Large Language Models and their use in research

Pierre Lison
Norsk Regnesentral (NR) & UiO
plison@nr.no

TIK9015 – STS Methodologies: Practice-Oriented Document Analysis
November 5, 2024, Oslo

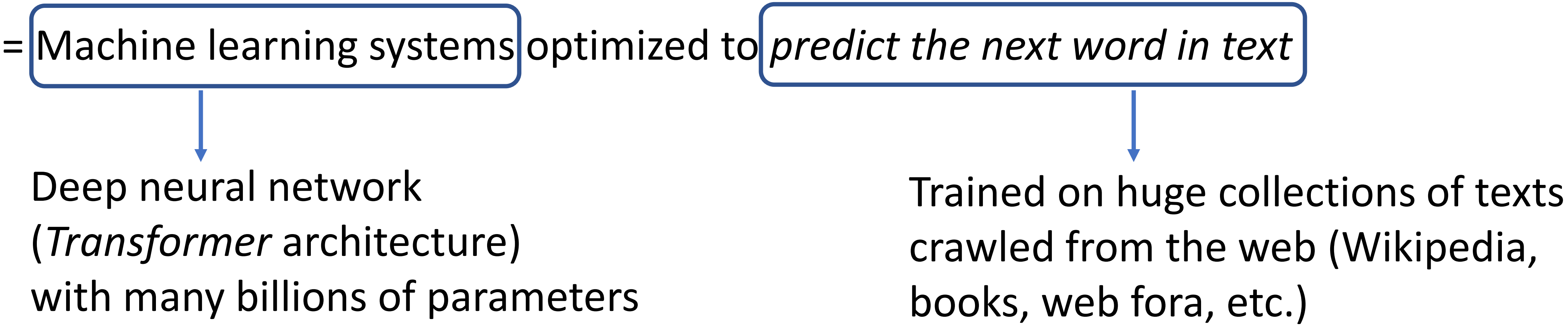
Outline

1. What are Large Language Models?
2. Using LLMs in practice
3. Open questions

Outline

1. **What are Large Language Models?**
2. Using LLMs in practice
3. Open questions

Large Language Models?



= Machine learning systems optimized to *predict the next word in text*

Deep neural network (*Transformer* architecture) with many billions of parameters

Trained on huge collections of texts crawled from the web (Wikipedia, books, web fora, etc.)

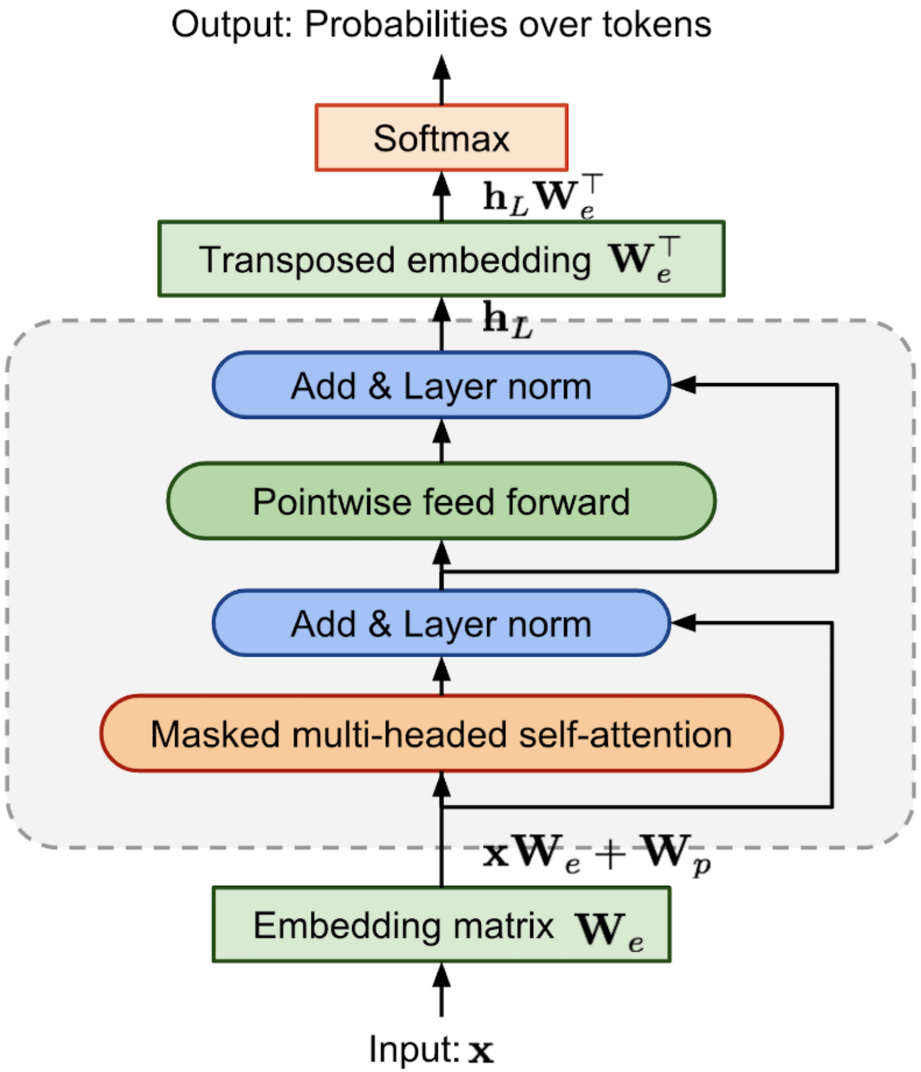
As a by-product of this “guessing game”, the neural model will gradually build up **representations** of both *linguistic* and *world* knowledge

This model can then be *adapted* for various tasks

Transformers

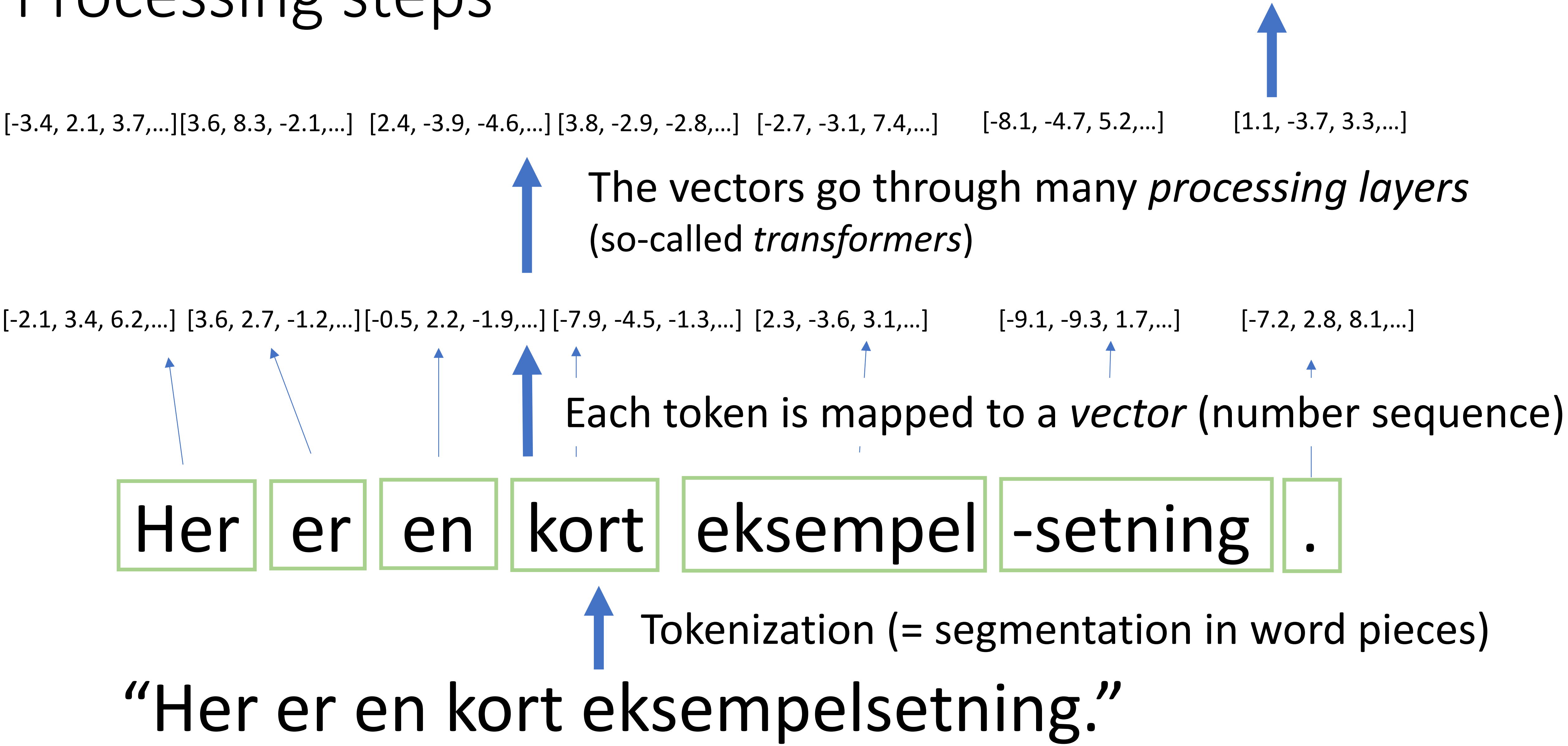
- Today's LLMs are all variants of the same architecture: the **transformer**
- Many processing layers on top of each other, each with their own parameters
- *Key idea*: let the vector of each token be influenced by its **context** (=the tokens that precede/surround it)

How tokens "influence" each other's vectors is something that is learned from data


Output: Probabilities over tokens
Softmax
$\mathbf{h}_L\mathbf{W}_e^\top$
Transposed embedding $\mathbf{W}_e^\top$
$\mathbf{h}_L$
Add & Layer norm
Pointwise feed forward
Add & Layer norm
Masked multi-headed self-attention
$\mathbf{x}\mathbf{W}_e + \mathbf{W}_p$
Embedding matrix $\mathbf{W}_e$
Input: $\mathbf{x}$

Processing steps

And the result is used to predict the next word



[-3.4, 2.1, 3.7,…][3.6, 8.3, -2.1,…] [2.4, -3.9, -4.6,…] [3.8, -2.9, -2.8,…] [-2.7, -3.1, 7.4,…] [-8.1, -4.7, 5.2,…] [1.1, -3.7, 3.3,…]

The vectors go through many *processing layers*
(so-called *transformers*)

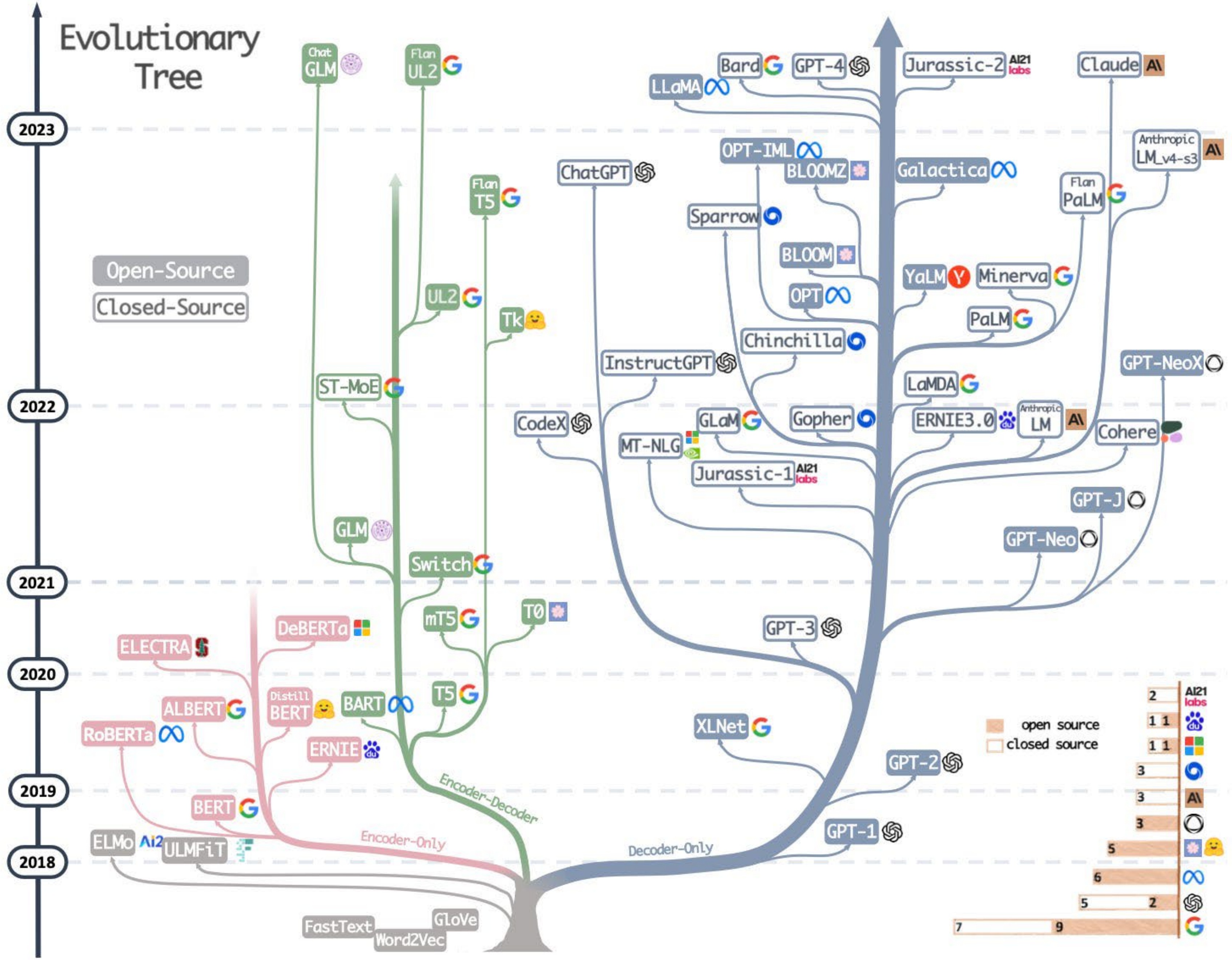
[-2.1, 3.4, 6.2,…] [3.6, 2.7, -1.2,…][-0.5, 2.2, -1.9,…] [-7.9, -4.5, -1.3,…] [2.3, -3.6, 3.1,…] [-9.1, -9.3, 1.7,…] [-7.2, 2.8, 8.1,…]

Each token is mapped to a *vector* (number sequence)

Her | er | en | kort | eksempel | -setning | .

Tokenization (= segmentation in word pieces)

“Her er en kort eksempelsetning.”

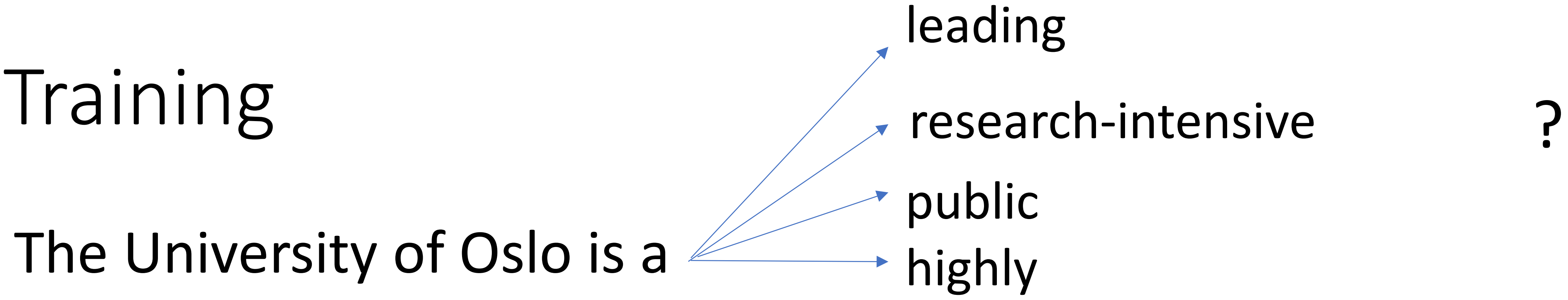

Evolutionary Tree
2023
2022
2021
2020
2019
2018
Open-Source
Closed-Source
Chat GLM
Flan UL2
Flan T5
UL2
Tk
ST-MoE
GLM
Switch
mT5
T0
T5
BART
DeBERTa
ELECTRA
ALBERT
Distill BERT
RoBERTa
ERNIE
BERT
ELMo
Ai2
ULMFiT
Encoder-Only
Encoder-Decoder
Decoder-Only
FastText
GloVe
Word2Vec
LLaMA
Bard
GPT-4
Jurassic-2
AI21 labs
Claude
OPT-IML
BLOOMZ
Galactica
Anthropic LM_v4-s3
ChatGPT
Flan PaLM
Sparrow
BLOOM
YaLM
Minerva
OPT
PaLM
Chinchilla
InstructGPT
GPT-NeoX
LaMDA
CodeX
GLaM
Gopher
ERNIE3.0
Anthropic LM
Cohere
MT-NLG
Jurassic-1
AI21 labs
GPT-J
GPT-Neo
GPT-3
XLNet
GPT-2
GPT-1
open source
closed source
2
AI21 labs
1 1
1 1
3
3
3
5
6
5 2
7 9

Multilingual or not?

- Some LLMs are trained for a specific language
  - For Norwegian: NorBERT, NorMistral, etc.
  - Need a sufficiently large training set
    (the Norwegian Collossal Corpus from the National Library)
- Others are trained on generic web content
  - Can work on many languages
  - But high imbalance when it comes to cultural references etc.

“ChatGPT is multilingual but monocultural” (Jill Walker Rettberg)

Training



The University of Oslo is a

leading

research-intensive

public

highly

?

- LLMs are optimized (or *trained*) to predict the next word in a text
- **Training** = gradual change of parameters so that the model predictions get a little closer to the “ground truth“
- Modern LLMs have hundreds of billions of parameters and are trained on special hardware called GPUs

Multimodality

- Generative AI models are increasingly **multi-modal**:
  - Not only text, but also images, audio, video, structured databases, etc.
  - Both input and output side
  - “Foundation models”
- Still “web data” (no connection to situated, embodied experience)

Model alignment

- A “standard” LLM is optimized to find likely continuations of an input string

  Input: Translate to French: The small dog

  Output: crossed the road → A perfectly plausible continuation, but not what we were looking for

- We often want our LLMs to **follow instructions** and generate outputs that satisfy certain **quality criteria**

↓

Be relevant and informative, do not spread disinformation, avoid offensive or abusive language, refrain from producing dangerous/illegal content, etc.

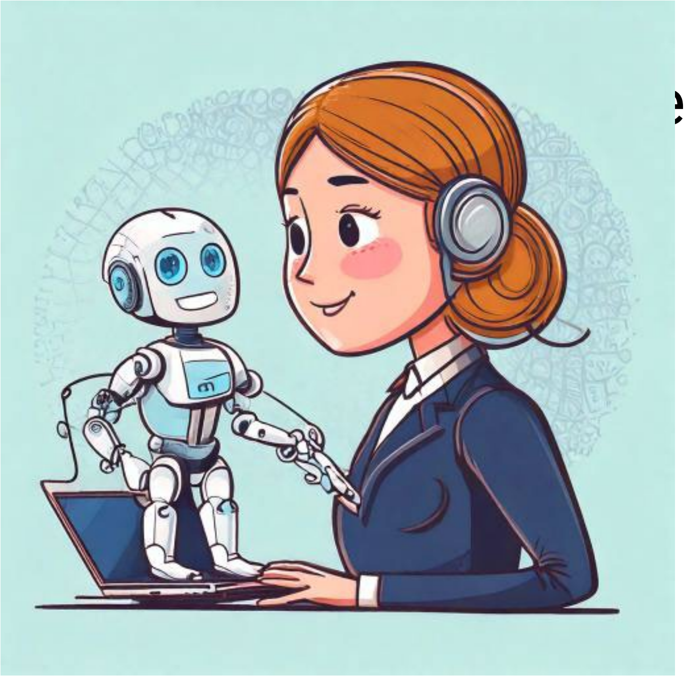
Model alignment

- Getting the model to output “what we want” is often called alignment
  - We seek to “align” the model with certain values & rules of conduct
- This is not a pure technological question!
  - Ethical/political questions about what LLM should “align” towards
- Two alignment methods:

Instruction tuning (from examples of correct responses)

Preference alignment

👍 👎



Outline

1. What are Large Language Models?
2. **Using LLMs in practice**
3. Open questions

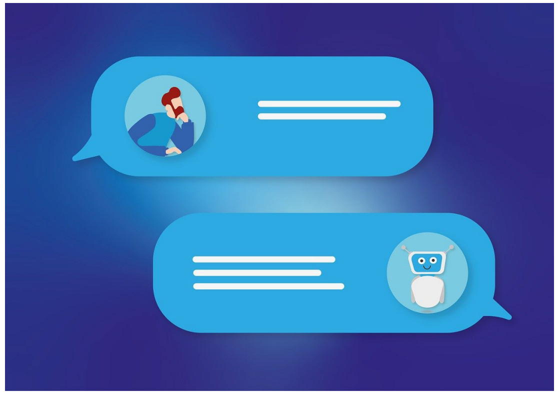
Some applications

- Classify, filter, organize documents
- Information extraction & annotation
- Summarisation, translation
- Transcription of audio transcripts

- Writing assistance (get a first draft or improve the form of a given text)
- Autocomplete++ (particularly for coding)
- Question answering, “sparring partner”

Prompting

- A prompt is simply the **input to the LLM**, which can be a question or instructions on the task to perform
- Many LLMs distinguish between the *system* and *user prompt*:
  - **System prompt**: Generic instructions, defined by the system provider, on how the system should respond (be helpful, avoid offensive language, etc.) and how it should present itself to the user (“You are ChatGPT, a large language model trained by OpenAI…”)
  - **User prompt**: Specific instructions provided by the system



In-context learning

Adding a few **examples / demonstrations** can make it easier for the model to “understand” what output is expected

➔ Called “in-context learning”

⚠ Not really “learning” (the model stays unchanged), but biasing the LLM towards certain types of outputs

Given a text, find all named entities and return them in a list. Do not output anything else.

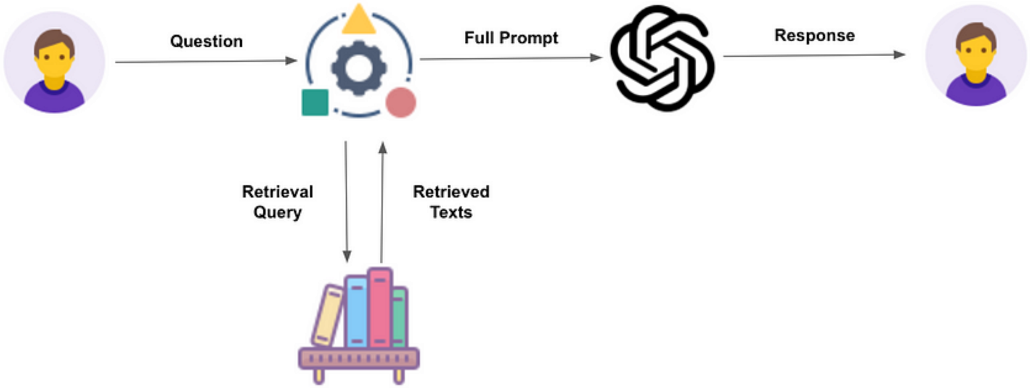
Examples:
Text: “My name is Pierre Lison and I work at the Norwegian Computing Center”
Output: [Pierre Lison, Norwegian Computing Center]
…

Now find the named entities in this text:
“Hilde Reinertsen is an Associate Professor at the University of Oslo.”

Retrieval-augmented generation

- A standard LLM has all of its knowledge stored in its weights
- Many shortcomings:
    - Cannot *add*, *remove* or *update* factual knowledge without retraining the model
    - Lack of *transparency*: difficult to inspect the model knowledge, or understand where it comes from (attribution)
    - How can we connect LLM to *internal databases* or to *specialized knowledge sources* ?
- **Retrieval-augmented generation** (RAG) provides an elegant solution

Retrieval-augmented generation



Basic idea:

1. Given a user prompt, we first *search* for relevant documents in a text database
2. The most relevant texts are then retrieved and added to the prompt to generate the response:

Retrieval stage

Reading stage

«Given the following information: [RETRIEVED TEXTS], answer the following question: [USER QUESTION]»

Chain-of-Thought

- LLMs are still quite limited in their abilities to perform complex, “multi-hop” reasoning
    - Math, science, advanced coding, planning, etc.
- But those “reasoning” capabilities can be improved by forcing the LLM to **generate longer, step-by-step responses**
    - “Let’s think step by step”
    - Can also criticize or debate its own responses
- OpenAI’s new o1 model is build precisely on those tricks

Reasoning abilities

How good are LLMs at tasks requiring reasoning skills?

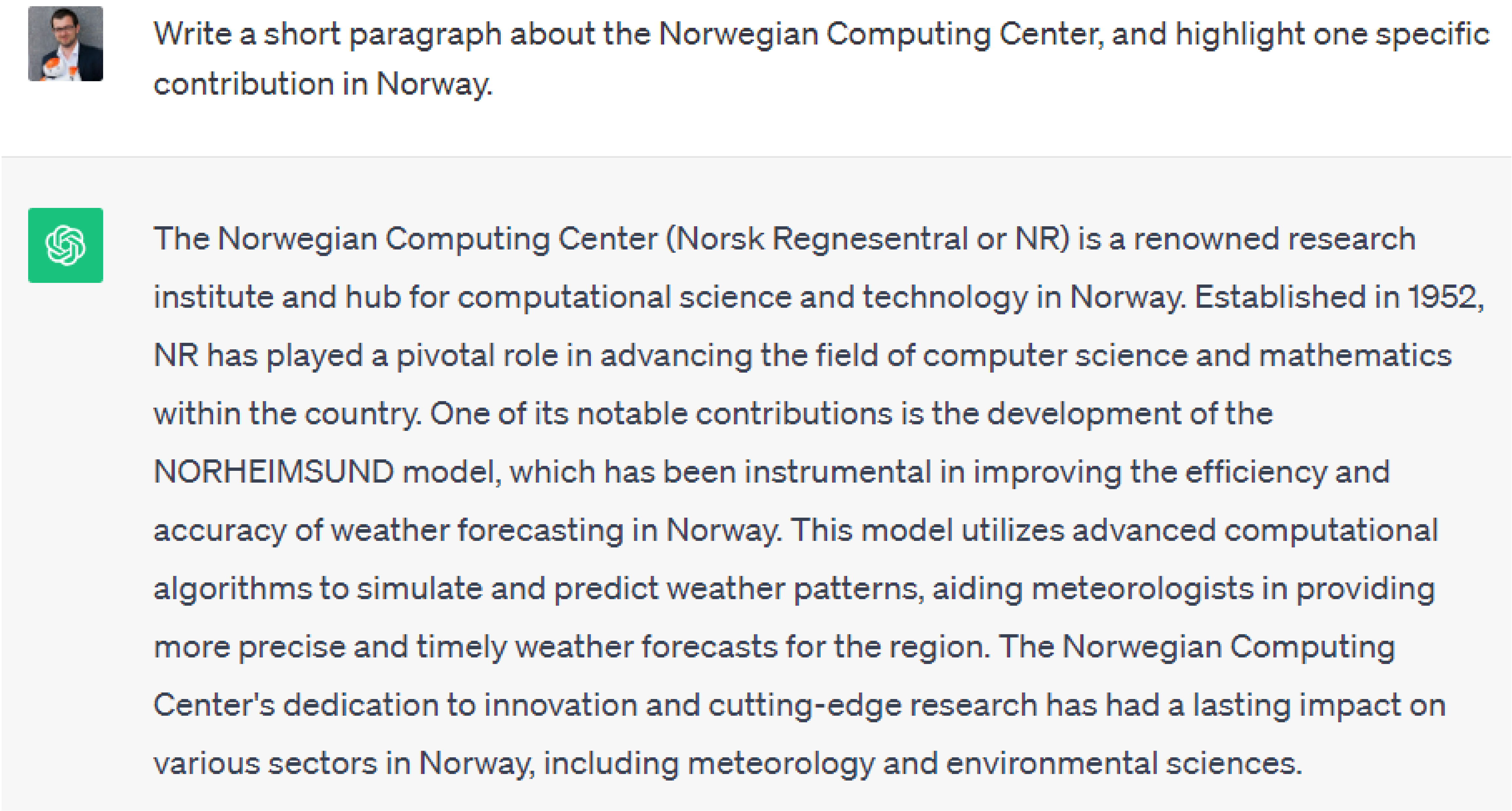
- Active research topic with sometimes conflicting results
- LLMs are experts at **pattern-matching**, also at abstract levels
- But they also rely a lot on **memorization**, which mean they will excel on “familiar” problems, and struggle on novel ones
- Better at *inductive* than *deductive* reasoning
- Catastrophic reasoning failures may happen

Mirzadeh, Iman, et al. "GSM-Symbolic: Understanding the Limitations of Mathematical Reasoning in Large Language Models." *arXiv preprint arXiv:2410.05229* (2024).

Cheng, Kewei, et al. "Inductive or Deductive? Rethinking the Fundamental Reasoning Abilities of LLMs." *arXiv preprint arXiv:2408.00114* (2024).

Outline

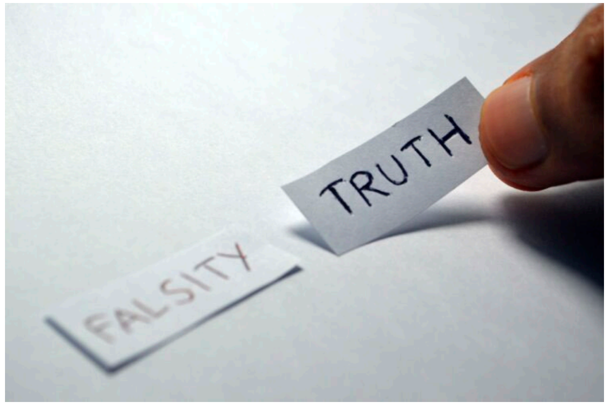
1. What are Large Language Models?
2. Using LLMs in practice
3. **Open questions**



Write a short paragraph about the Norwegian Computing Center, and highlight one specific contribution in Norway.

The Norwegian Computing Center (Norsk Regnesentral or NR) is a renowned research institute and hub for computational science and technology in Norway. Established in 1952, NR has played a pivotal role in advancing the field of computer science and mathematics within the country. One of its notable contributions is the development of the NORHEIMSUND model, which has been instrumental in improving the efficiency and accuracy of weather forecasting in Norway. This model utilizes advanced computational algorithms to simulate and predict weather patterns, aiding meteorologists in providing more precise and timely weather forecasts for the region. The Norwegian Computing Center's dedication to innovation and cutting-edge research has had a lasting impact on various sectors in Norway, including meteorology and environmental sciences.

Hallucinations



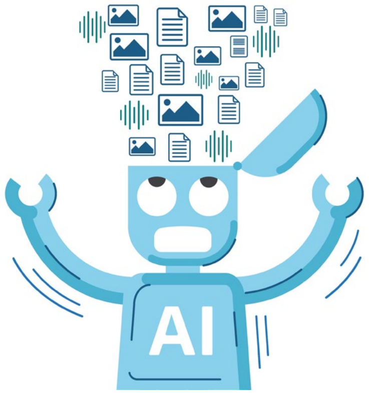
- Language Models are optimized to produce *plausible* texts, not necessarily *correct* ones!
- Incorrect answers may come from the training data, which may contain various types of errors or disinformation…
- … But language models may still hallucinate with a “perfect” training set!
- And often do so in an overly *confident tone*

Broader questions

- Biases and limitations of LLMs
  - And how those are shaped by the underlying training data
- Societal aspects
  - The role of content creators
  - Loss of autonomy
  - Privacy
  - Anthropomorphism
  - Transparency

- ➔ We need researchers from the humanities and social sciences to help us with those questions!

Take-home messages



- Large Language Models are **powerful tools** that can be applied to a wide array of language processing tasks:
  - *"Document intelligence":* information extraction, classification, filtering
  - *Writing assistance*: transcription, translation, summarization, "co-pilots"
- But we also need to be aware of their **limitations**
- … And discuss how to **regulate** their deployment (notably in terms of *accountability*, *transparency* & *data protection*)