Under panseret til ChatGPT

En ikke-teknisk innføring i teknologien og terminologi

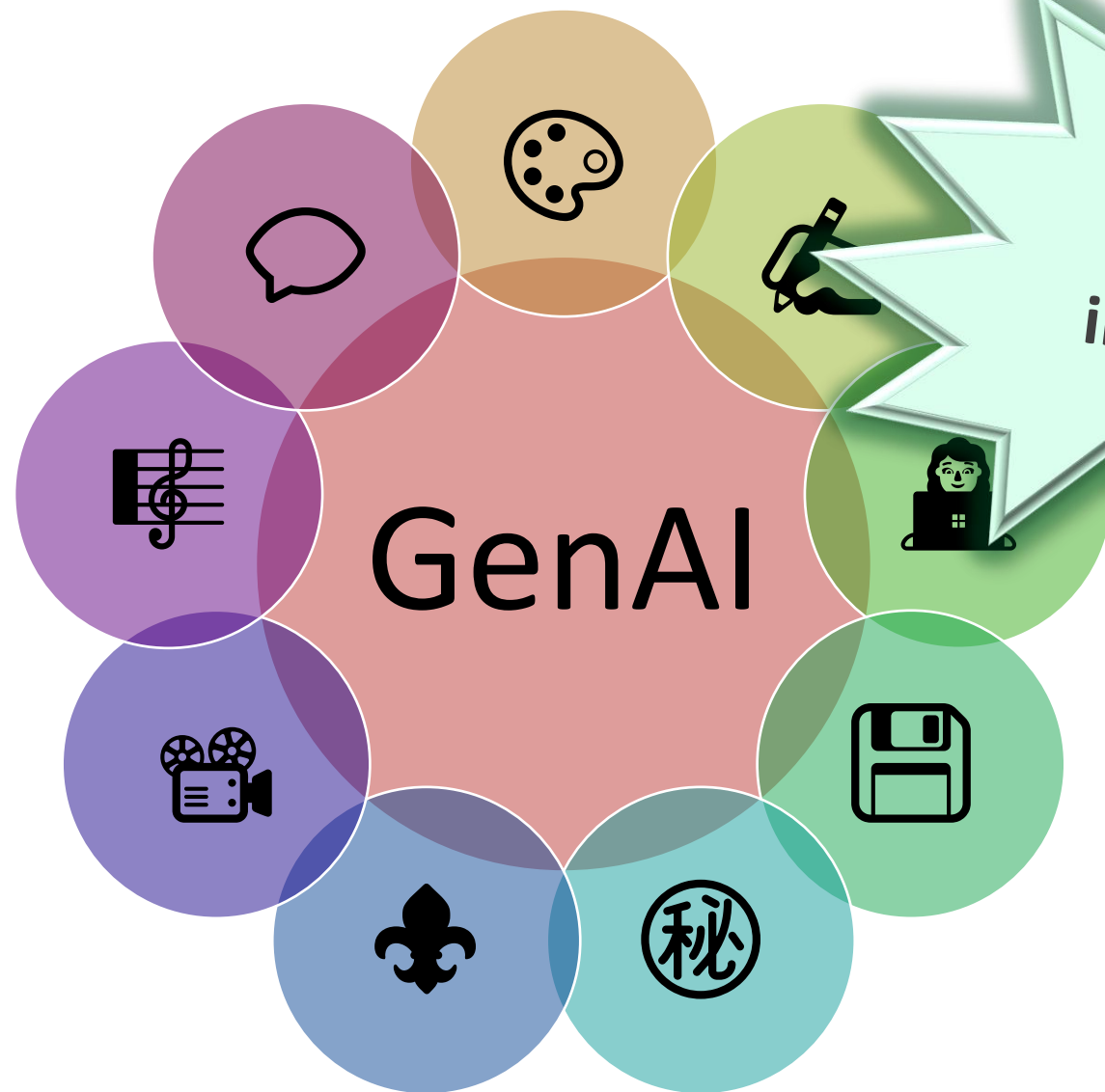**Buzzwords!**

SINTEF

Teknologi for et bedre samfunn

GenAI

ChatGPT

Modell

$y=f(x)$

x

Maskin-læring

SINTEF

Teknologi for et bedre samfunn
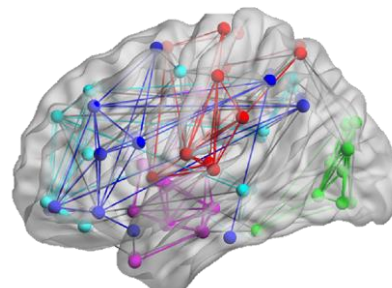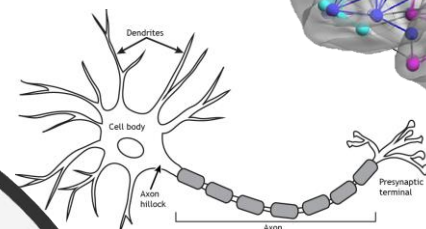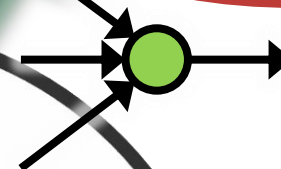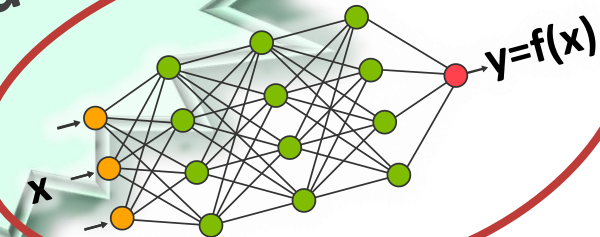
Trening

Uten trening

"You must decide for yourself,"
said Elizabeth
L;_FfsqeW9*M)4OCg4AAm-
x_4t:h
WXkm3i5qEt*f0btTf!1,VKle[T?V5
238
qw_34g"?4iQ(C79UoK"dq3SEHC"
L.(tEx9CY0wutdgjj!hZl2.Q[G
A'D6gsm
aP6!!s*Kq4(lA";ZNA9]OS4y":U,RR
bss;MNkhONT7Owpp_!Gl!yprd8b
: REwOPPUn4c-
Scaa_kKQXDD3bS)S?9pN(MXN
?G9?!;Y2Dz[[aZa:

250 runder med trening

"You must decide for yourself,"
said Elizabeth gra buseriteand
the utharth s atouchanders nd
shadeto the se owrerer o thino
athe athetlf w wad asire tutor t
m wo sinast othe wo t Mr m ve
the Shersour us mass ho'sed t tha
shan all athok I rs ofit yo t
onchthe axpld ndise fas e trnglad
fere sed, orou ke as thelf se wn
the tnn cenneng. f hachaplll I fmy
n p oth ther o mbe butane.
hastamt ok, f warot Bere a unge
sen t t they op "Yo she aise tho
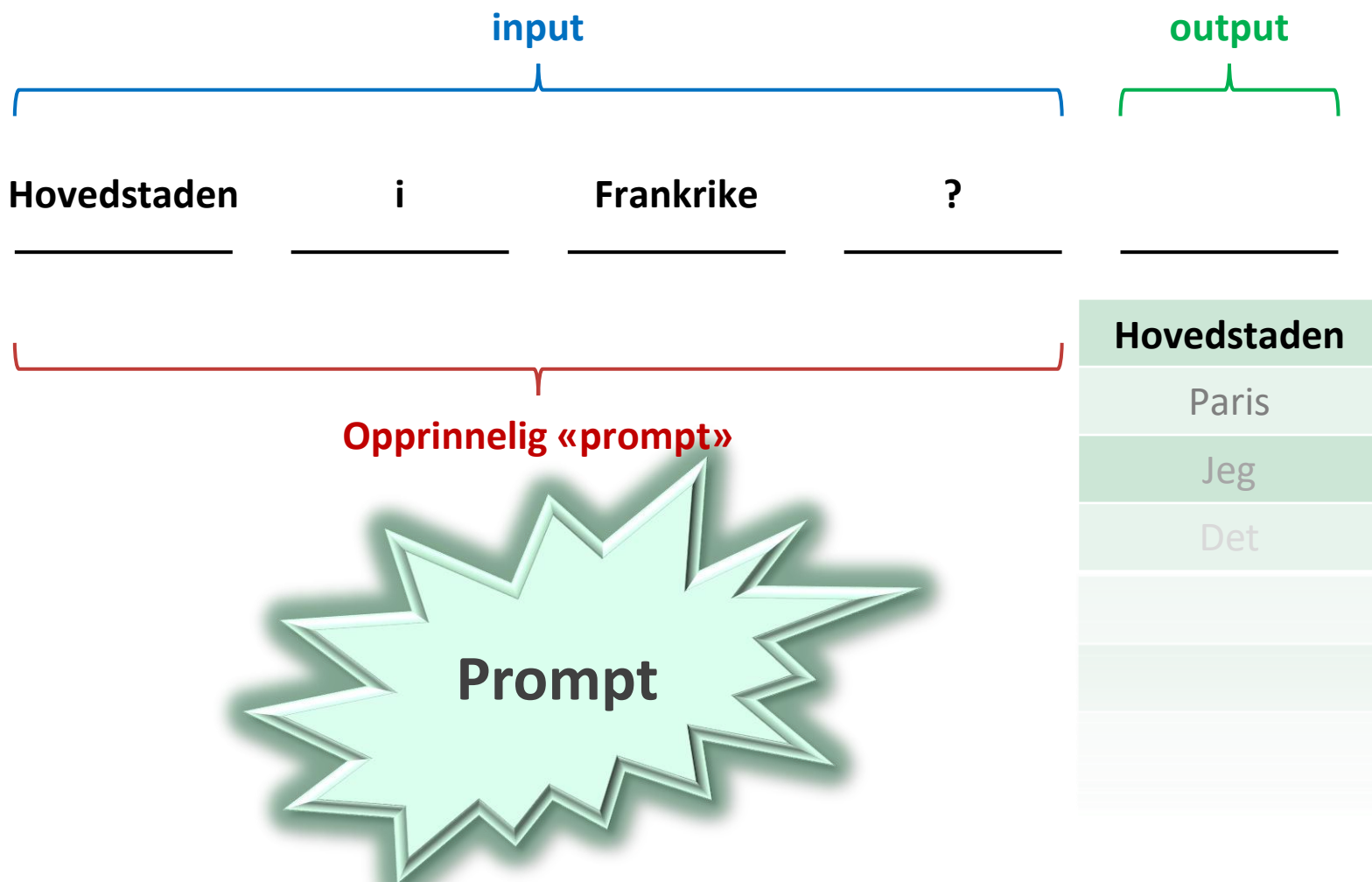hiven tise He sus y athire wald
wal be, o akrer etlas o

5.000 runder med trening

"You must decide for yourself,"
said Elizabeth, rather repeated
an unhappy confirmed, "as now
it is a few eyes," were there at
Anne's like walking to meet that
she always were quite married
that some uncumstant's decent
the imagination of side, for their
two secondness which going at
Here fault would be to the same
between the necessary. "He was
never known that she began and
interest of situation."

30.000 runder med trening

"You must decide for yourself,"
said Elizabeth, rather repeatedly;
"that is very agreeable
displeasure, they will ever be a
lively young woman as it will be
more disagreeable." "My dear
Fanny, who is a match of your
present satisfaction, and I am at
liberty and dinner, for everybody
can be happy to you again; and
now when I think I used to be
capable of other people, by being
hastily used to be forgotten in
something of the little first
possibility of my usual taste
which such a party as this word.
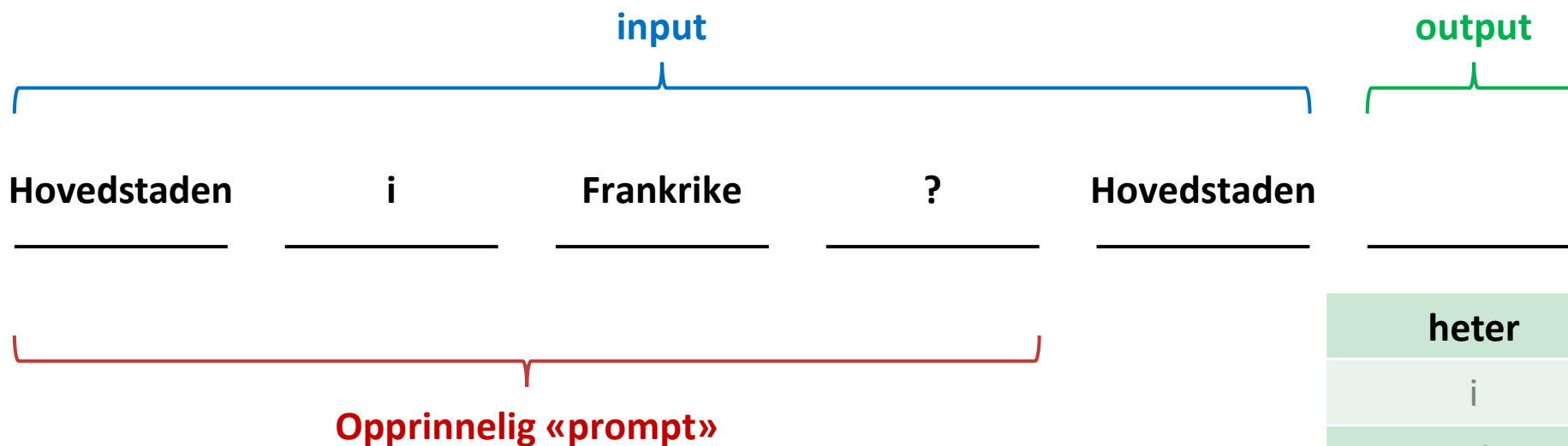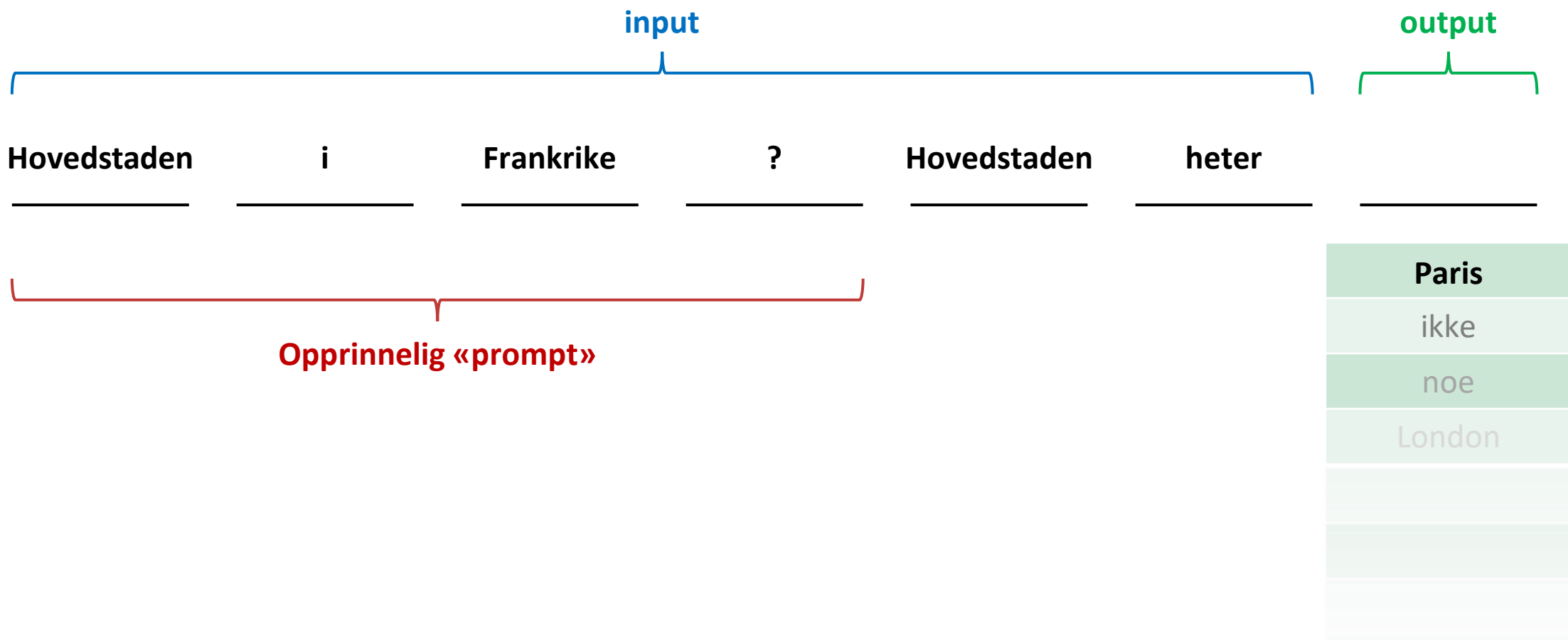
NY Times | GPT from scratch

Teknologi for et bedre samfunn

Hvordan svarer ChatGPT?

input

output

Hovedstaden	i	Frankrike	?	

Opprinnelig «prompt»

Prompt

Hovedstaden
Paris
Jeg
Det

Hvordan svarer ChatGPT?

input

output

Hovedstaden   **i**   **Frankrike**   **?**   **Hovedstaden**

Opprinnelig «prompt»

heter
i
til
er

Teknologi for et bedre samfunn

Hvordan svarer ChatGPT?

input

output

Hovedstaden   i   Frankrike   ?   Hovedstaden   heter

Opprinnelig «prompt»

Paris
ikke
noe
London

Hvordan svarer ChatGPT?

Hovedstaden	i	Frankrike	?	Hovedstaden	heter	Paris

Opprinnelig «prompt»

Svar
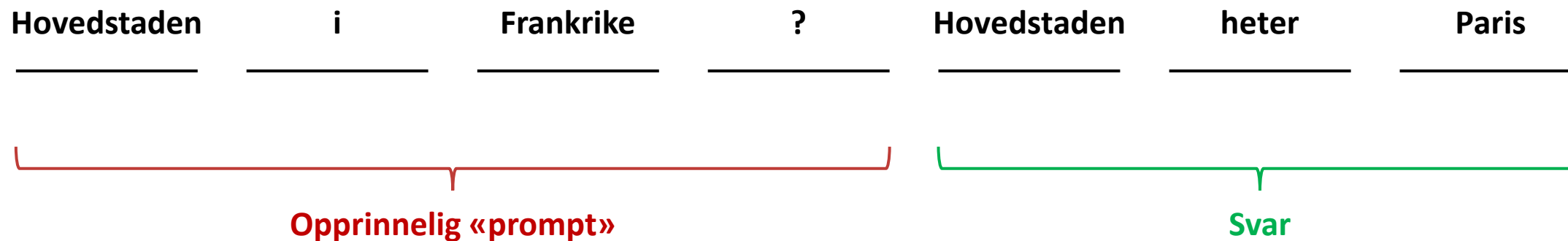
Hvordan lage ChatGPT?

Deployable model

Non-deployable model

Commercial product

Train a large language model

Supervised Fine Tuning

Reward Modeling

Reinforcement Learning

Content Moderation Filter

ChatGPT

Train a large language model
- ☀ : Raw internet texts
- ➡ : A Foundation Model (FM)
- 📝 : Requires extremly large dataset and resources (low quality, high quantity)
- 🎯 : Adequate text completion
- ⏱ : Takes 99% of the time/resources

Supervised Fine Tuning
- ☀ : Ideal Q/A examples
- ➡ : An Assistant, a Supervised Fine Tuned (SFT) model
- 📝 : 10-100K pairs of [prompt, responses], written by humans (high quality, low quantity)
- 🎯 : Prompt response proficiency
- ⚙ : Refines the FM

Reward Modeling
- ☀ : Ranked completions
- ➡ : A Reward Model (RM)
- 📝 : 100K-1M comparisons ranked by humans (high quality, low quantity)
- 🎯 : Model that can assess prompt reply quality

Reinforcement Learning
- ☀ : Prompts
- ➡ : An Assistant, a Re-inforcement Learning (RL) model
- 📝 : 10-100K prompts written by humans (high quality, low quantity)
- 🎯 : Prompt response mastery
- ⚙ : Refines the SFT model using RL and the RM.

Content Moderation Filter
- ➡ : Content filtering on user input and model output
- 📝 : OpenAI provides its own Moderation API
- 🎯 : Content Quality and Compliance Enforcement

Teknologi for et bedre samfunn

Til slutt...

- «Stokastisk pappegøye»: ChatGPT er en språkmodell, ikke en kunnskapsmodell
  - Veien blir til mens den går
  - Hallusinering
  - Output reflekterer treningssettet sitt

- OpenAI: ikke så åpne som de høres ut
  - Mye ukjent
  - Samtalene brukes sannsynligvis til videre trening

- Men... ikke kast barnet ut med badevannet

Teknologi for et bedre samfunn