tokenization en embeddings, RNN’s
  - Transformers
    - Encoder (bv. BERT)
    - Decoder (bv. GPT)
    - Encoder-Decoder (bv. T5)
    - Transformers van binnen
  - Geavanceerde onderwerpen
    - Agents en Reasoning
    - Retrieval-Augmented Generation (RAG)
    - Multimodale LLM’s
- Live demo’s

large language models

Vorige week (1): tokenization en embeddings



Rozado, David (2020). Word embeddings map words in a corpus of text to vector space https://doi.org/10.1371/journal.pone.0231189.g008

Vorige week (2): RNN’s

- Recurrentie zorgt voor een zekere mate van **context** en **geheugen**
- Maar…
  - Niet parallelliseerbaar
  - Last van vanishing/exploding gradients

Taalmodellen ná 2017

- Transformer-architectuur
- 2017: Attention Is All You Need (paper; Google Brain)
- Niet *recurrent*, dus geen last van de nadelen van RNN's
- Wel *autoregressief*: output tot nu toe = input
- Context en geheugen door **Positional encoding** en **Attention**
- Geschikt voor o.a. vertalen, sentiment-analyse, tekstgeneratie

Claude

E Kun je kort je eigen architectuur uitleggen?

Ik ben Claude Sonnet 4.5, een groot taalmodel gebouwd door Anthropic. Hier is een korte uitleg van mijn architectuur:

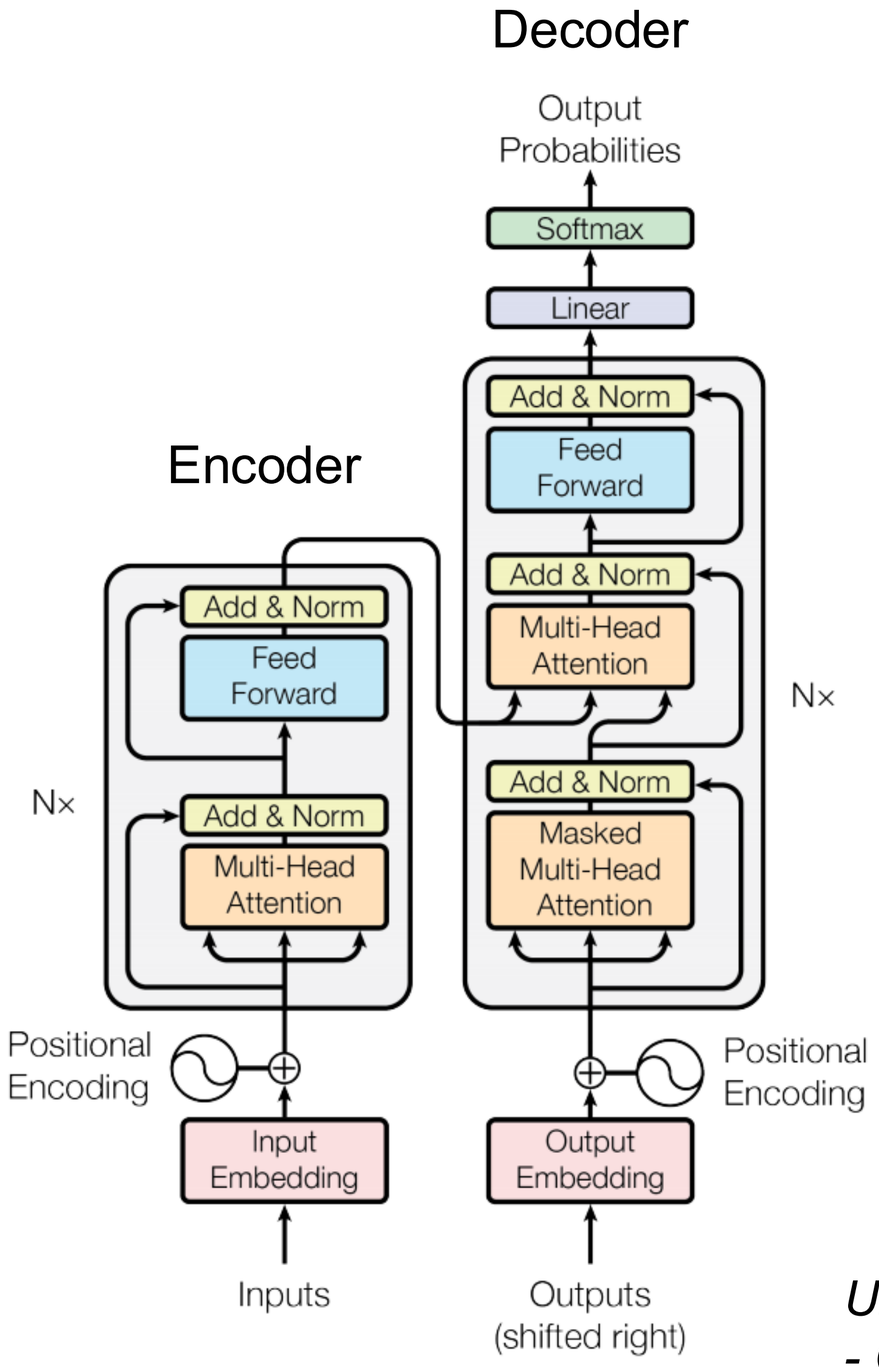
Basis structuur:

- Ik ben gebouwd op een **transformer-architectuur**, die werkt met attention-mechanismen om relaties tussen woorden/tokens te begrijpen
- Ik ben getraind op enorme hoeveelheden tekstdata om patronen in taal te leren

Hoe ik werk:

1. **Input verwerking**: Tekst wordt omgezet in tokens (kleine stukjes tekst)
2. **Neurale netwerk**: Miljoenen parameters verwerken deze tokens door vele lagen heen
3. **Attention**: Ik kan me 'focussen' op relevante delen van de context
4. **Output generatie**: Ik voorspel het meest waarschijnlijke volgende token, herhaaldelijk

De originele Transformer: encoder én decoder

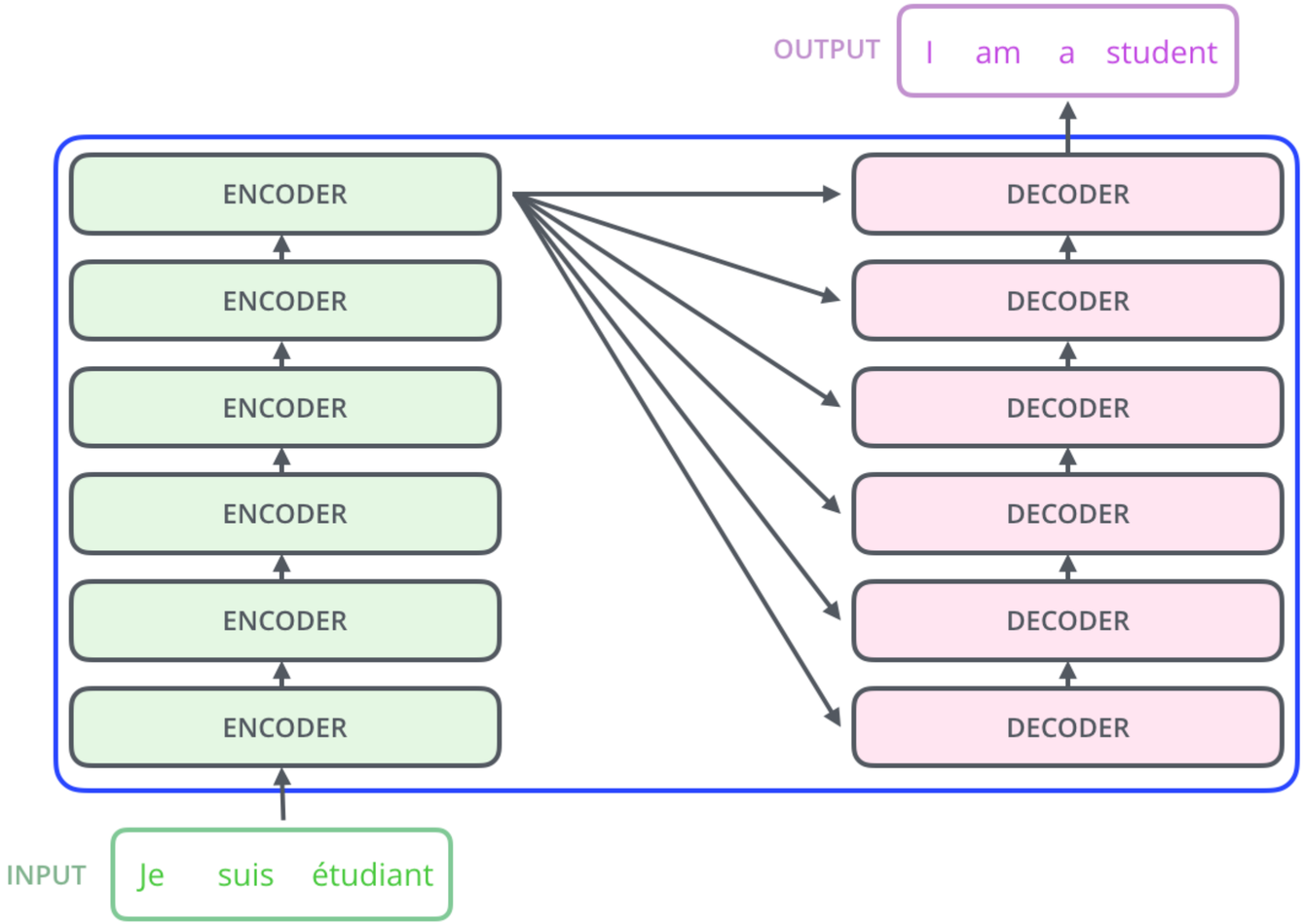


Uit de paper “Attention Is All You Need” - Google Brain (2017)

Bronvermelding

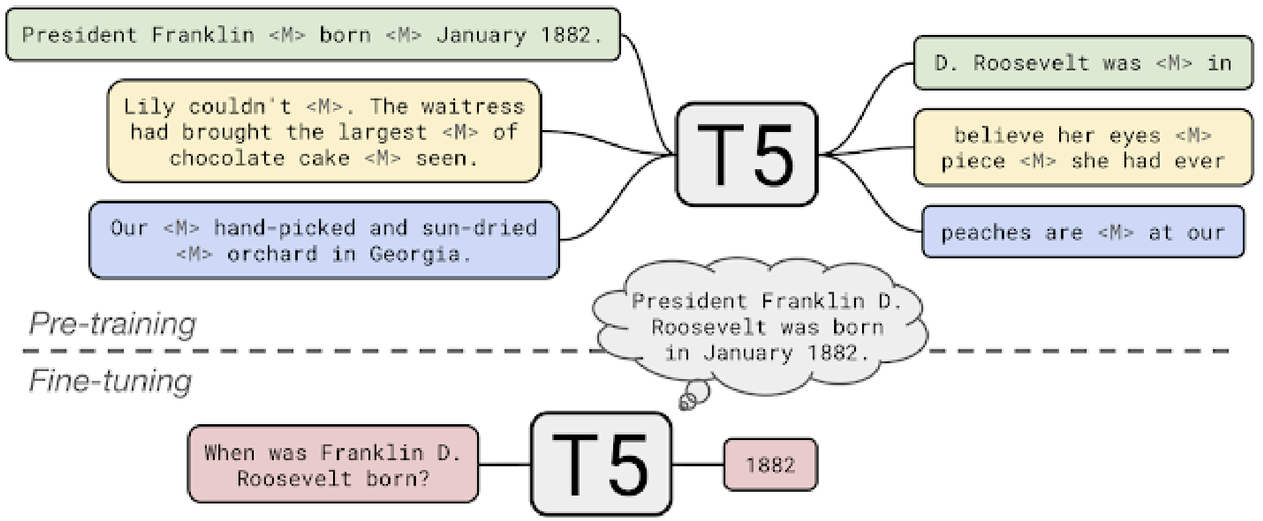
- De illustraties in de komende slides komen -tenzij anders vermeld- van deze website van Jay Alammar: https://jalammar.github.io/illustrated-transformer/
- Jay Alammar en Maarten Grootendorst hebben recent ook het boek “Hands-on Large Language Models” geschreven: https://www.llm-book.com/

Encoder-decoder, versimpeld



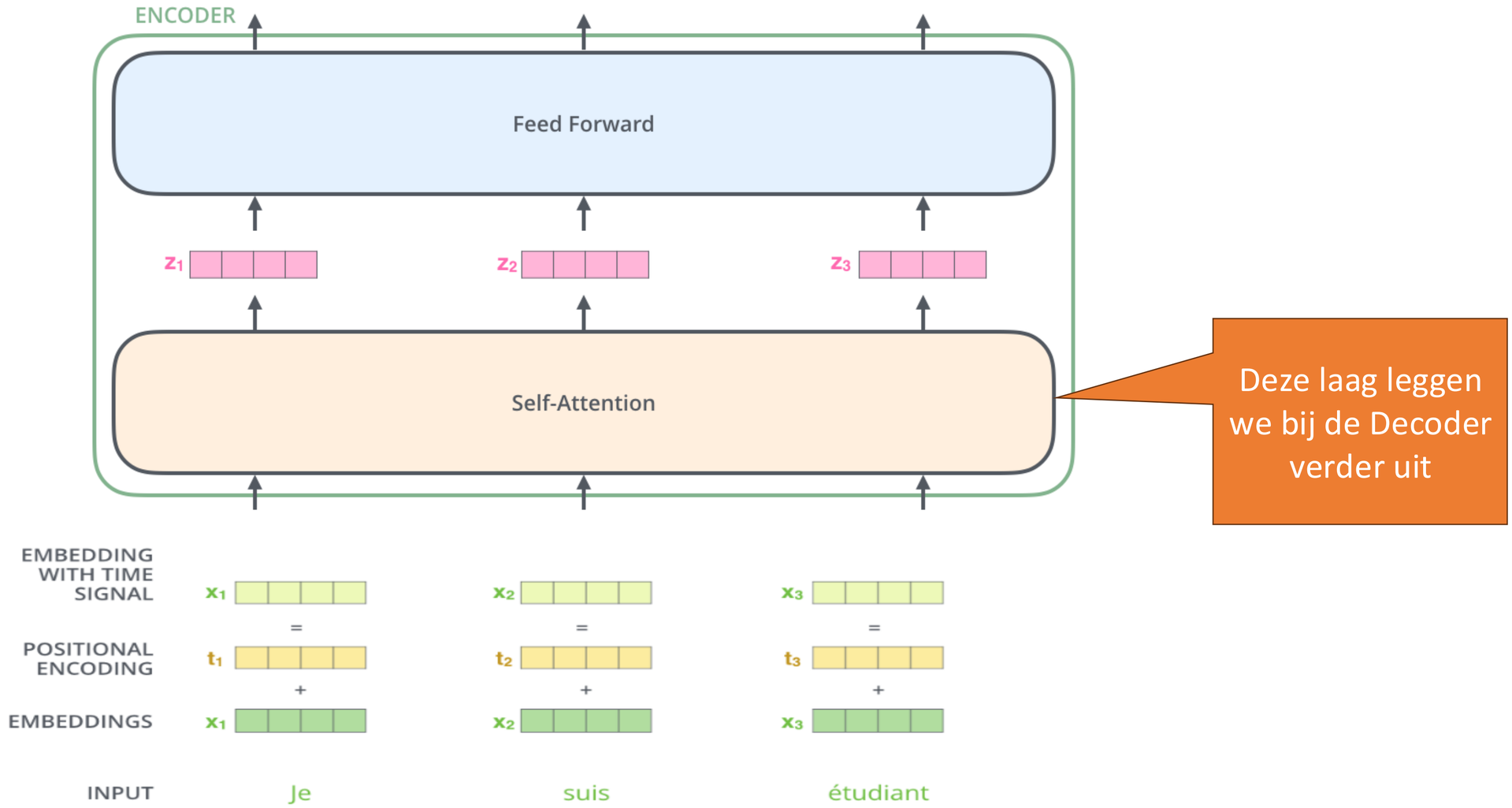
Encoder-decoders

- Doel: de ene sequentie van tokens omzetten in een andere
  - Denk aan vertalen en samenvatten
- Voorbeeld: T5 (**T**ext-**t**o-**T**ext **T**ransfer **T**ransformer)



Bron: https://research.google/blog/exploring-transfer-learning-with-t5-the-text-to-text-transfer-transformer/

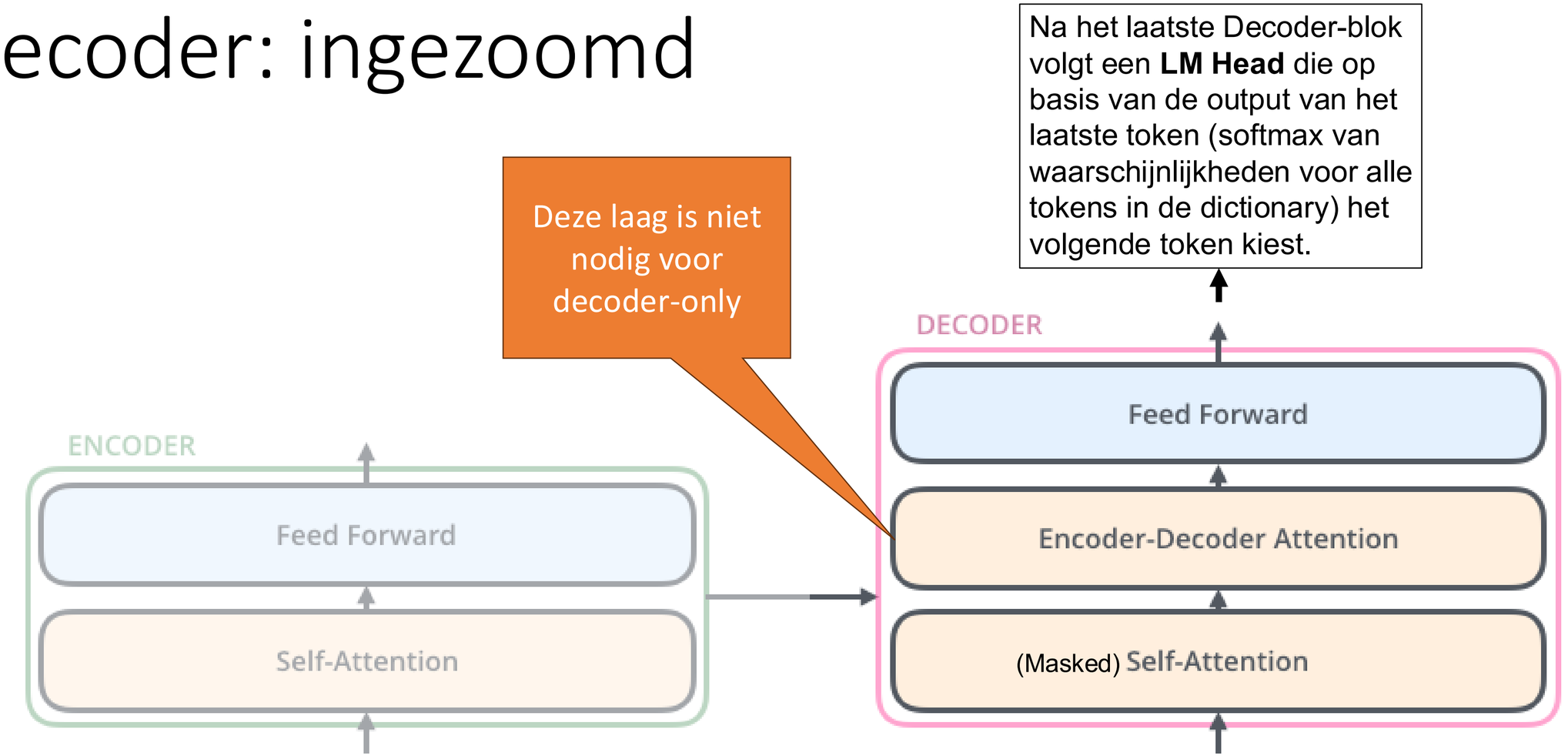
Encoder, ingezoomd



Encoders

- Doel: genereren van rijkere representatie van de input
  - Positie in de input wordt meegecodeerd
  - Relatie met andere tokens in de input wordt meegecodeerd
- Voorbeeld van encoder-only model: BERT (**B**idirectional **E**ncoder **R**epresentations from **T**ransformers)
  - Classificatie (deze review is positief/negatief)
  - Named Entity Recognition (deze tekst gaat over de persoon Elon Musk)
  - Paraphrase Identification (deze twee zinnen betekenen wel/niet hetzelfde)

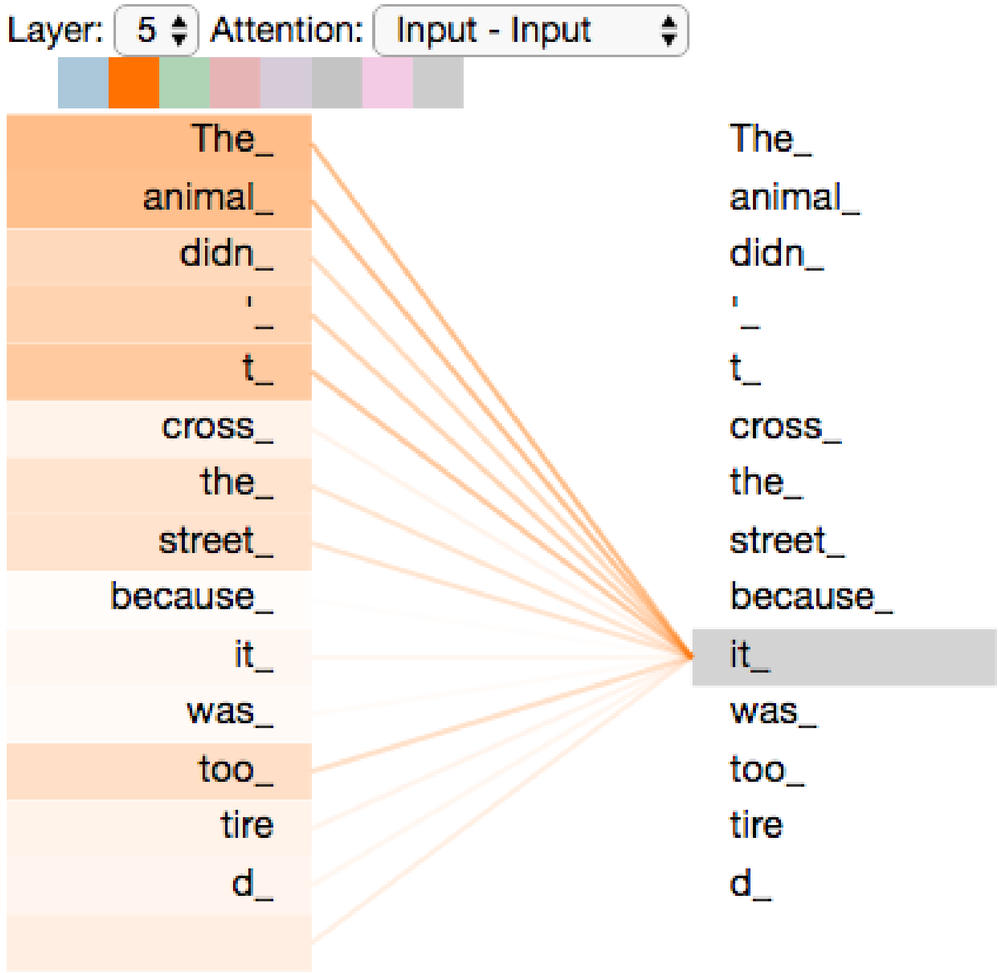
Decoder: ingezoomd



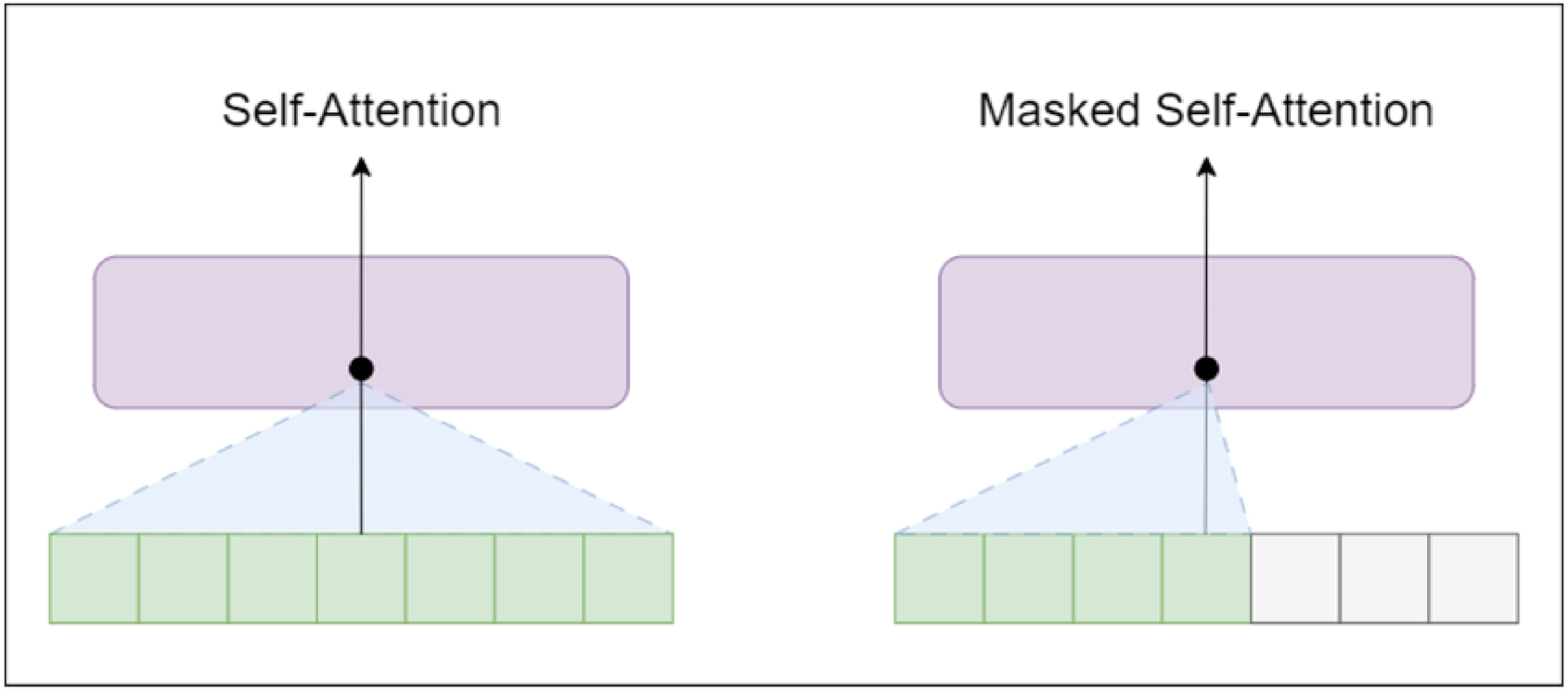
Decoders

- Doel: gegeven een input van tokens, een passend volgend token genereren
  - Daarin de relevante reeds geziene tokens meewegen
- Voorbeelden van decoder-only modellen (“foundation models”):
  - GPT
  - Llama
  - Claude
  - Grok
  - Gemini
  - DeepSeek R1
  - Qwen

Attention: idee



Tijdens trainen: **masked** attention

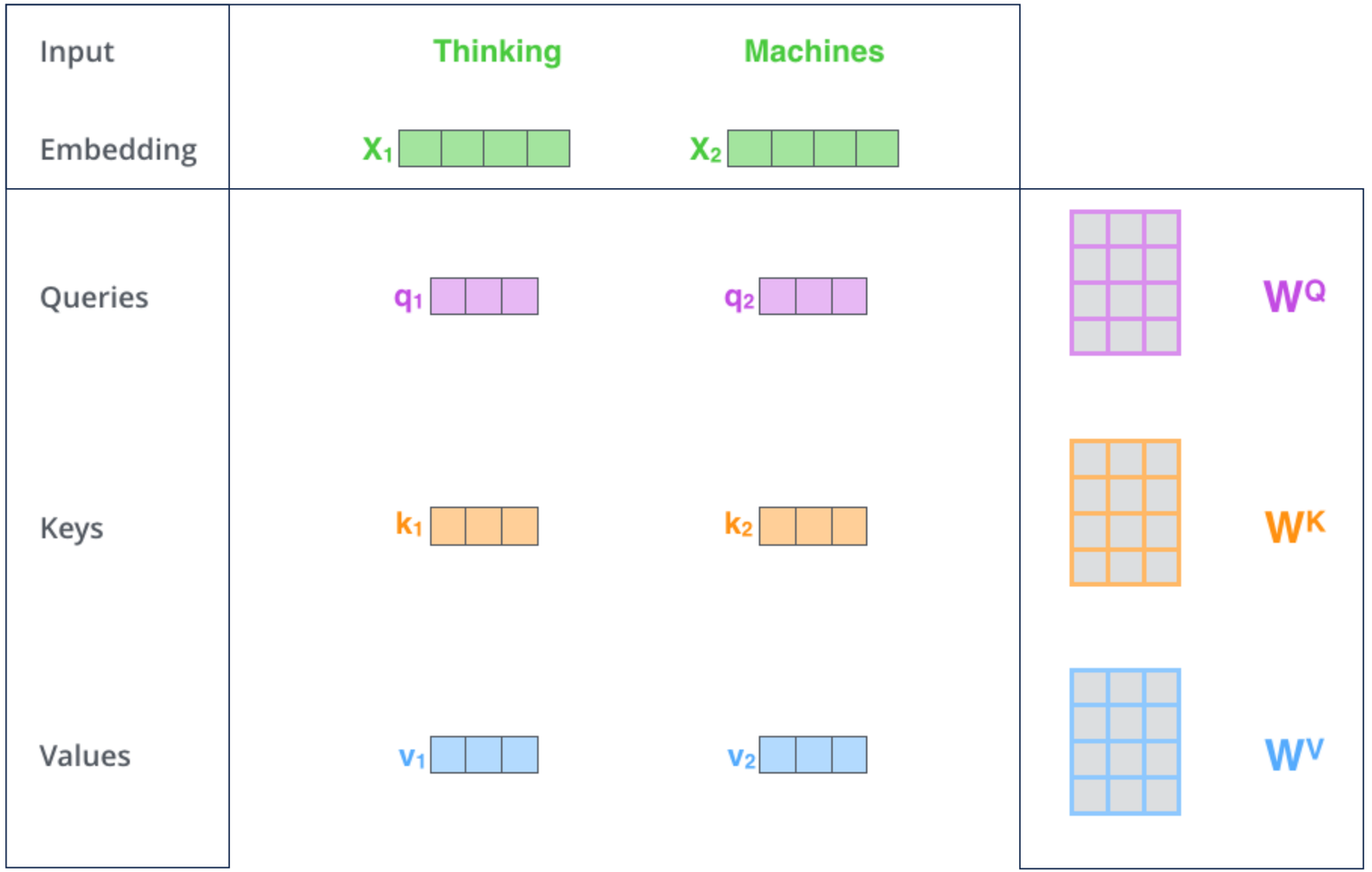


Bron: https://pytorch.org/blog/interactive-chat-gen-model/

Attention in detail (1/5)

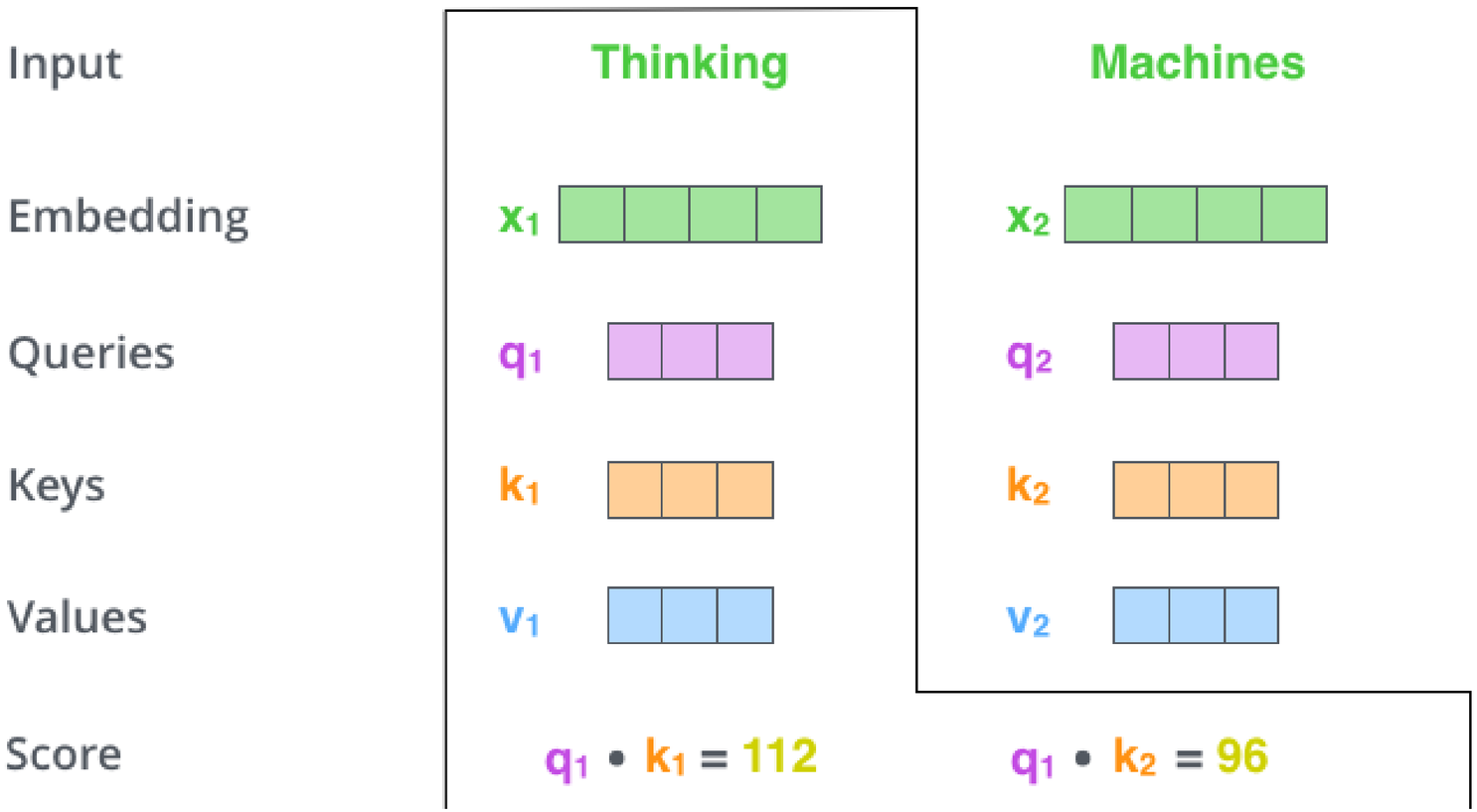
- Grafische uitwerking in de komende slides
- Weergegeven met *vectoren*; in werkelijkheid geparallelliseerd door gebruik van *matrices*
- Berekening van attention wordt gedaan met 3, tijdens het trainen geleerde, matrices:
  - Queries (**Q**): *waarop wil het huidige token letten*
  - Keys (**K**): *welk token is wanneer relevant*
  - Values (**V**): *welke informatie heeft elk token te bieden*
  - Dus Q en K bepalen samen op welke andere tokens we nu willen letten, met V kunnen we daarna de informatie van die tokens verkrijgen
- Resultaat: een verrijkte embedding van het huidige token, met relevante andere tokens (context) erin meegewogen

Attention in detail (2/5)



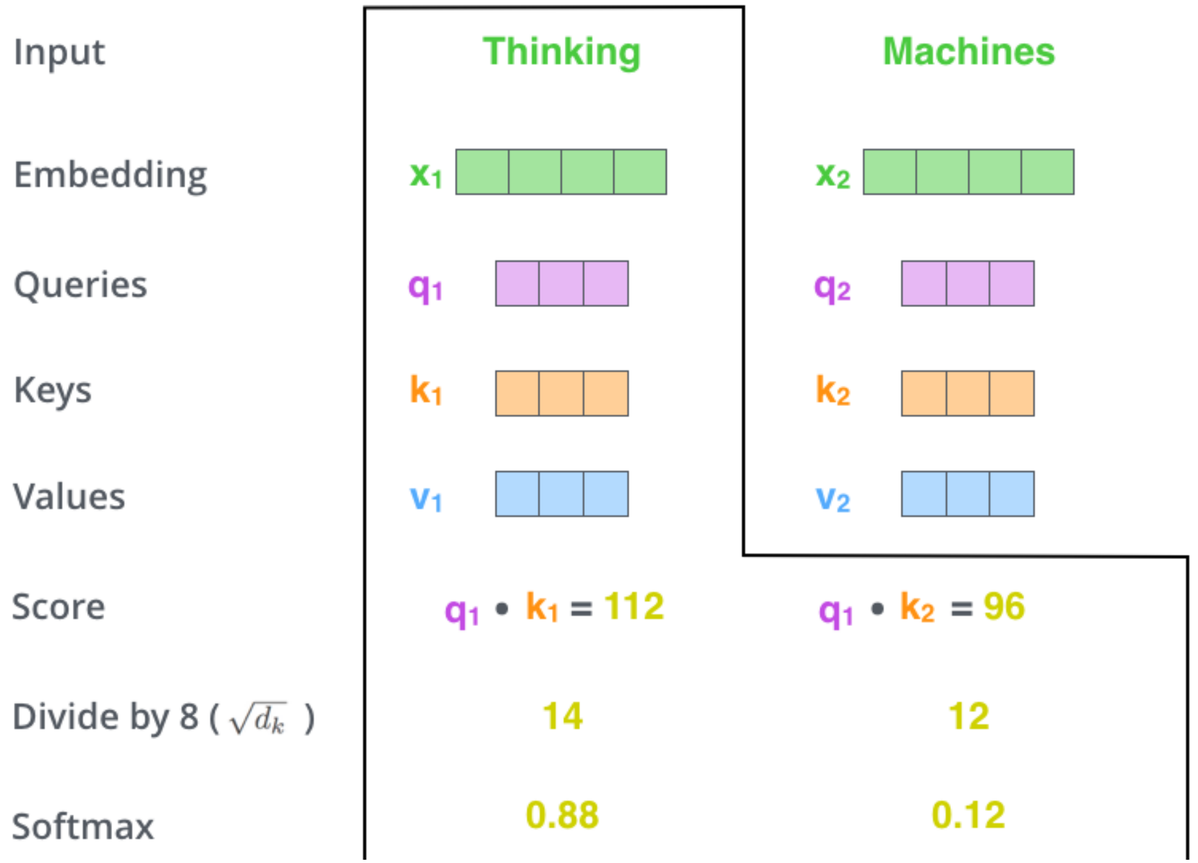
geleerd tijdens training

Attention in detail (3/5)

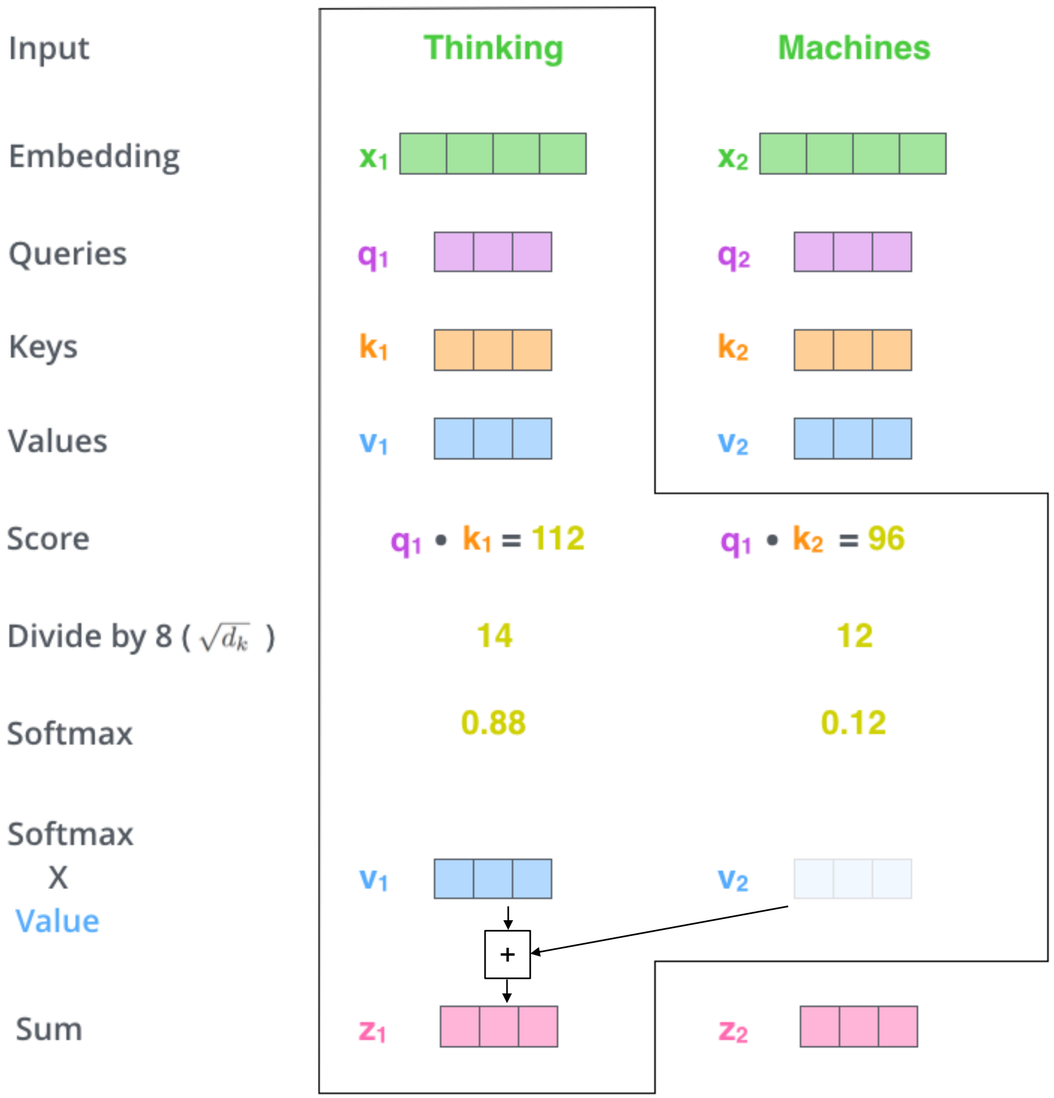


Attention in detail (4/5)

Wortel van de lengte van de key-vectors; geeft stabielere *gradients*



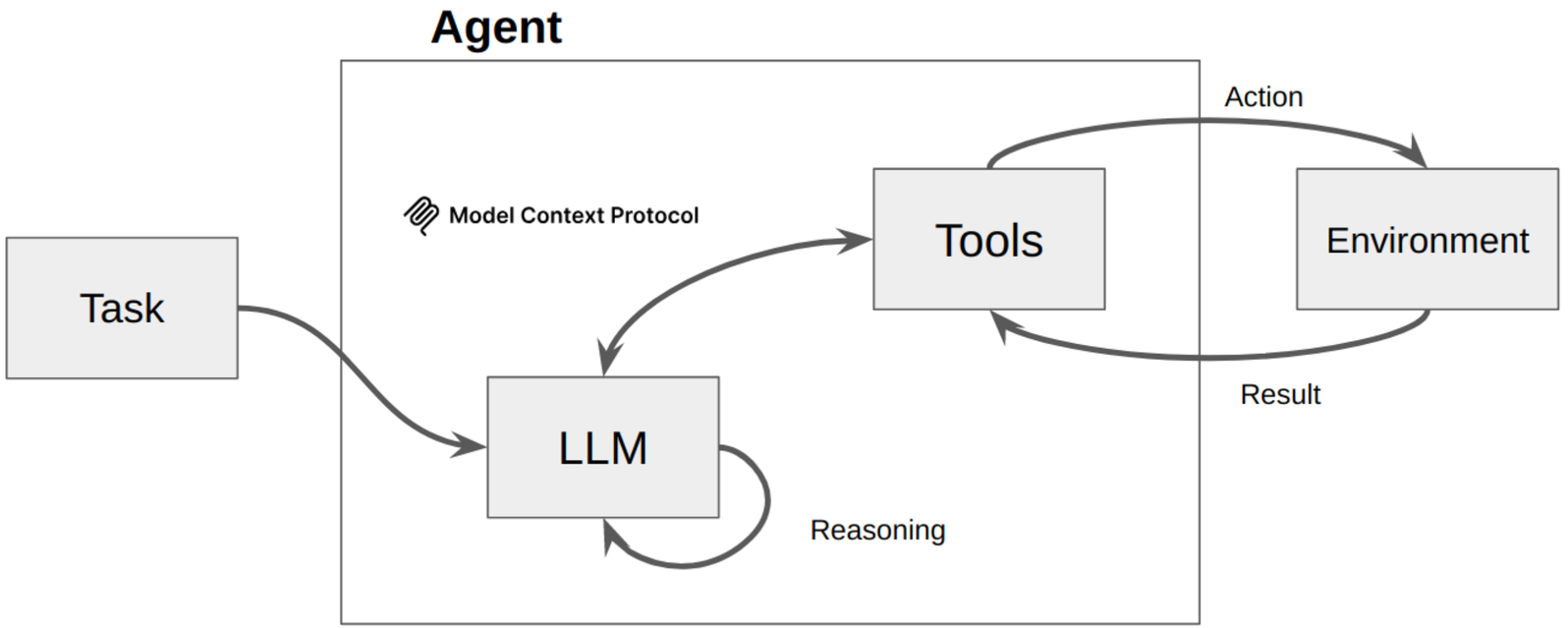
Attention in detail (5/5)



geavanceerde onderwerpen

Agents en Reasoning

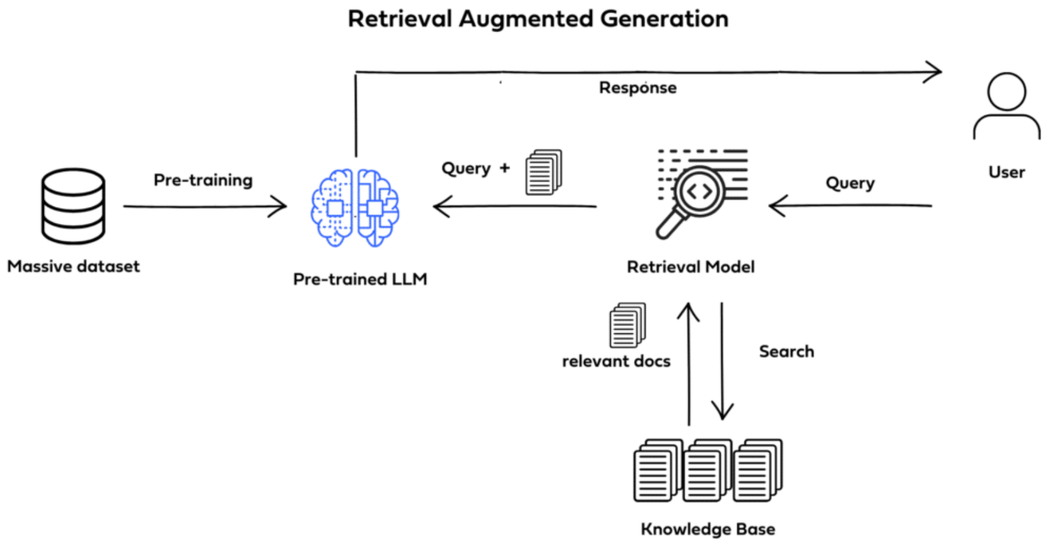
- LLM’s worden uitgebreid met *geheugen* en koppelingen naar *tools*
- Thought - Action - Observation



Bron: https://betterprogramming.pub/implement-generative-agent-with-local-llm-guidance-and-langchain-full-features-fa57655f3de1

Retrieval-Augmented Generation (RAG)

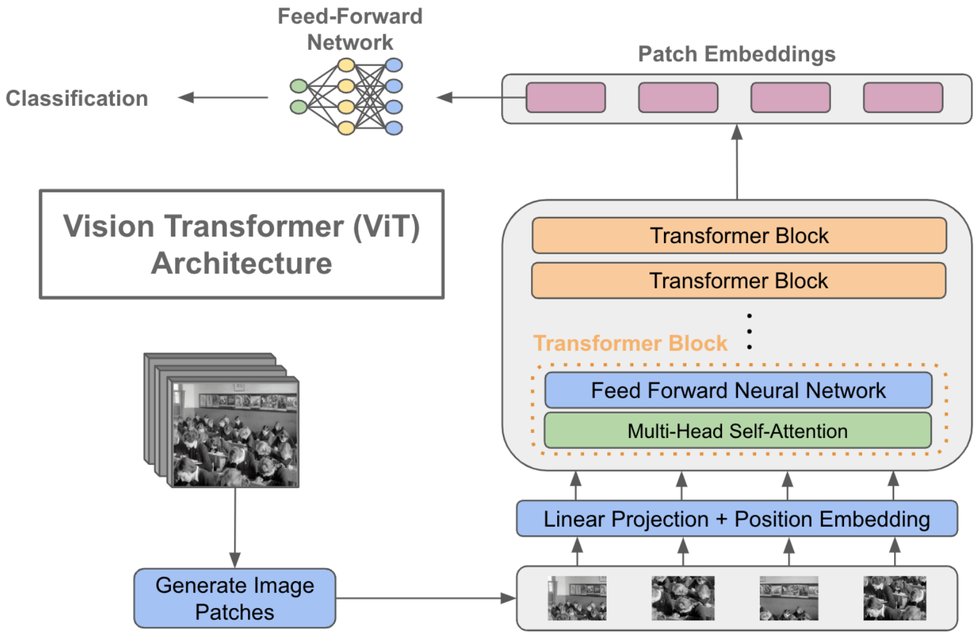
- Semantic search
  - Dense retrieval
  - Reranking
- RAG
  - Zoekresultaten gebruiken als aanvulling op de prompt
  - Om *hallucineren* te verminderen



Bron: https://medium.com/@krtarunsingh/introduction-to-retrieval-augmented-generation-rag-and-its-transformative-role-in-ai-c07e35da7f01

Multimodale LLM's

- Modaliteiten: tekst, afbeeldingen, audio, video, code, sensorwaarden…
- Voorbeeld: Vision Transformer (ViT)
  - **Encoder** die plaatjes omzet naar embeddings
- Tekst én afbeeldingen in dezelfde embedding-space: CLIP
  - Contrastive Language-Image Pre-training
  - Mogelijkheden:
    - Classificeren
    - Clusteren
    - Zoeken
    - Genereren (stable diffusion)



Bron: https://medium.com/artificial-corner/vision-transformer-embrace-convolutional-neural-networks-tec-net-791366f95c2c

live demo’s

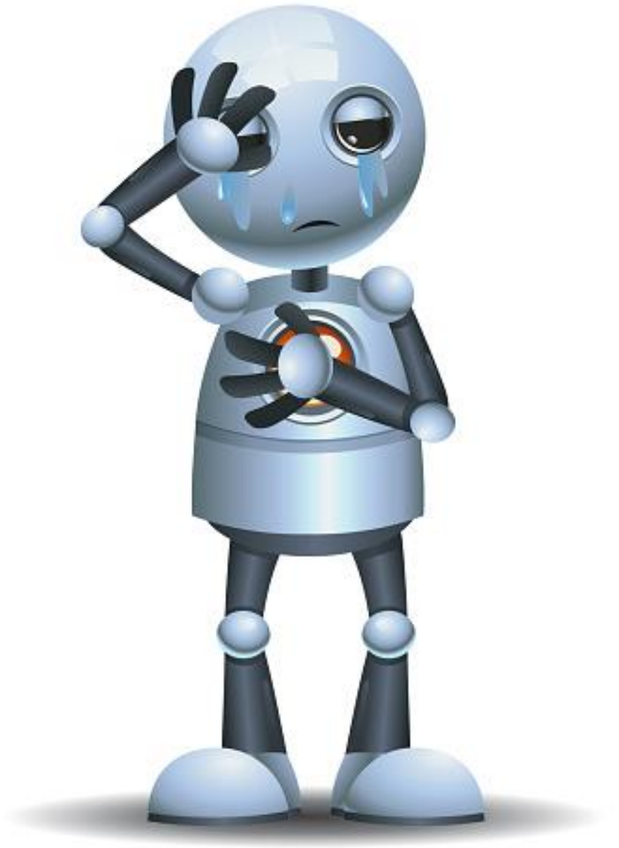
1. Tokens en embeddings
2. Een LLM in actie: Microsoft Phi (3,8 miljard parameters)
3. Decoder en LM Head
4. Classificatie met BERT

Zelf mee aan de slag?

- Keras en Natural Language Processing
  - https://keras.io/api/keras_nlp/
- Hugging Face
  - https://huggingface.co/models

Einde van de hoorcolleges Machine Learning…

- Vragen over de stof?
- Over de laatste opdrachtset?




That's all Folks!