

Large Language Models

**Hoe ontstaat
een antwoord?**

Wat is een LLM?

L

Large

Groot netwerk,
getraind op veel tekst

L

Language

Patronen in taal
en informatie

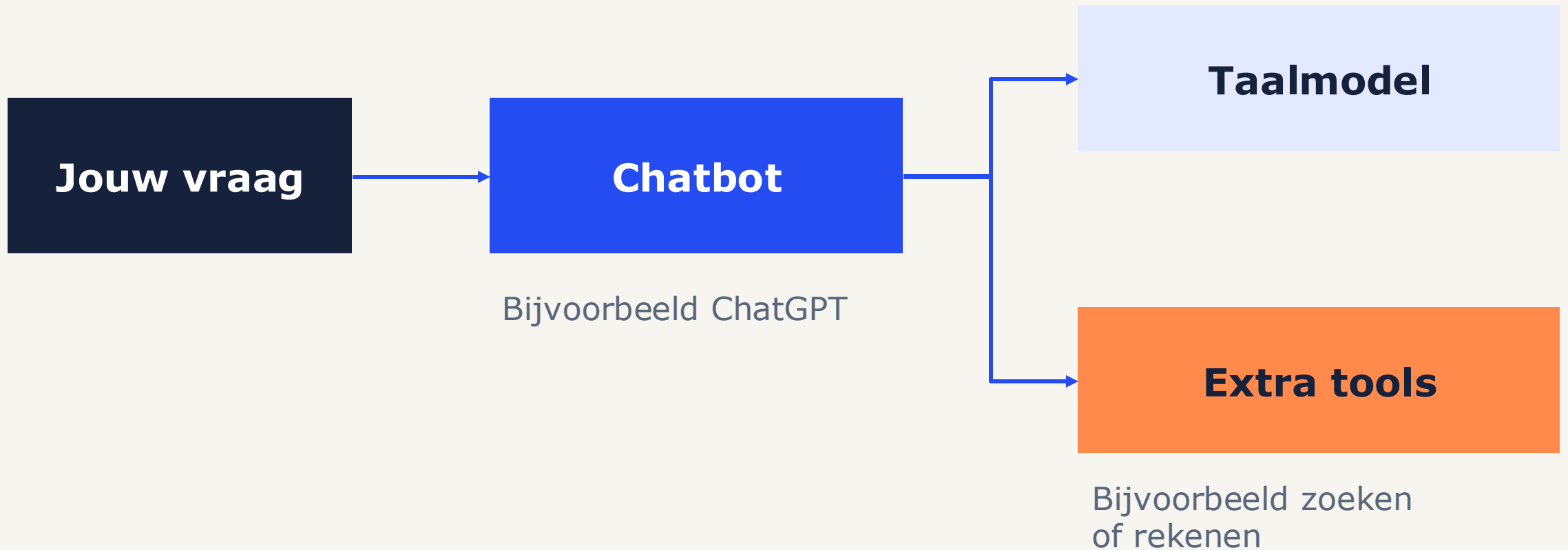
M

Model

Een wiskundig
systeem

Een LLM leert taalpatronen uit tekst en gebruikt die om tekst te genereren.

Het model en de chatbot



Welk woord komt hierna?

**Ik smeer mijn
boterham met ...**

Drie suggesties, daarvan gaan we er op 1 stemmen

Wat verandert door extra context?

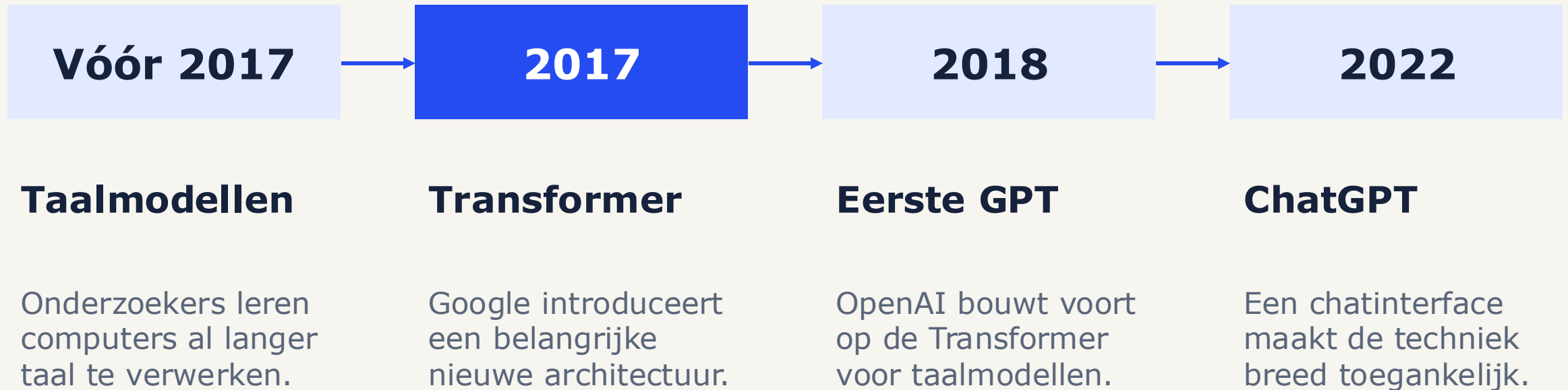
“Ik heb zin in iets zoets.”

**Ik smeer mijn
boterham met ...**

Verandert onze voorspelling? Waarom?

Ook bij een LLM beïnvloedt de context welk tekststukje volgt.

Een korte geschiedenis



Google, OpenAI en andere onderzoeksgroepen bouwen verder op elkaars werk.

Een antwoord begint met tokens

Ik ga morgen naar school.

Ik

ga

morgen

naar

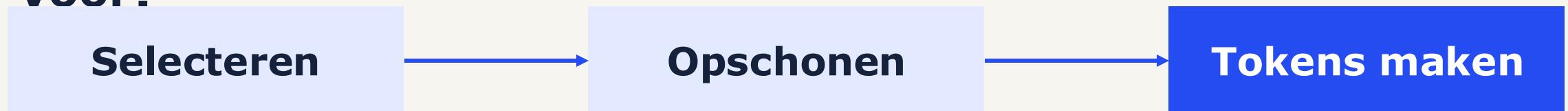
school

.

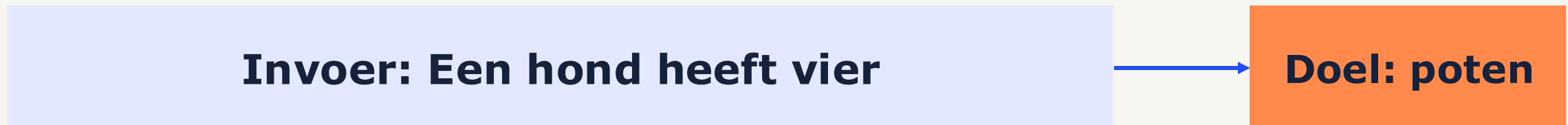
Tekststukjes krijgen een ID en een rij getallen.
Daarmee kan het model rekenen.

1. Pretraining

Bv. websites, boeken en code. Hebben ze daar toestemming voor?



Trainingstekst: "Een hond heeft vier poten."



Dit heet zelfgesuperviseerd leren.

2. Een foutscore stuurt het leren

Werkelijk vervolg: poten

VOORSPELDE KANSEN

poten	20%
vleugels	50%
overig	30%

Loss = foutscore

Hoe lager de kans op "poten",
hoe groter de foutscore.

Backpropagation

Berekent hoe veranderingen
in getallen de fout beïnvloeden.

Optimizer

Past die getallen een beetje aan.
Daarna opnieuw voorspellen.

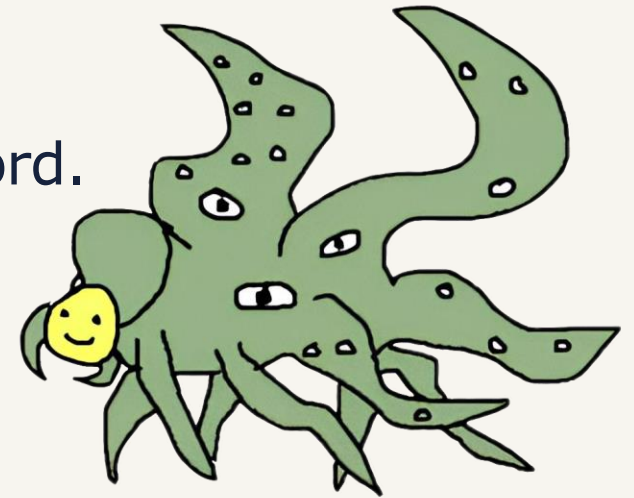
3. Extra training voor een chatbot

Instructietraining

Het model krijgt veel voorbeelden van een goed antwoord.

Training met feedback

Beoordelaars vergelijken antwoorden. Hun voorkeuren helpen het model bruikbaarere antwoorden te leren geven. Dit heet reinforcement training.

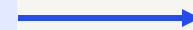


Het antwoord groeit token voor token

Elke stap: kansen berekenen, een token kiezen en toevoegen.

STAP 1

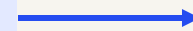
De hoofdstad van Frankrijk is



Parijs

STAP 2

De hoofdstad van Frankrijk is Parijs



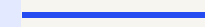
.

Het nieuwe token telt direct mee bij de volgende voorspelling.

Attention weegt de context mee

Attention helpt het model relevante woorden mee te wegen.

“Ik stort geld op de bank.”



Financieel

“Ik zit op de bank.”



Meubel

“Geld” en “zit” geven hetzelfde woord een andere context.

Betrouwbaarheid

Een LLM geeft niet altijd betrouwbare of correcte informatie.

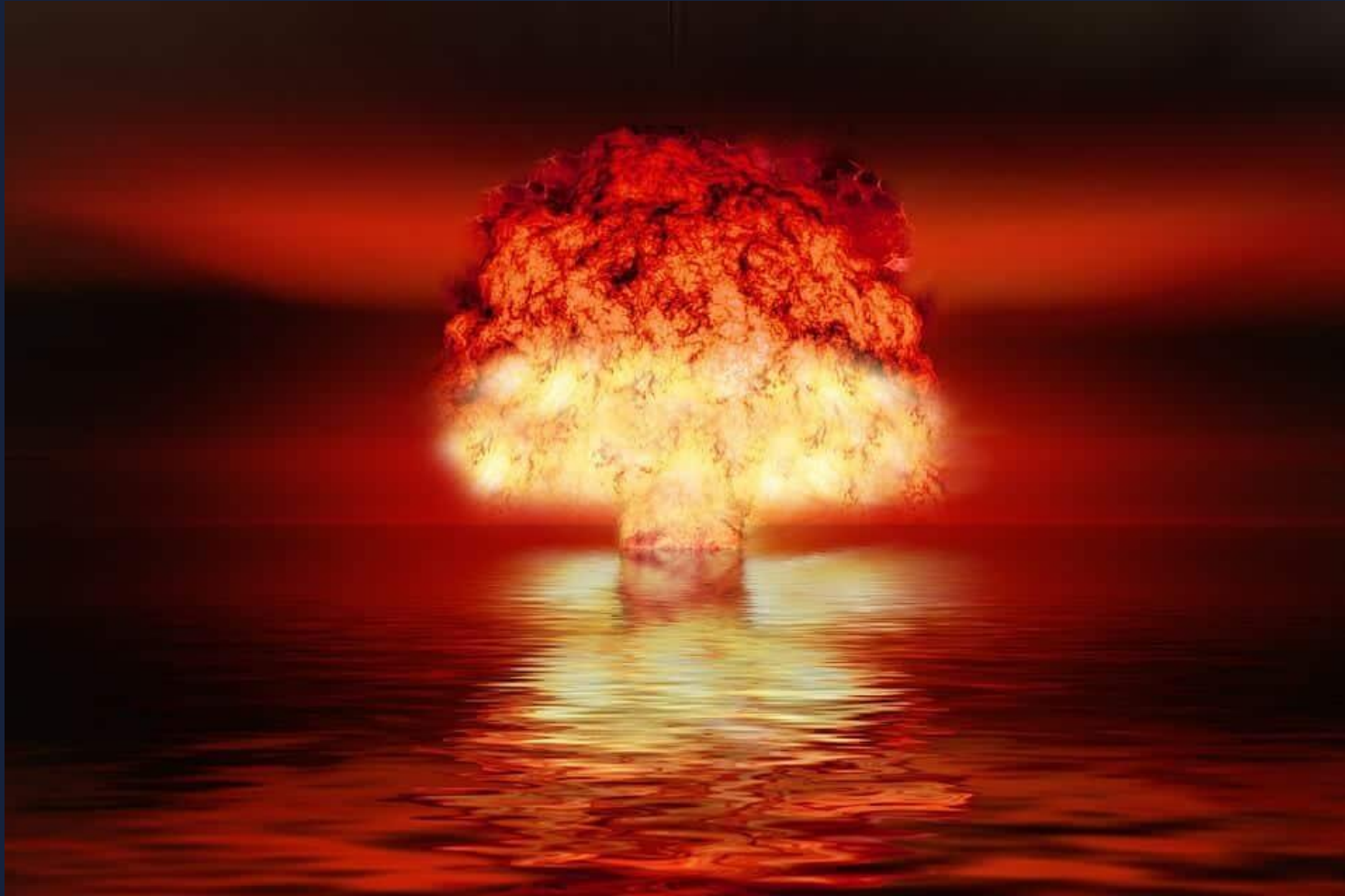
Een passende formulering kan ook een verzonnen bron bevatten.

Tekst voorspellen controleert op zichzelf niet of iets waar is.
Een onjuist antwoord kan daardoor heel overtuigend klinken.

Ook risico's: vooroordelen, privacyproblemen en misbruik.

**Hoe groot kunnen AI-
risico's worden?**

Vrij groot



Hoe groot kunnen AI-risico's worden?

Misbruik

Krachtige systemen kunnen helpen bij grootschalige schade. Zoals cybersecurity.

Sommige onderzoekers waarschuwen zelfs voor uitstervingsrisico van de mensheid.

Controleverlies

Blijven toekomstige LLM's aligned met de mensheid en zullen die naar ons luisteren.

Hoeveel controle willen we houden?

**Eerst veiliger?
Of sneller verder?**

Wie mag bepalen welk risico aanvaardbaar is?

Bronnen en betrouwbaarheid

Transformer / Google en Vaswani et al. (2017)

arxiv.org/abs/1706.03762

Eerste GPT / OpenAI (2018)

openai.com/index/language-unsupervised/

Introductie ChatGPT / OpenAI (2022)

openai.com/index/chatgpt/

RAG / Lewis et al. (2020)

arxiv.org/abs/2005.11401

Risico's / NIST (2024)

doi.org/10.6028/NIST.AI.600-1

Bronnen: training en grotere risico's

Basistraining / Brown et al. (2020)

arxiv.org/abs/2005.14165

Menselijke feedback / Ouyang et al. (2022)

arxiv.org/abs/2203.02155

Backpropagation / Google Machine Learning Crash Course

developers.google.com/machine-learning/crash-course/neural-networks/backpropagation

Risico's en onzekerheid / International AI Safety Report (2025)

internationalaisafetyreport.org/publication/international-ai-safety-report-2025

Waarschuwing van experts / Center for AI Safety

safe.ai/work/statement-on-ai-extinction-risk