Natural Language Processing

LLM architectures

Marco Kuhlmann
Department of Computer and Information Science

LINKÖPING UNIVERSITY

This session

- LLM architectures (Q&A)
- Deep-dive into attention
- Attention is explanation? (in-class assignment)
- Looking ahead (lab 2, unit 3)

LLM architectures (Q&A)

Overview of Unit 2

2.1 Introduction to machine translation

2.2 Neural machine translation

2.3 Attention

2.4 The Transformer architecture

2.5 Decoder-based language models (GPT)

2.6 Encoder-based language models (BERT)

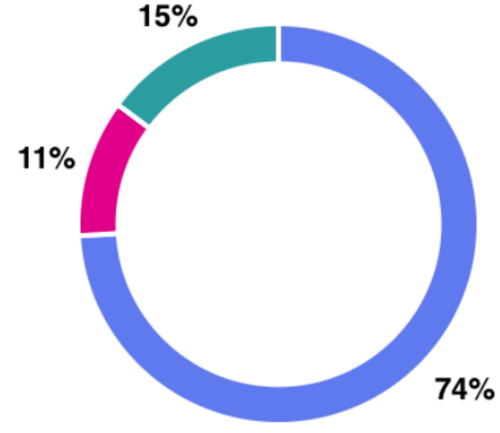
Lecture 2.2, question 5

5. Why do we use length normalisation together with beam search in decoding? (1 point)

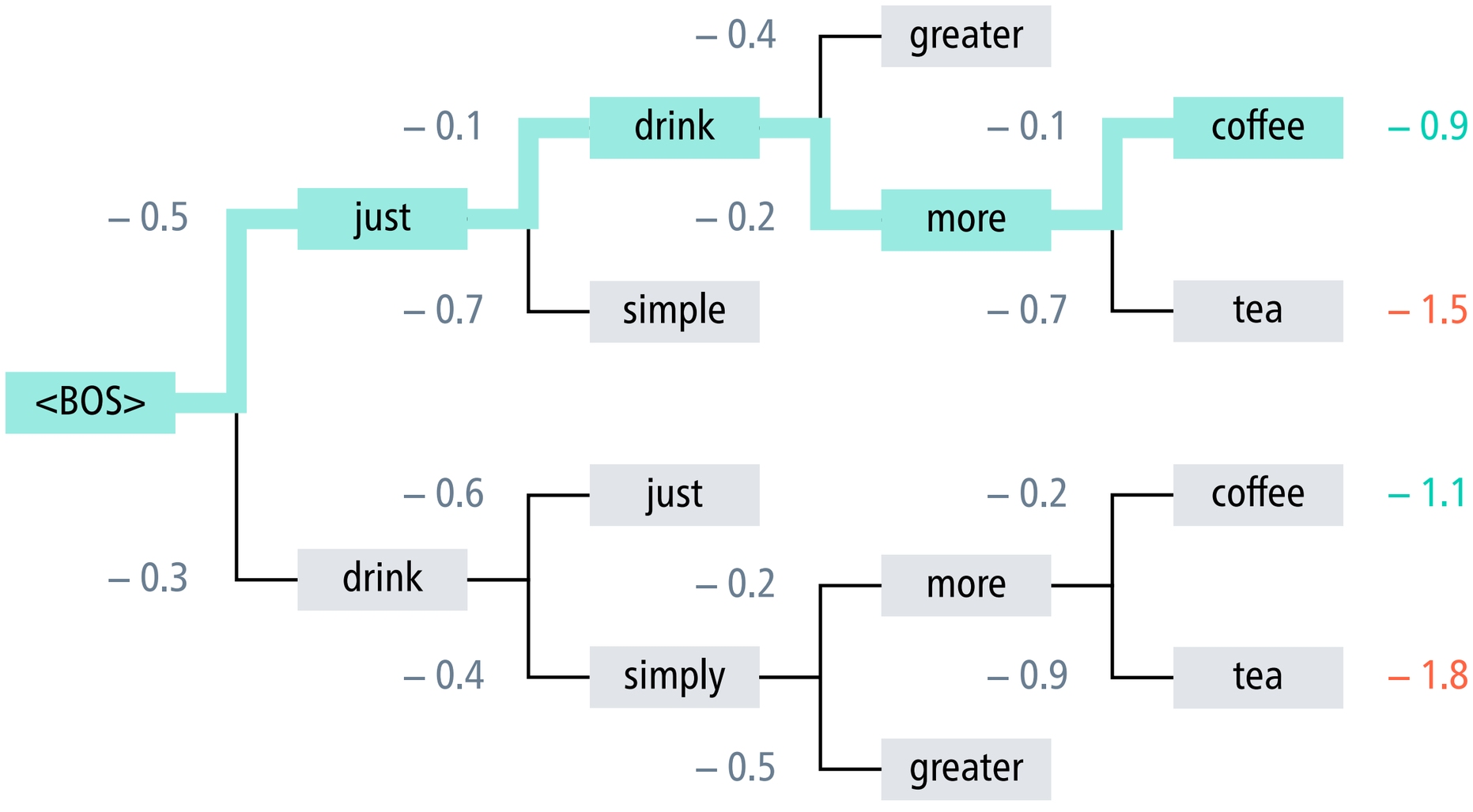
More details

74% of respondents answered this question correctly.

We do not want to penalise long translations.	20 ✓
We do not want to penalise short translations.	3
We want to avoid numerical overflow.	4



Beam search example



Lecture 2.3, question 3

3. Consider following values for the example of “Contextual word embeddings via attention”.

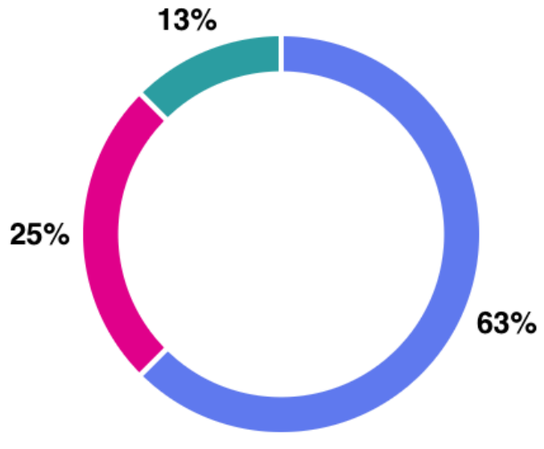
h1 = [0.5539, 0.7239], h2 = [0.4111, 0.3878], h3 = [0.2376, 0.1264]

More details

Assuming that the attention score is computed using the unscaled dot product, what is the refined representation for h2? (1 point)

63% of respondents answered this question correctly.

[0.4198, 0.4488]	15 ✓
[0.5084, 0.3194, 0.1467]	6
[0.3962, 0.3279, 0.2759]	3



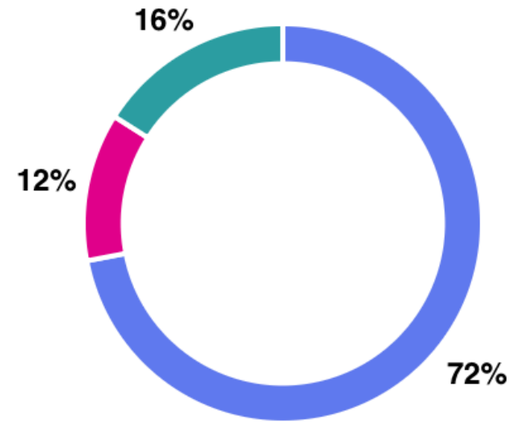
Lecture 2.3, question 4

4. Which of the following statements about the more general characterisation of attention in terms of queries, keys and values is true? (1 point)

More details

72% of respondents answered this question correctly.

The output has the same length as each value.	18 ✓
The query has the same length as each value.	3
Each key has the same length as each value.	4



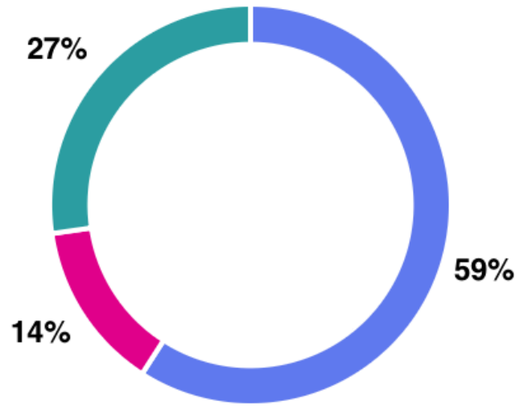
Lecture 2.4, question 2

2. Consider the example translation used to illustrate the Transformer architecture. Which of the following statements is false? (1 point)

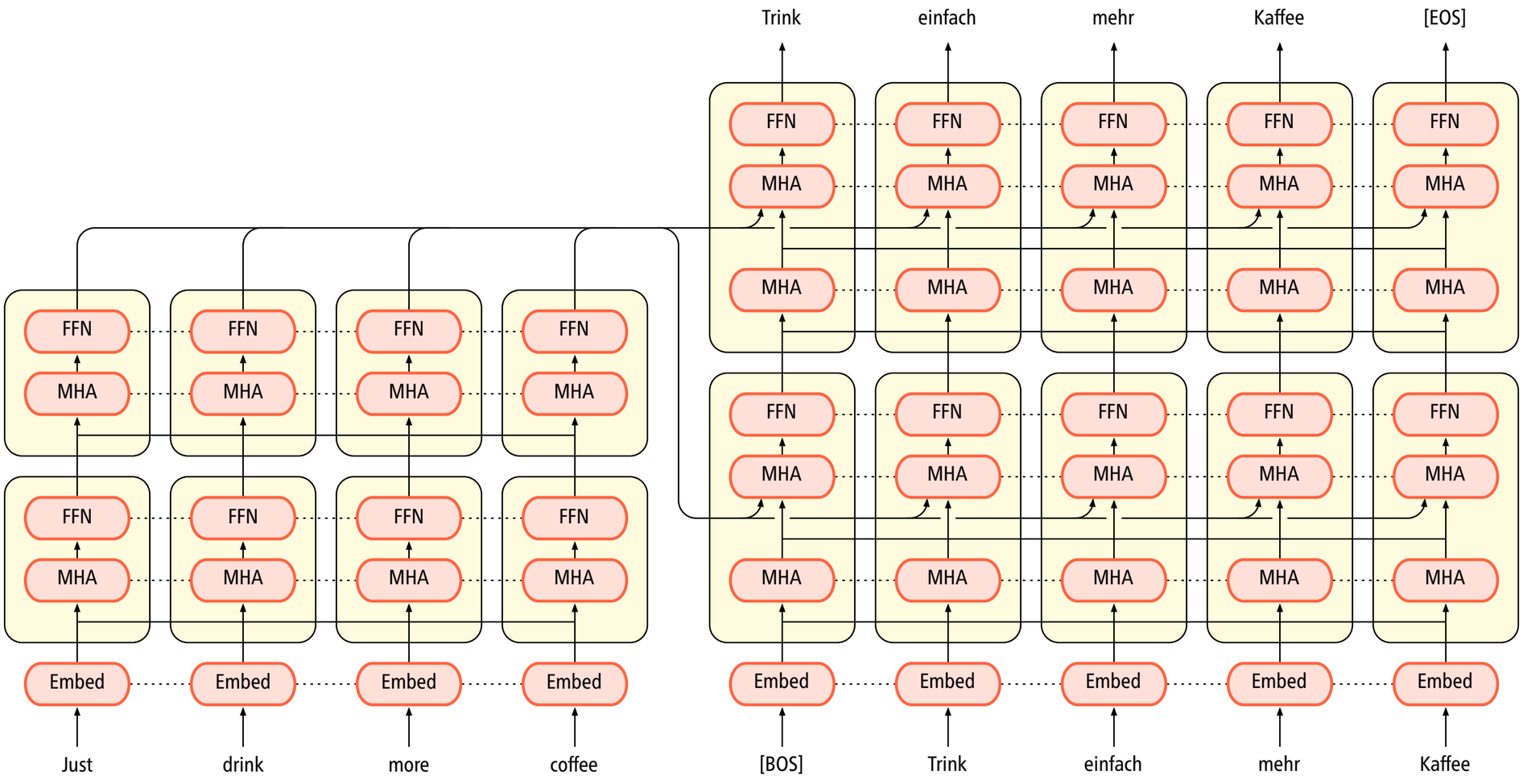
More details

59% of respondents answered this question correctly.

The final encoder representation of *drink* depends on the token embedding of *Kaffee*.	13 ✓
The final encoder representation of *coffee* depends on the token embedding of *drink*.	3
The final decoder representation of *Kaffee* depends on the final encoder representation of *coffee*.	6



Example translation



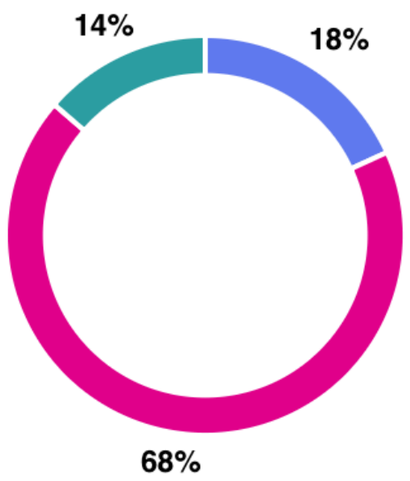
Lecture 2.4, question 5

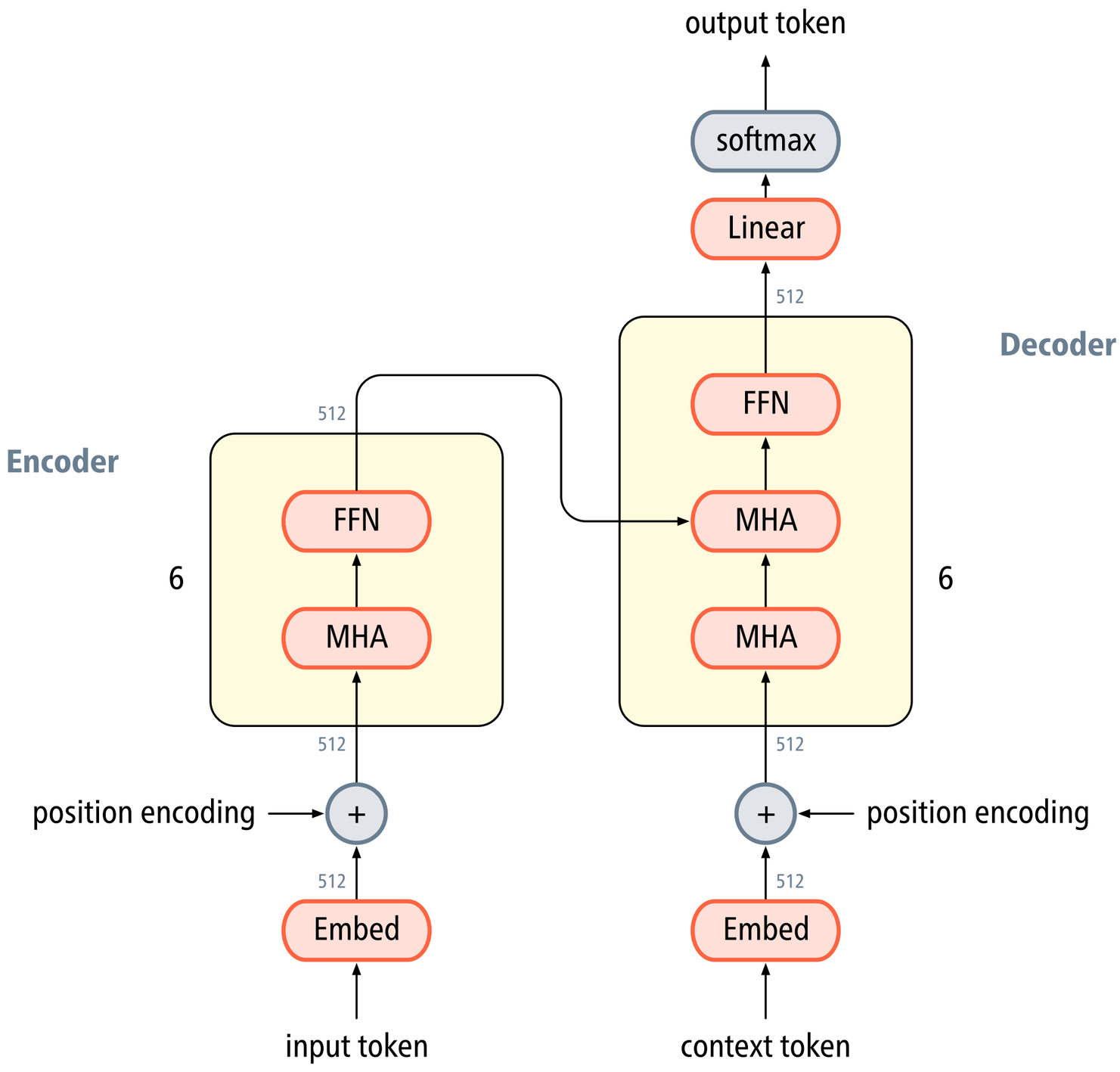
5. True or false: Permuting the input tokens to a Transformer encoder does not change the final token representations. (1 point)

More details

68% of respondents answered this question correctly.

True	4	
False	15	✓
Depends on the input tokens	3	




output token
softmax
Linear
512
Decoder
FFN
MHA
6
512
Encoder
FFN
MHA
6
512
position encoding
+
512
Embed
input token
512
+
position encoding
512
Embed
context token

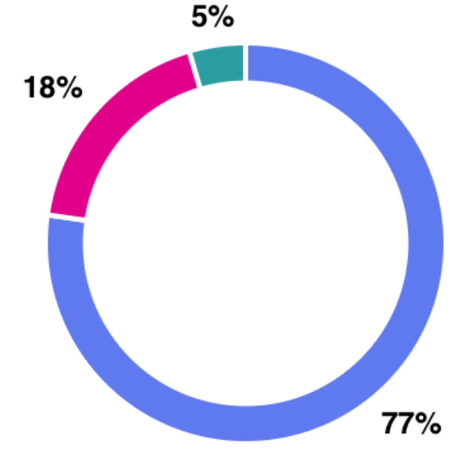
Lecture 2.5, question 2

2. Looking at the original GPT model architecture (Radford et al., 2018), what is the approximate number of trainable parameters in the FNN? (1 point)

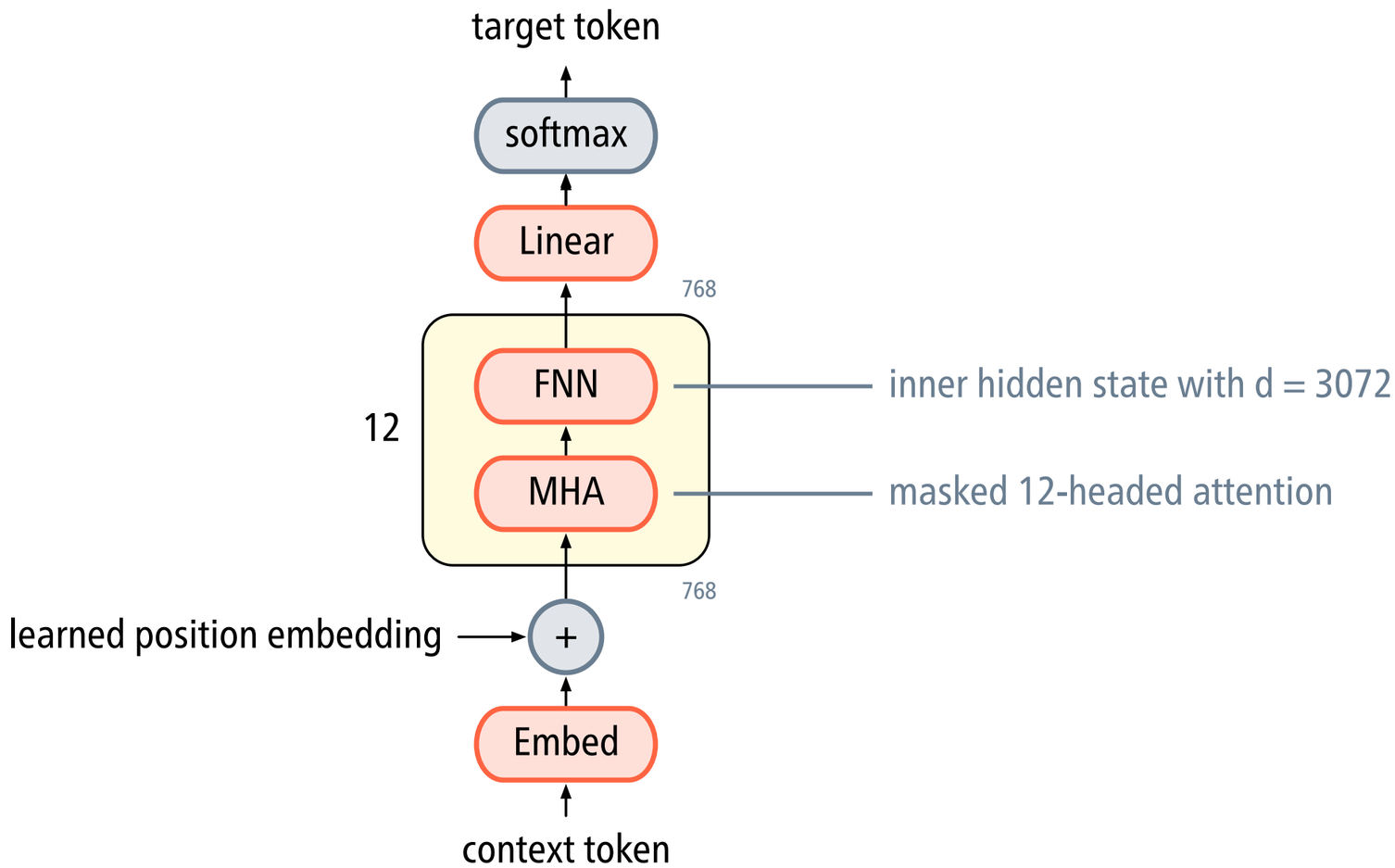
More details

77% of respondents answered this question correctly.

4718592	17 ✓
589824	4
9216	1



GPT model architecture



Radford et al. (2018)

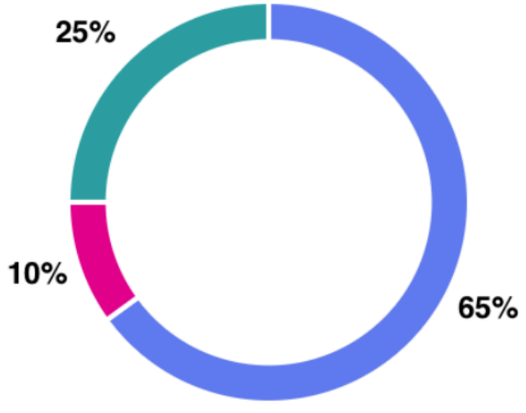
Lecture 2.6, question 3

3. What is the main purpose of the [CLS] token in BERT? (1 point)

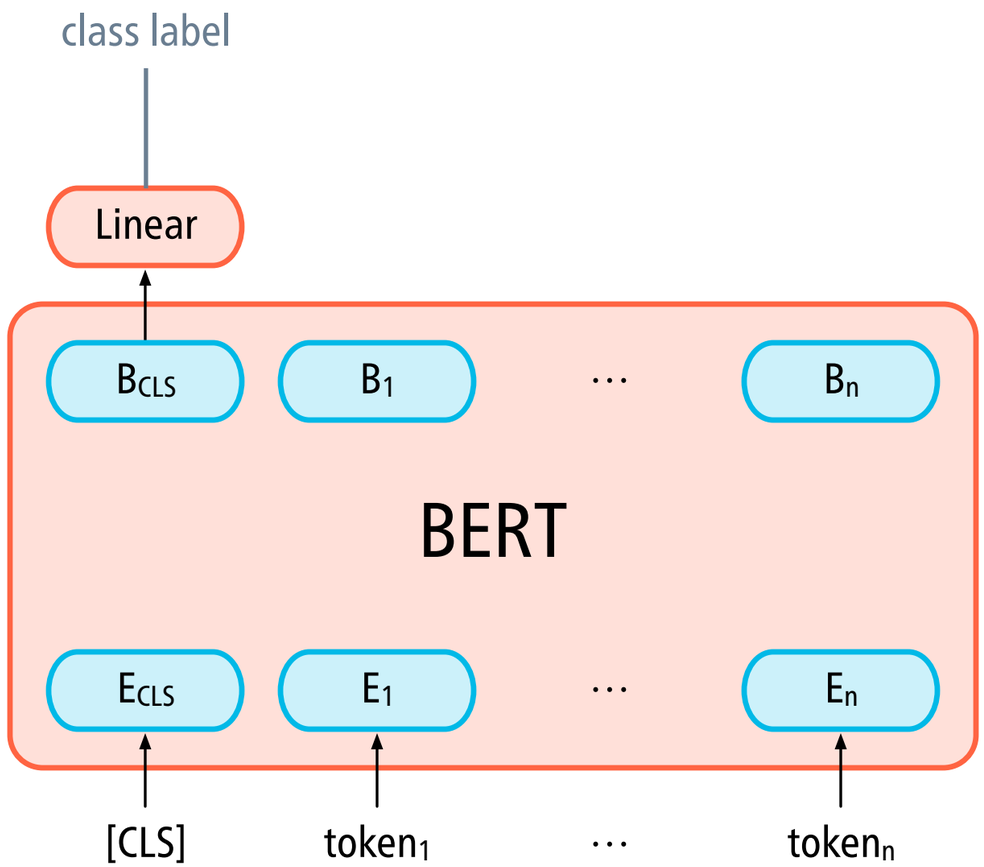
More details

65% of respondents answered this question correctly.

It is used as a representation of the complete input sentence pair.	13 ✓
It is used as a padding token.	2
It is used for the masked language modelling task.	5

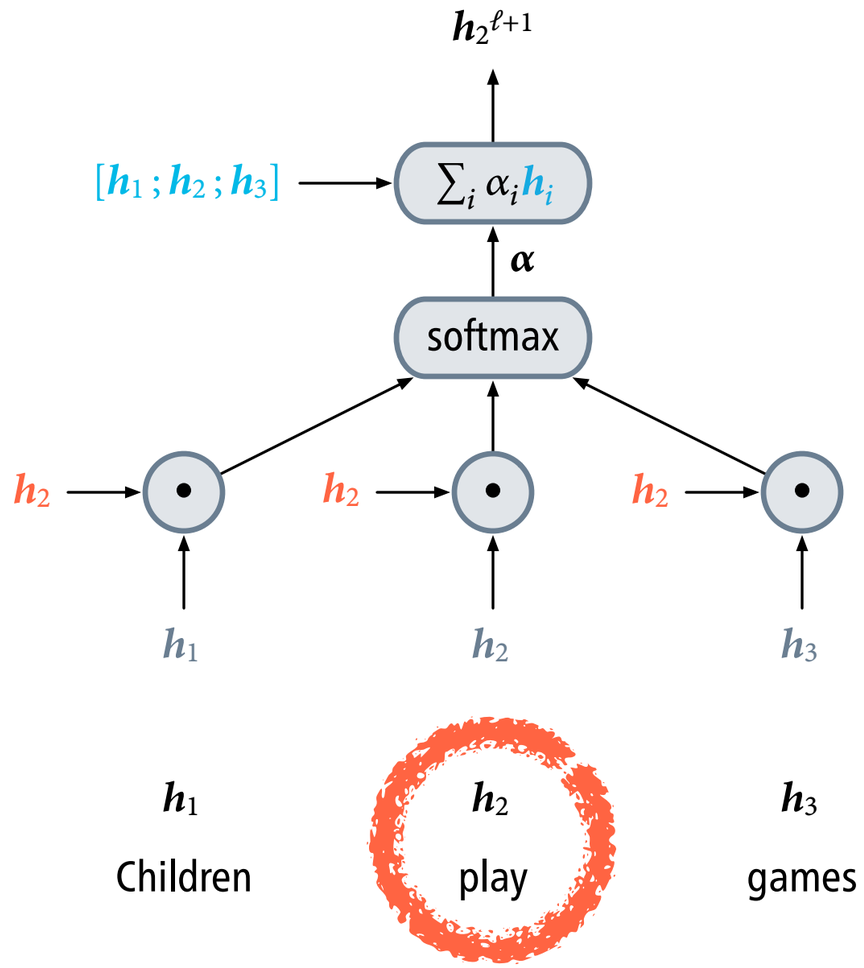


Fine-tuning on a single-sentence classification task



Deep-dive into attention

Contextual embeddings via attention



A general characterisation of attention

- In general, attention can be described as a mapping from a query $\boldsymbol{q}$ and a set of key–value pairs $\langle \boldsymbol{k}_i, \boldsymbol{v}_i \rangle$ to an output.
- The output is the weighted sum of the $\boldsymbol{v}_i$, where the weight of each $\boldsymbol{v}_i$ is given by the attention score between $\boldsymbol{q}$ and $\boldsymbol{k}_i$:

$$\boldsymbol{\alpha}_i = \text{softmax}(\text{score}(\boldsymbol{q}, \boldsymbol{K}))\boldsymbol{V}$$

$$\boldsymbol{q} \in \mathbb{R}^{d_Q},\ \boldsymbol{K} \in \mathbb{R}^{n \times d_K},\ \boldsymbol{V} \in \mathbb{R}^{n \times d_V},\ d_Q = d_K$$

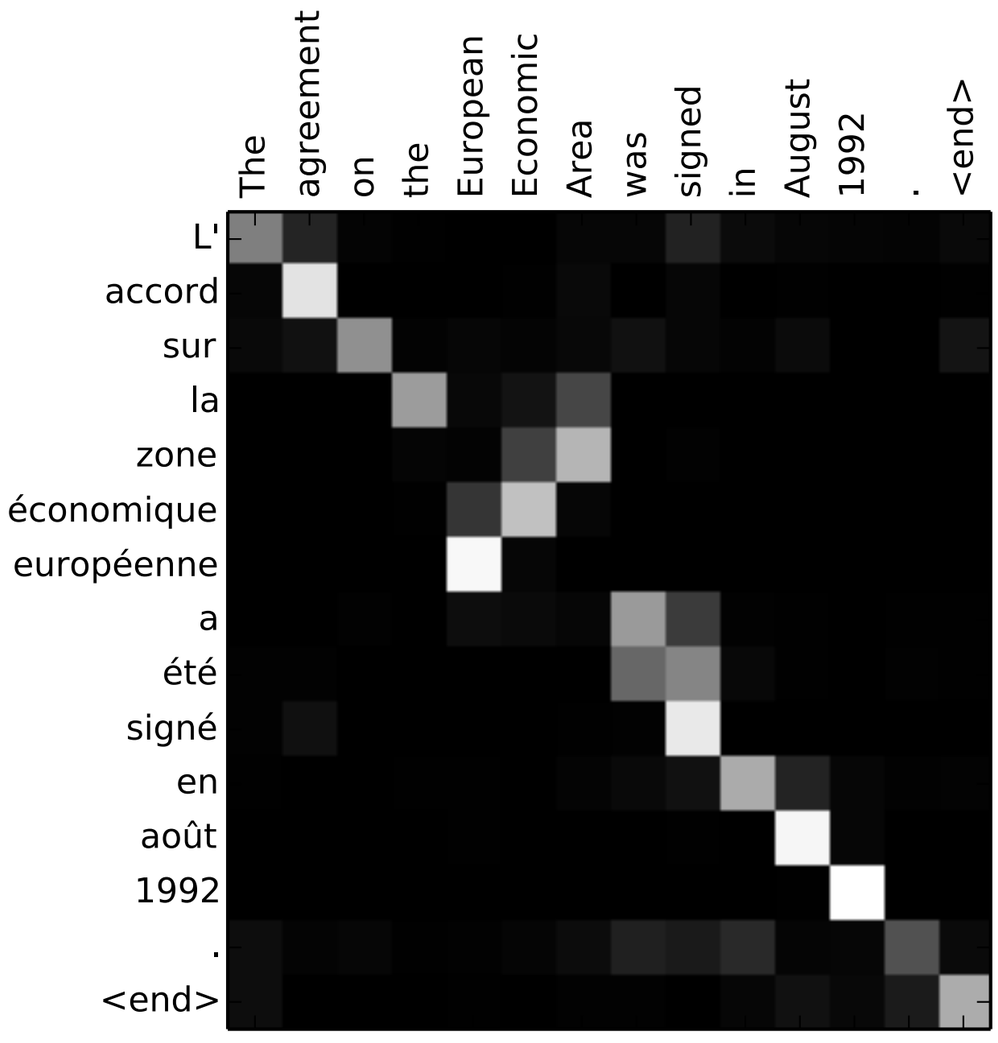
Vaswani et al. (2017)

In-class assignment

https://forms.office.com/e/wx071fCvwd

Attention is explanation?

Attention as word alignments



In the context of the encoder–decoder architecture for neural machine translation, attention weights resemble soft word alignments.

Image source: Bahdanau et al. (2015)

Attention is not Explanation

Sarthak Jain
Northeastern University
jain.sar@husky.neu.edu

Byron C. Wallace
Northeastern University
b.wallace@northeastern.edu

Abstract

Attention mechanisms have seen wide adoption in neural NLP models. In addition to improving predictive performance, these are often touted as affording transparency: models equipped with attention provide a distribution over attended-to input units, and this is often presented (at least implicitly) as communicating the relative importance of inputs. However, it is unclear what relationship exists between attention weights and model outputs. In this work we perform extensive experiments across a variety of NLP tasks that aim to assess the degree to which attention weights provide meaningful "explanations" for predictions. We find that they largely do not. For example, learned attention weights are frequently uncorrelated with gradient-based measures of feature importance, and one can identify very different attention distributions that nonetheless yield equivalent predictions. Our findings show that standard attention modules do not provide meaningful explanations and should not be treated as though they do. Code to reproduce all experiments is available at https://github.com/successar/AttentionExplanation.

1 Introduction and Motivation

Attention mechanisms (Bahdanau et al., 2014) induce conditional distributions over input units to compose a weighted context vector for downstream modules. These are now a near-ubiquitous component of neural NLP architectures. Attention weights are often claimed (implicitly or explicitly) to afford insights into the "inner-workings" of models: for a given output one can inspect the inputs to which the model assigned large attention weights. Li *et al.* (2016) summarized this commonly held view in NLP: "Attention provides an important way to explain the workings of neural models". Indeed, claims that attention provides interpretability are common in the literature, e.g., (Xu et al., 2015; Choi et al., 2016; Lei et al., 2017; Martins and Astudillo, 2016; Xie et al., 2017; Mullenbach et al., 2018).[1]

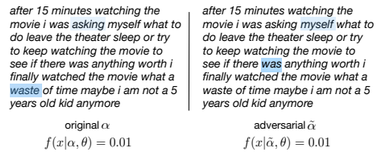


Figure 1: Heatmap of attention weights induced over a negative movie review. We show observed model attention (left) and an adversarially constructed set of attention weights (right). Despite being quite dissimilar, these both yield effectively the same prediction (0.01).

Implicit in this is the assumption that the inputs (e.g., words) accorded high attention weights are responsible for model outputs. But as far as we are aware, this assumption has not been formally evaluated. Here we empirically investigate the relationship between attention weights, inputs, and outputs.

Assuming attention provides a faithful explanation for model predictions, we might expect the following properties to hold. (i) Attention weights should correlate with feature importance measures (e.g., gradient-based measures); (ii) Alternative (or *counterfactual*) attention weight configurations ought to yield corresponding changes in prediction (and if they do not then are equally plausible as explanations). We report that neither property is consistently observed by a BiLSTM with a standard attention mechanism in the context of text classification, question answering (QA), and Natural Language Inference (NLI) tasks.

[1] We do not intend to single out any particular work; indeed one of the authors has himself presented (supervised) attention as providing interpretability (Zhang et al., 2016).

In-class assignment

- Answer the question on the form:

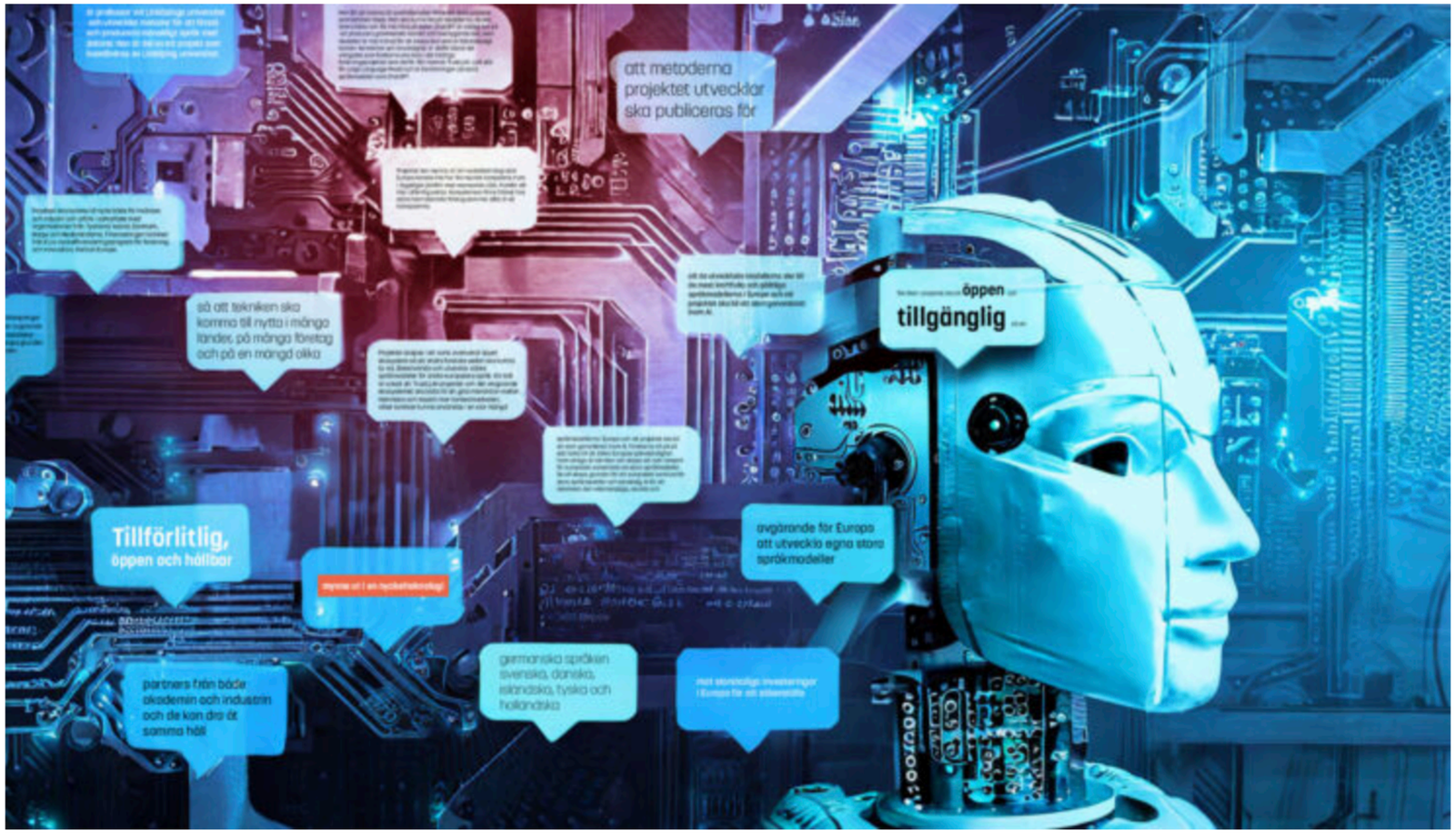
  Jain and Wallace (2019) criticise the claim that attention provides "insights into the 'inner-workings' of models". One technique they use to substantiate their criticism is the construction of counterfactual attention weight configurations. Explain the idea behind this technique in your own words.

Developing a reliable ChatGPT for European languages

01 November 2023

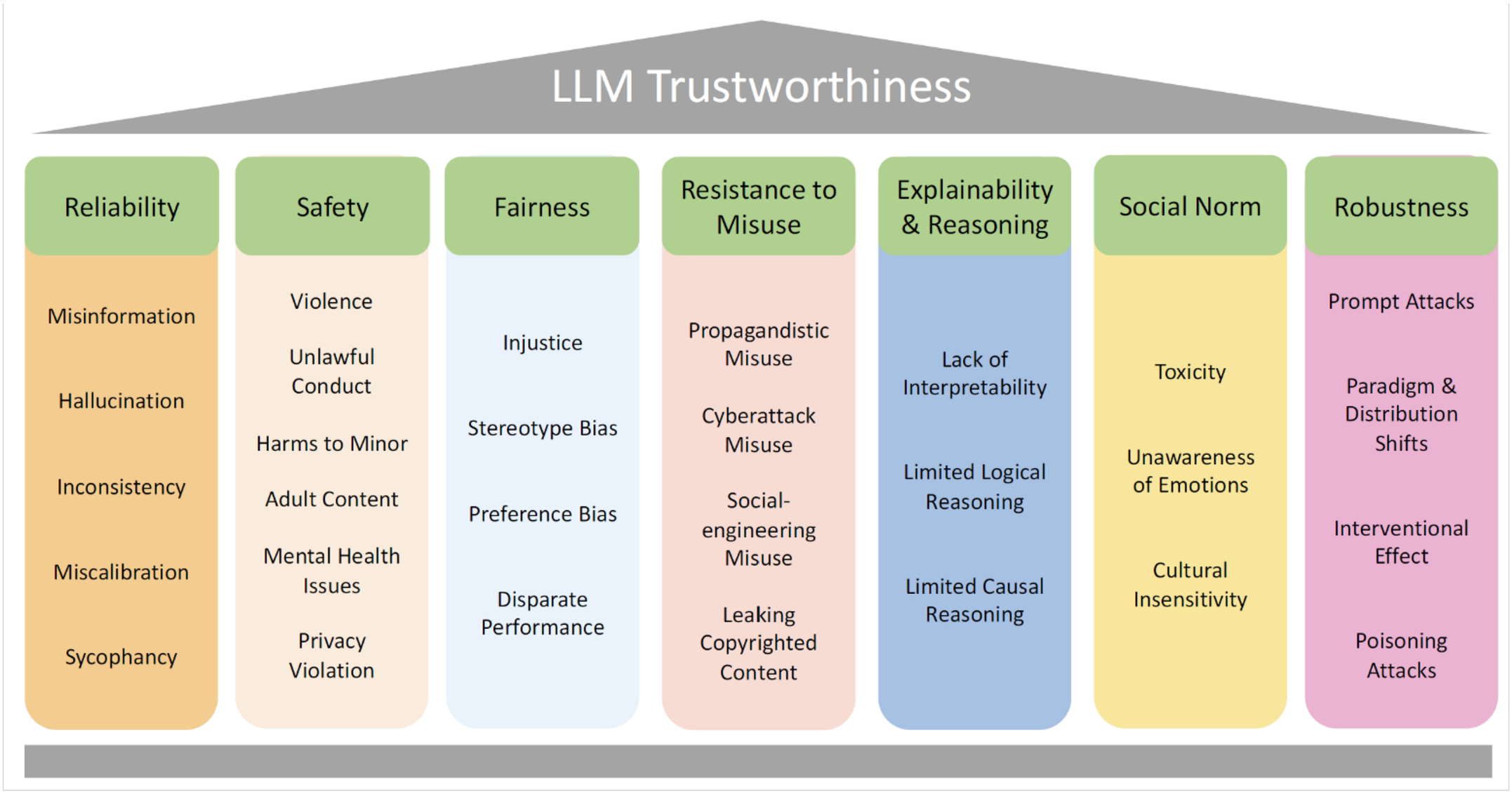
Sara Läthén

Reliable, open and sustainable – those are the qualities aimed for in a ChatGPT for European languages including Swedish. But there are still big problems that need to be solved before ChatGPT can be considered beneficial to society.



Bilden är delvis skapad med AI.

Marco Kuhlmann, professor at Linköping University, develops methods to understand and computer generate human speech. He is part of a project coordinated by Linköping University that has just been granted around SEK 70 million.


LLM Trustworthiness
Reliability
Misinformation
Hallucination
Inconsistency
Miscalibration
Sycophancy
Safety
Violence
Unlawful Conduct
Harms to Minor
Adult Content
Mental Health Issues
Privacy Violation
Fairness
Injustice
Stereotype Bias
Preference Bias
Disparate Performance
Resistance to Misuse
Propagandistic Misuse
Cyberattack Misuse
Social-engineering Misuse
Leaking Copyrighted Content
Explainability & Reasoning
Lack of Interpretability
Limited Logical Reasoning
Limited Causal Reasoning
Social Norm
Toxicity
Unawareness of Emotions
Cultural Insensitivity
Robustness
Prompt Attacks
Paradigm & Distribution Shifts
Interventional Effect
Poisoning Attacks

Source

In-class assignment

- Discuss in small groups:

  Inspecting attention patterns has been used as a technique for explaining NLP systems based on Transformer models. Give an example of (1) an aspect of trust which this technique can help build; (2) an aspect of trust that is beyond this technique.

- Summarise your discussion in the form by providing one example from each category: (1) an aspect of trust which attention patterns can build; (2) an aspect that is beyond.

In-class assignment

https://forms.office.com/e/wx071fCvwd

Looking ahead (lab 2, unit 3)