

# LLMs for classification and regression

Dirk Wulff & Zak Hussain



**MAX PLANCK INSTITUTE**  
FOR HUMAN DEVELOPMENT



# The problem

Classification and regression

## Regression

Input  
(e.g., sentence)

Output  
(e.g., sentiment)

St. Gallen is a  
beautiful city



8.2/10

## Classification

Input  
(e.g., sentence)

Output  
(e.g., topic class)

St. Gallen is a  
beautiful city



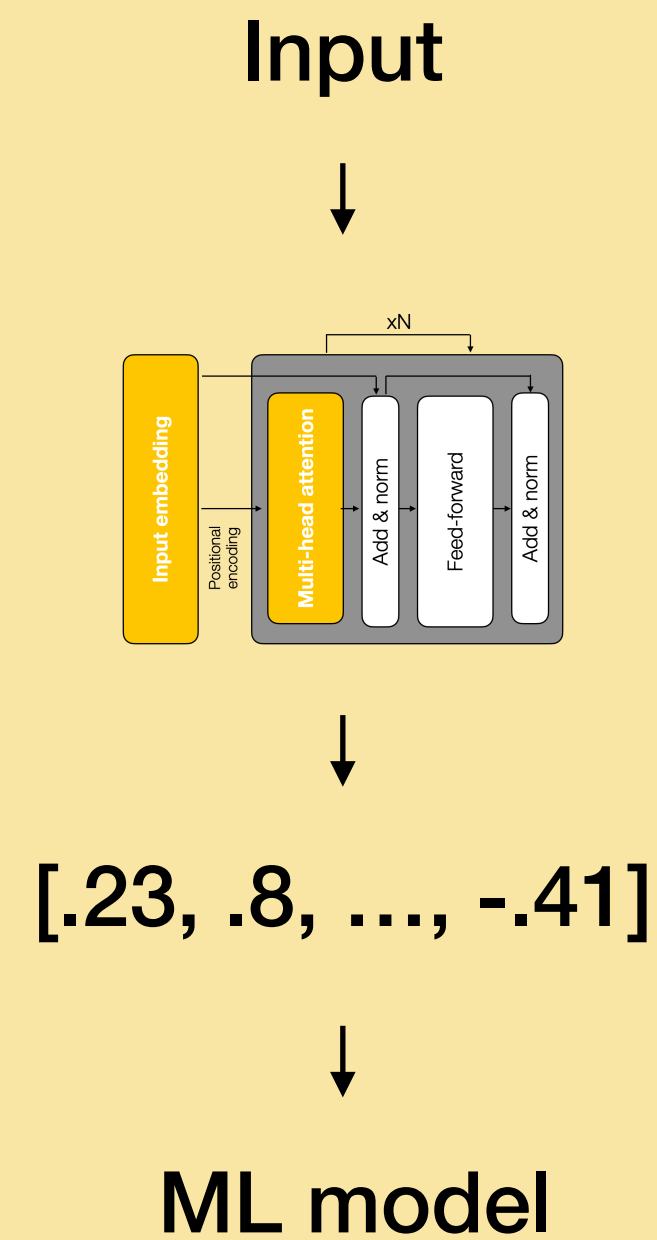
.82	Tourism
.1	Sports
.07	Politics
.01	Science

# Approaches

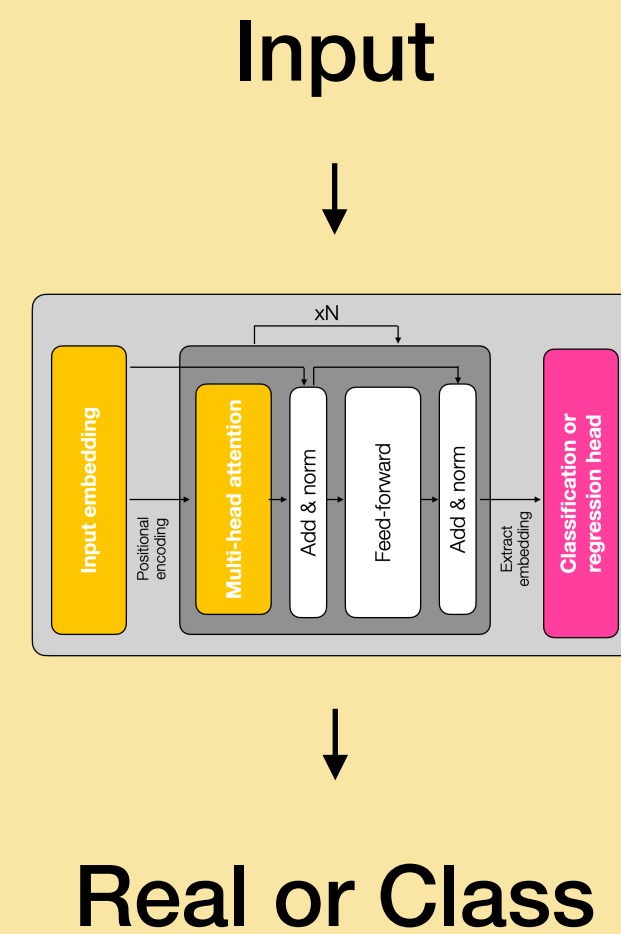
to classification and regression

require training data

## Feature extraction

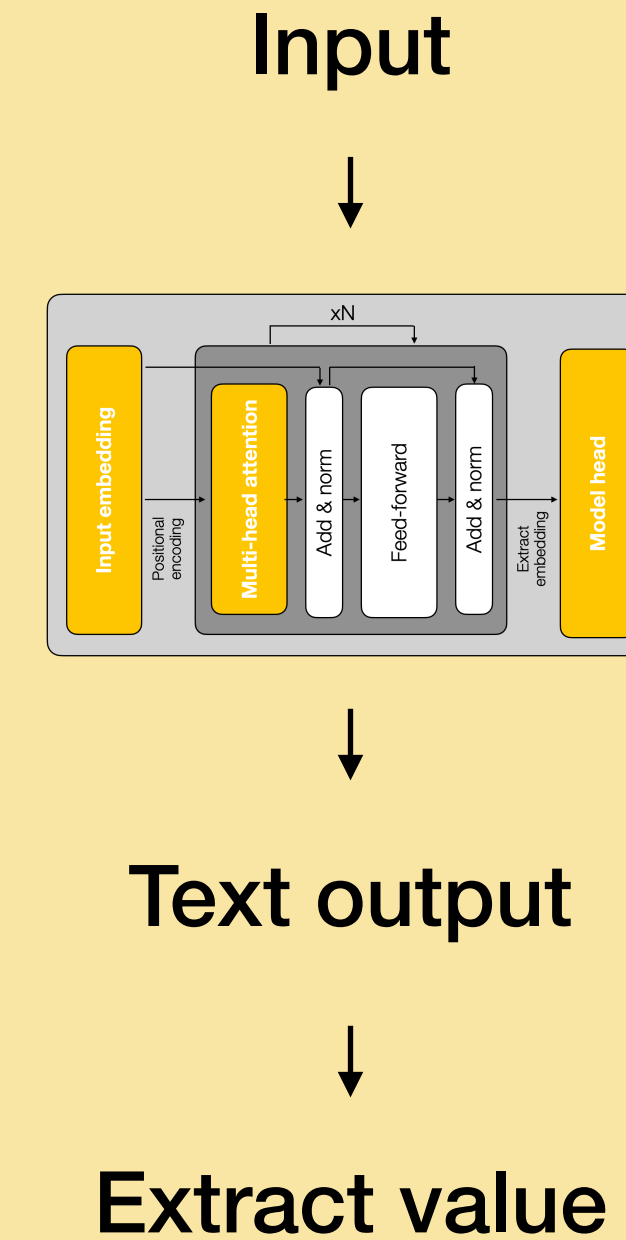


## Changing model head



require no training data

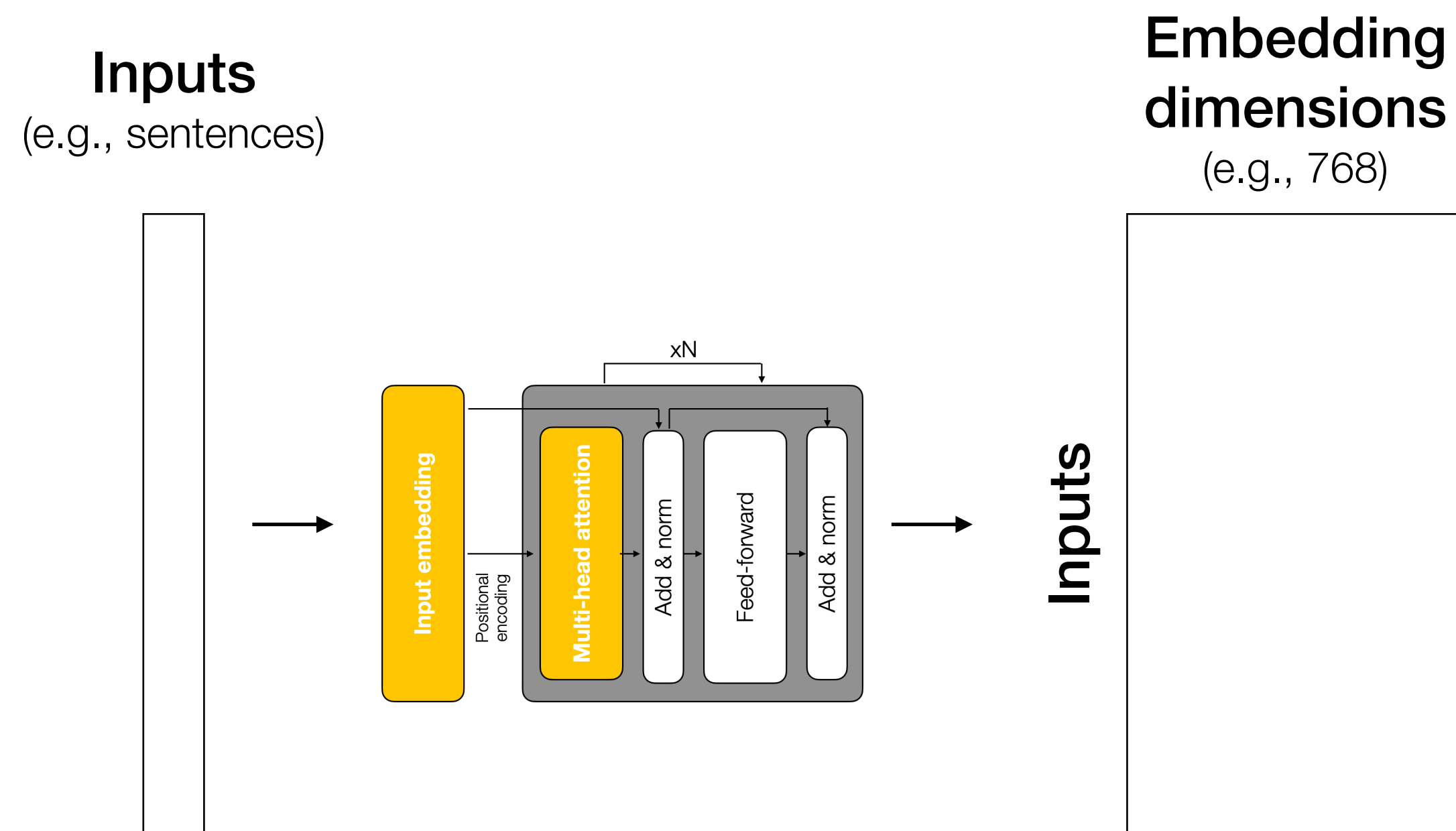
## Text generation



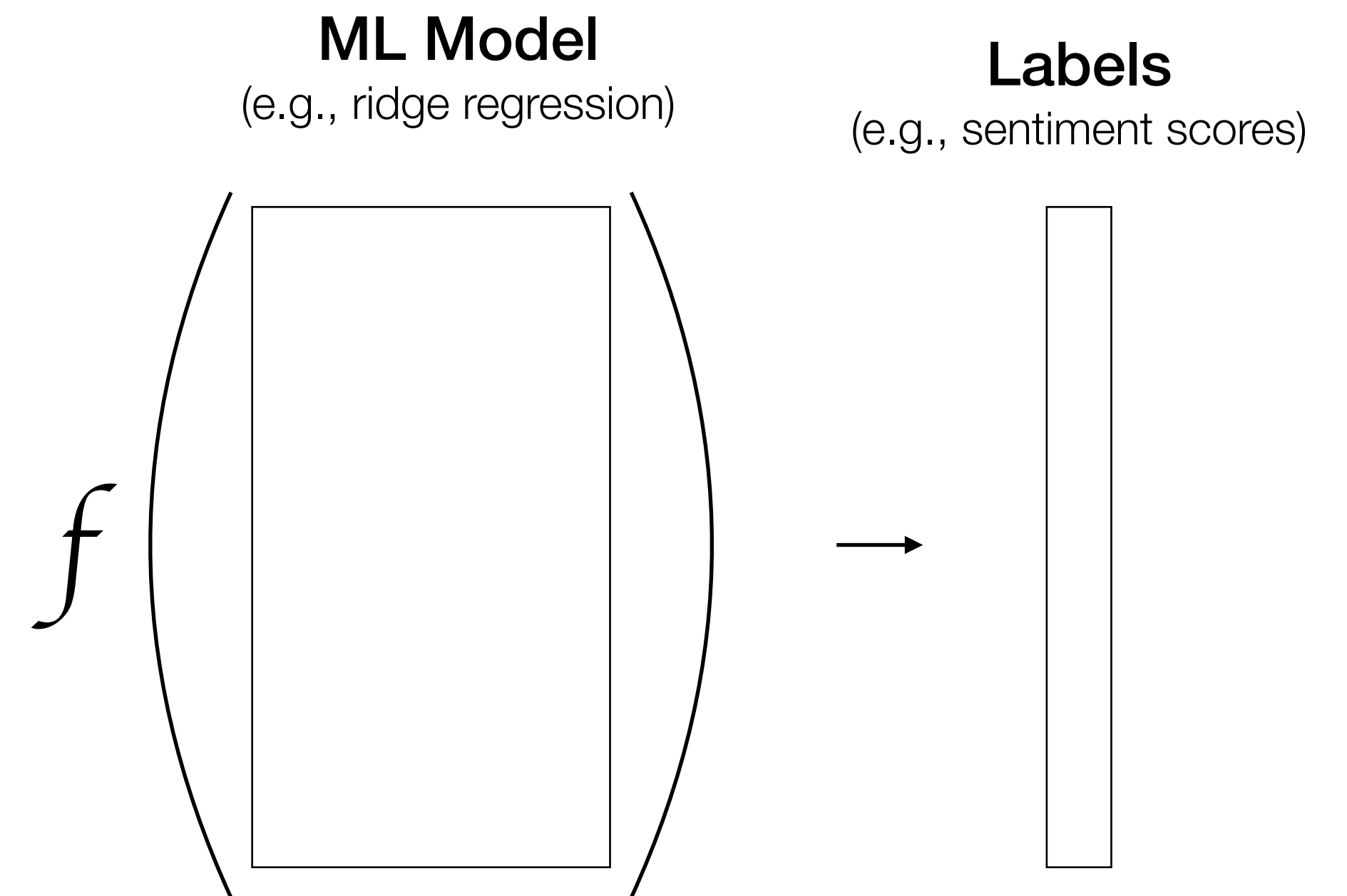
# Feature extraction

for regression and classification

Step 1 - feature extraction



Step 2 - train predictive model



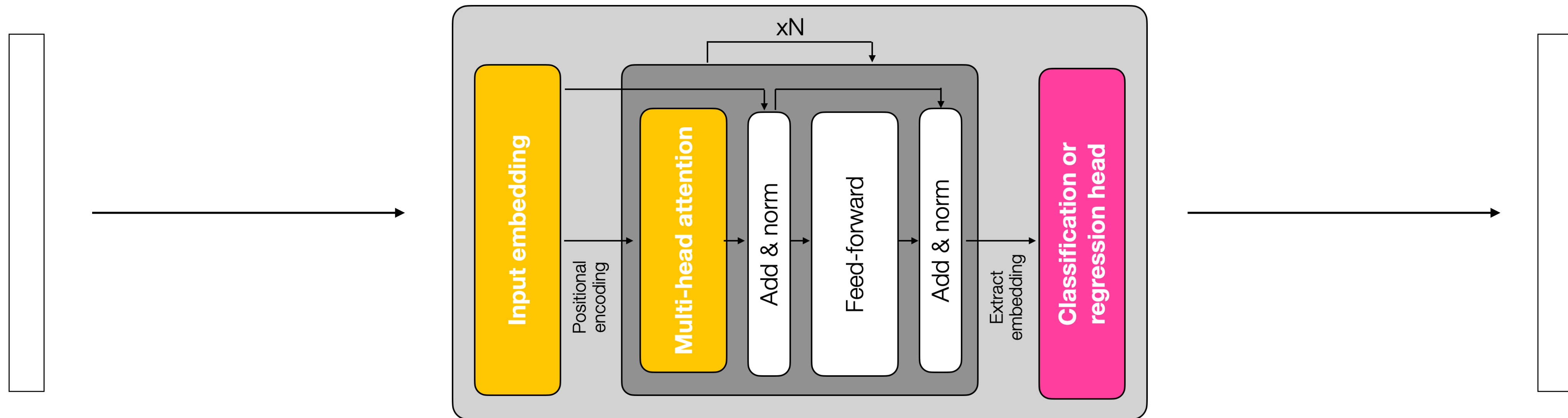
# Changing model head

for regression and classification

**Inputs**  
(e.g., sentences)

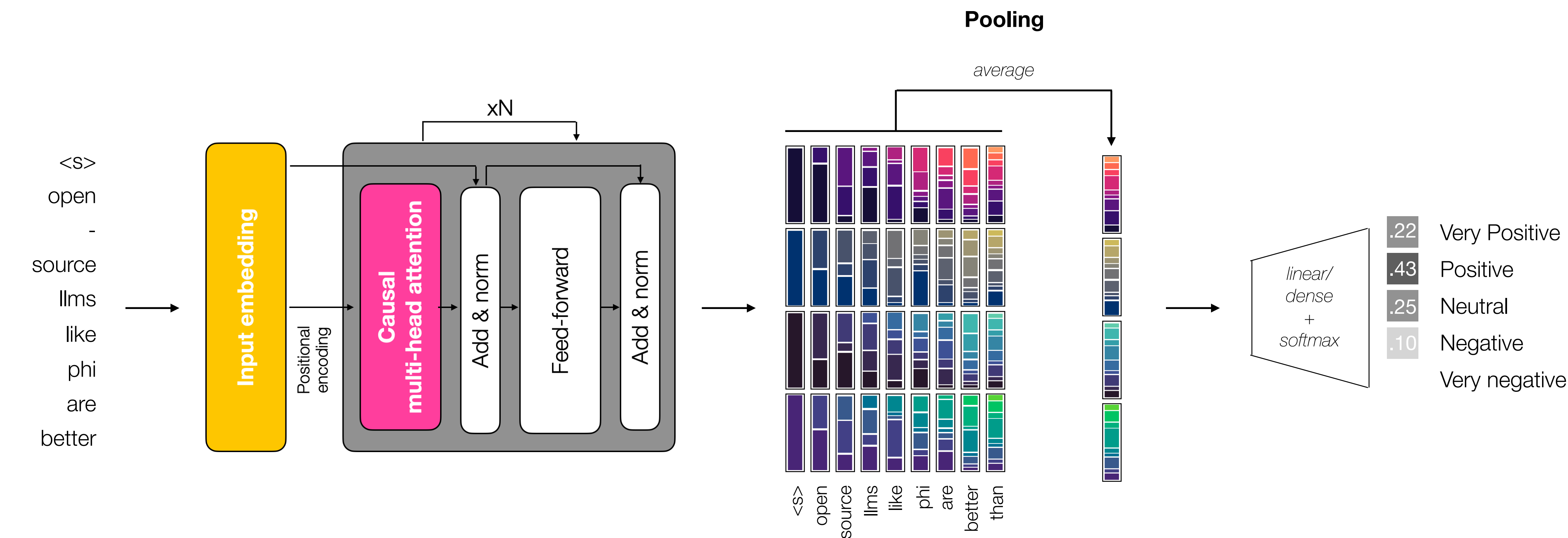
**Transformer**  
with classification or regression head

**Labels**  
(e.g., sentiment scores)



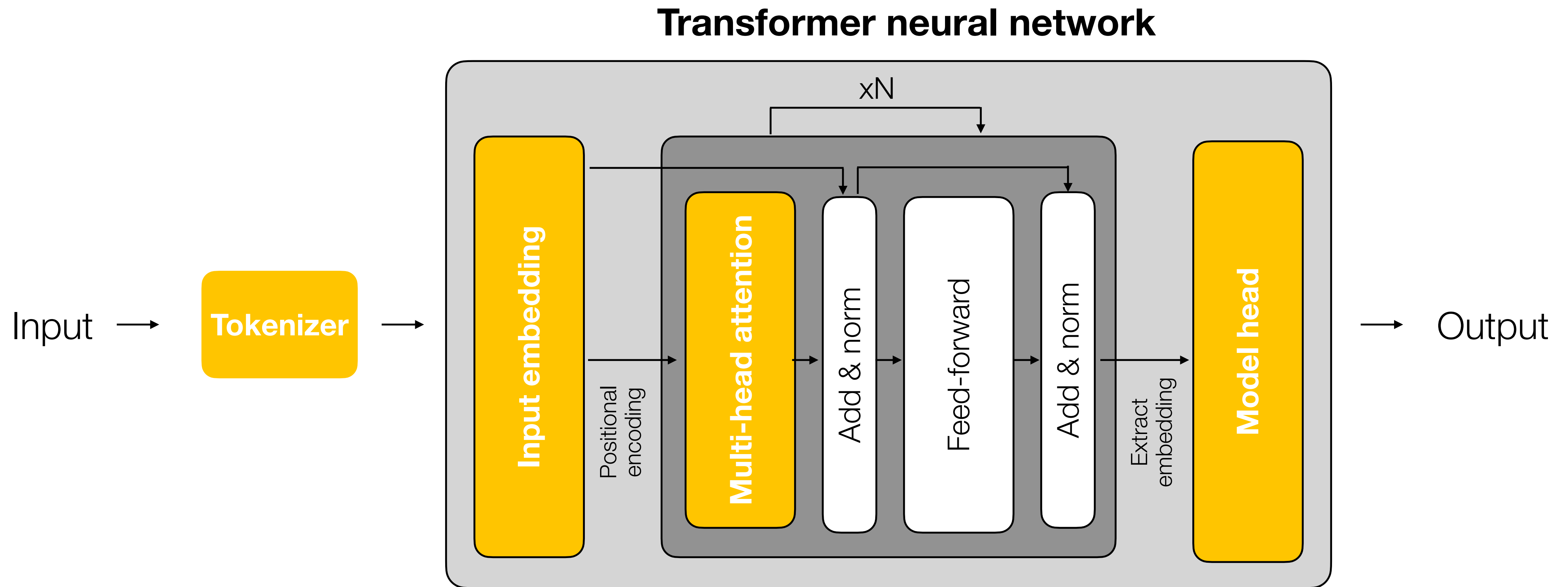
# Transformer

Model head for **classification**



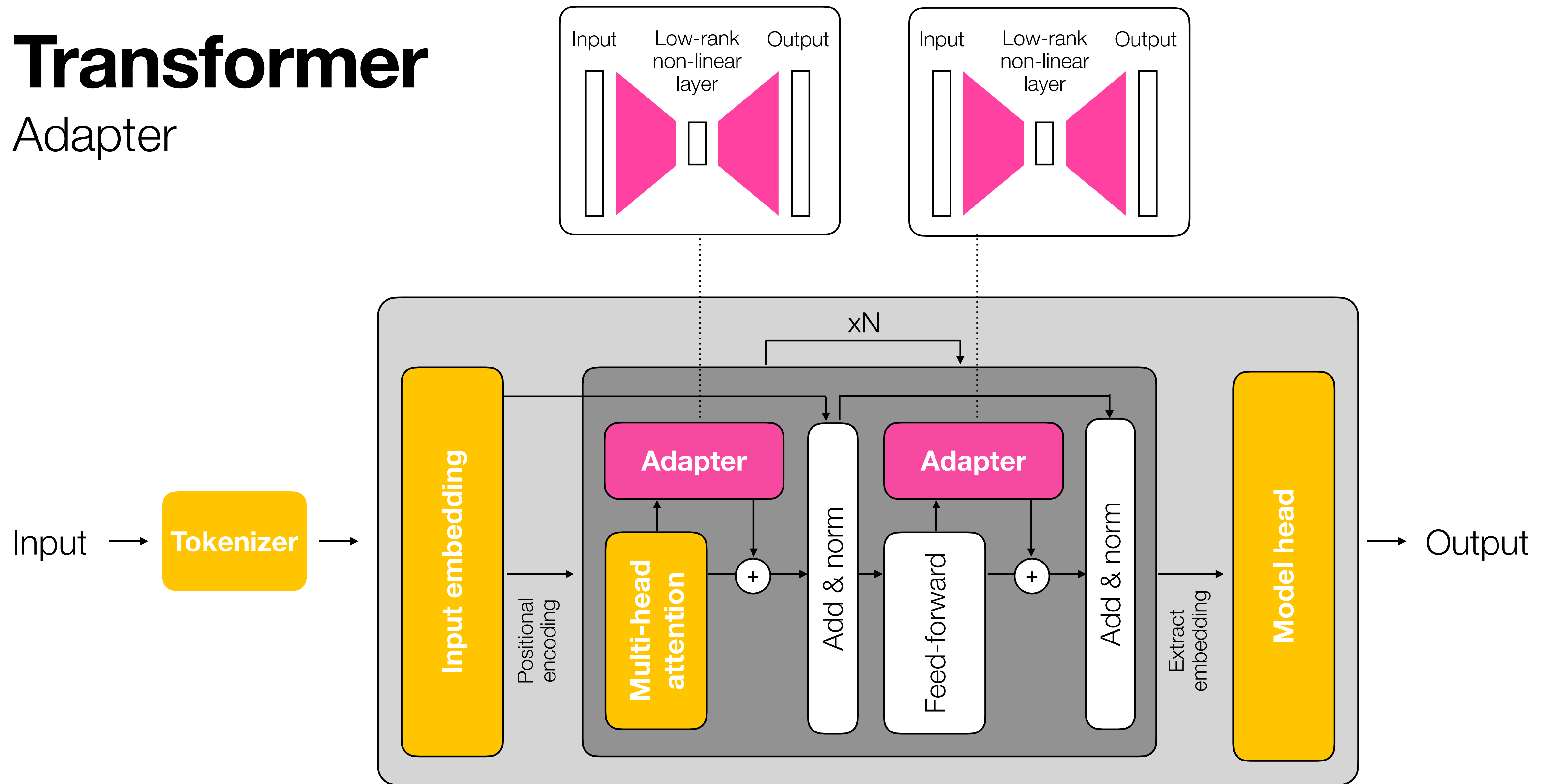
# Transformer

## Adapter



# Transformer

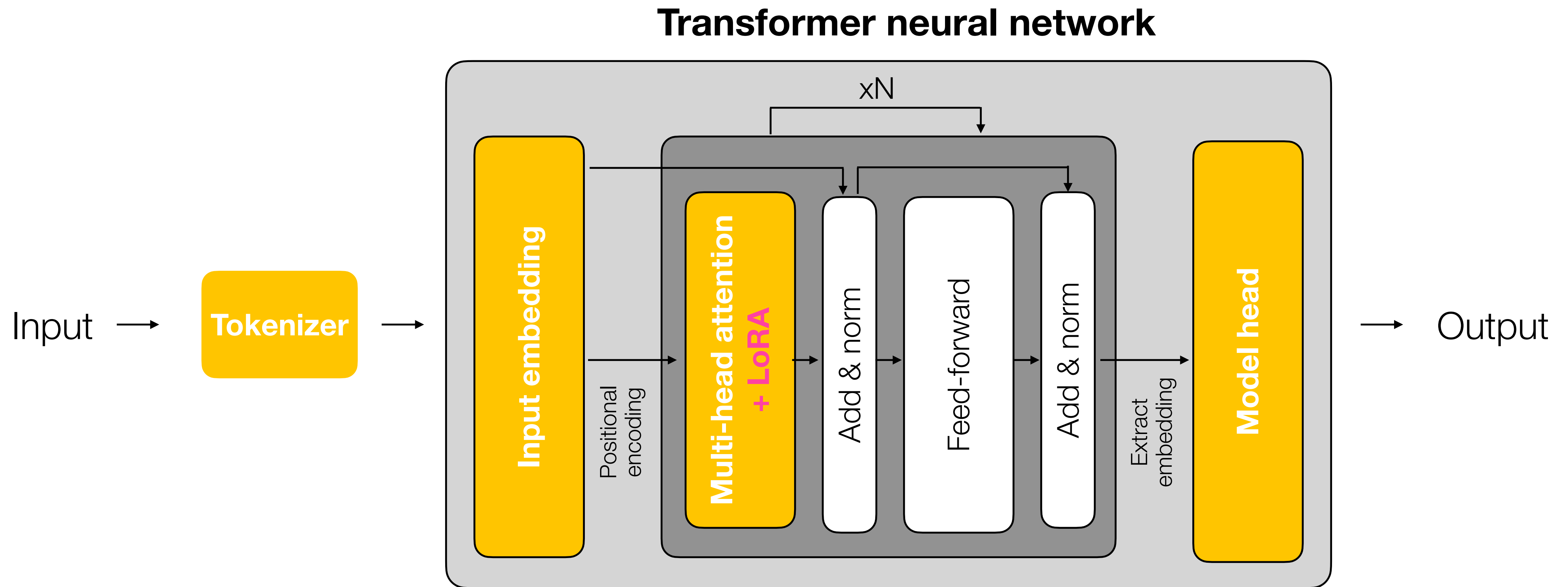
## Adapter





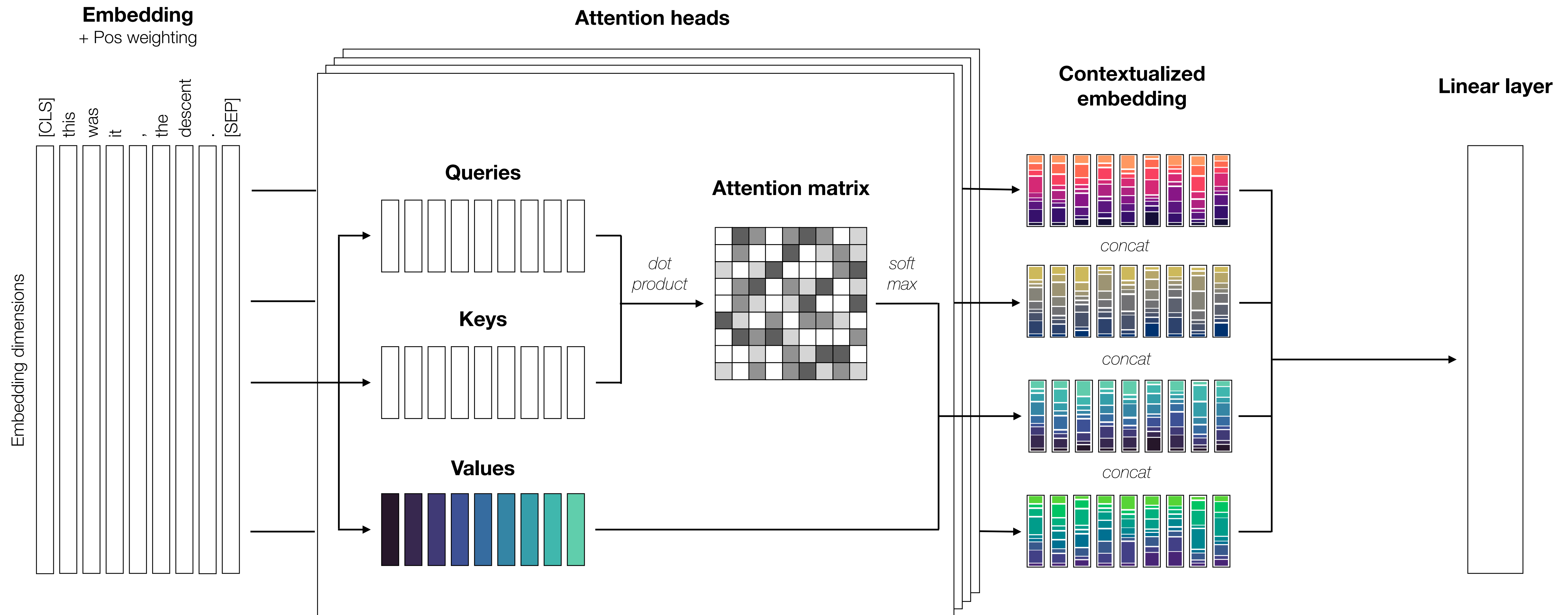
# Transformer

LoRA



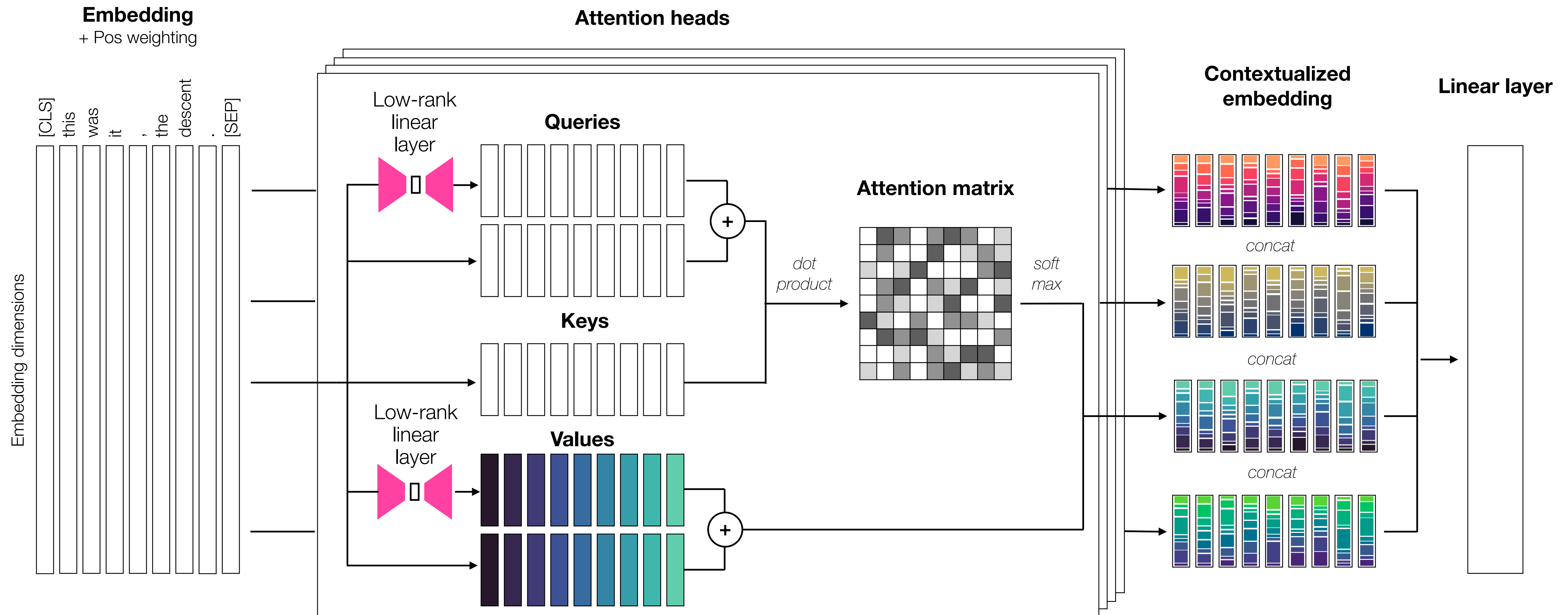
# Transformer

## Adapter



# Transformer

## Adapter



# Text generation

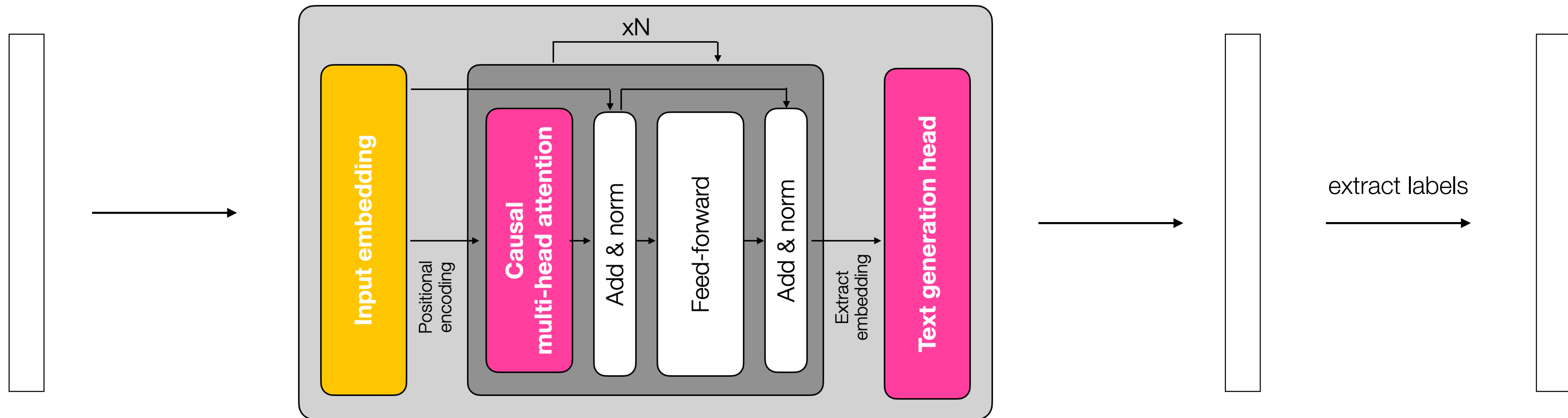
for regression and classification

**Inputs**  
e.g., sentences

**Transformer**  
for text generation

**Output**  
text response

**Labels**  
e.g., sentiment scores



# Example

## Extracting numerical estimates

(Can also be improved by better prompting)



On a scale of 0 to 100, assessing the risk of shampoo can be roughly quantified considering various factors. For most commercially available shampoos, here is a general risk assessment:

### Risk Factors:

#### 1. Chemical Composition (15-25):

- Presence of sulfates, parabens, and phthalates.
- Potential for skin and eye irritation.

#### 2. Health Effects (10-20):

- Allergic reactions.
- Long-term exposure to potentially harmful chemicals.

#### 3. Environmental Impact (5-10):

- Non-biodegradable ingredients.
- Plastic packaging waste.

### Overall Risk Assessment:

Considering these factors, most standard shampoos would fall within a risk range of:

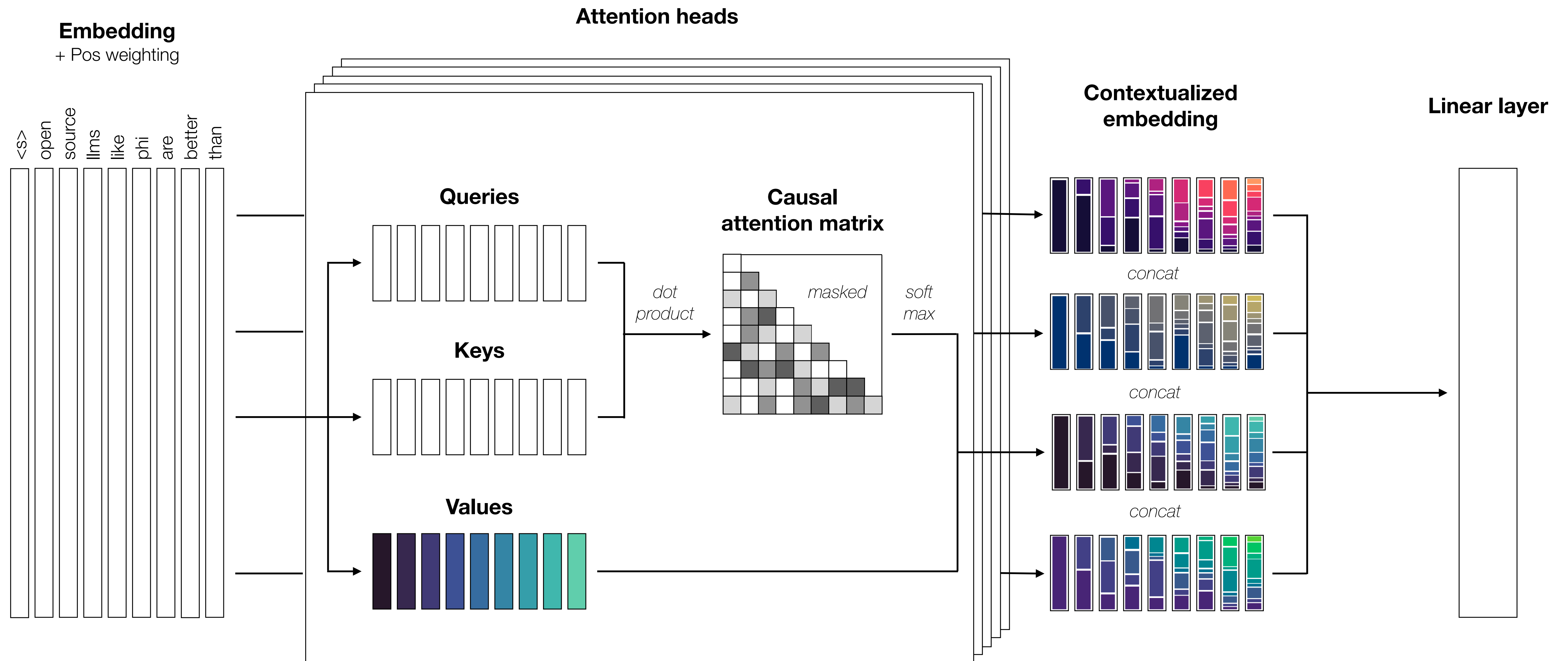
**30 to 55**

- **Lower end (30):** Shampoos with fewer synthetic chemicals, natural ingredients, and environmentally friendly packaging.
- **Higher end (55):** Shampoos with several synthetic chemicals, higher potential for irritation or allergies, and significant environmental concerns.

This assessment can vary depending on specific brands, formulations, and individual sensitivities.

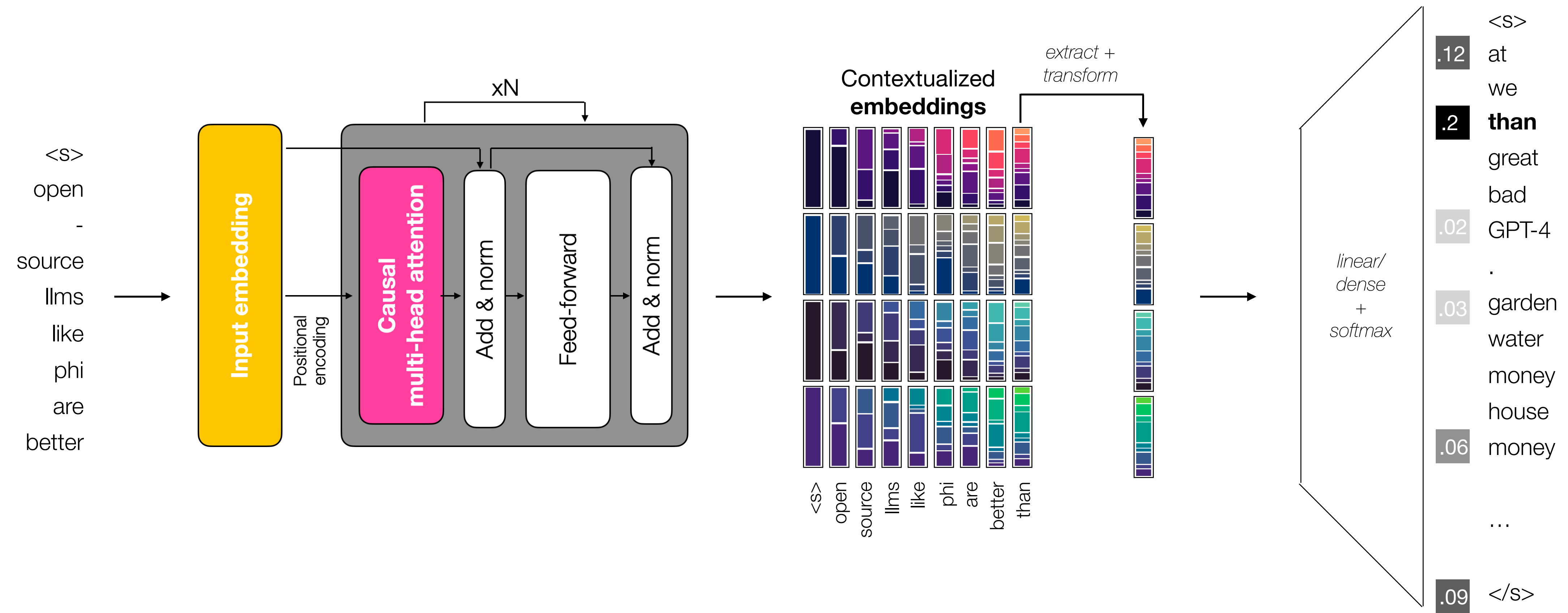
# Transformer

## Causal attention mechanism



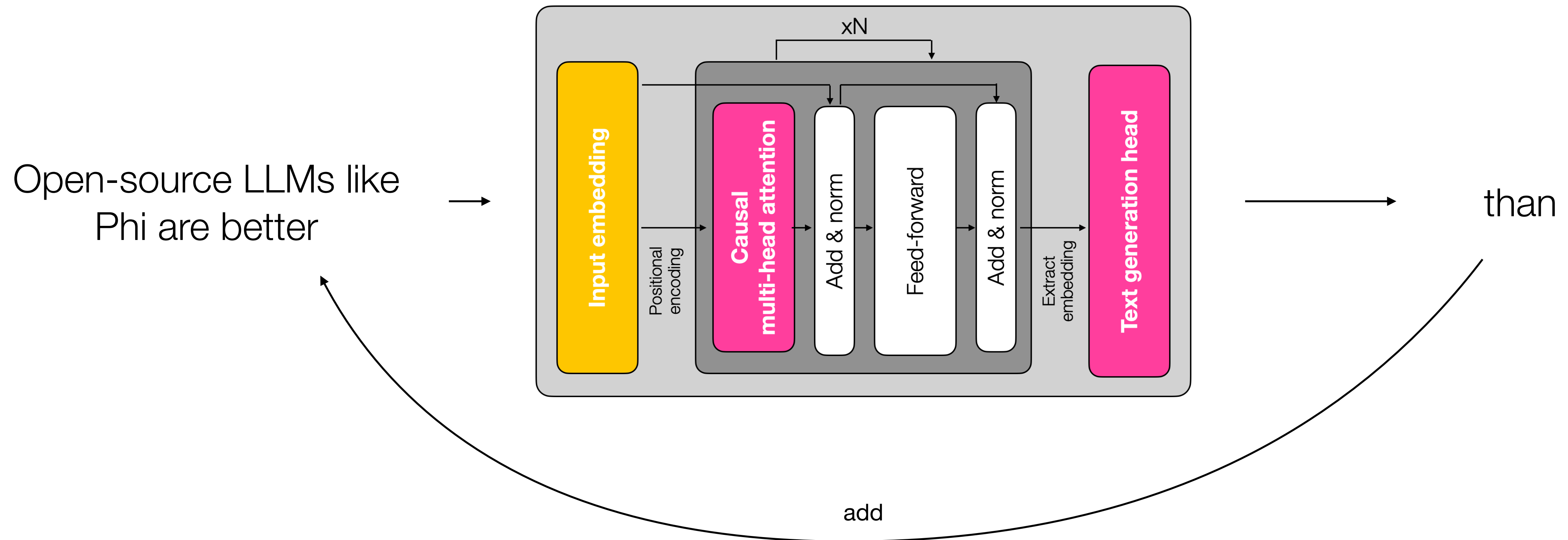
# Transformer

## Model head for causal language modeling



# Text generation

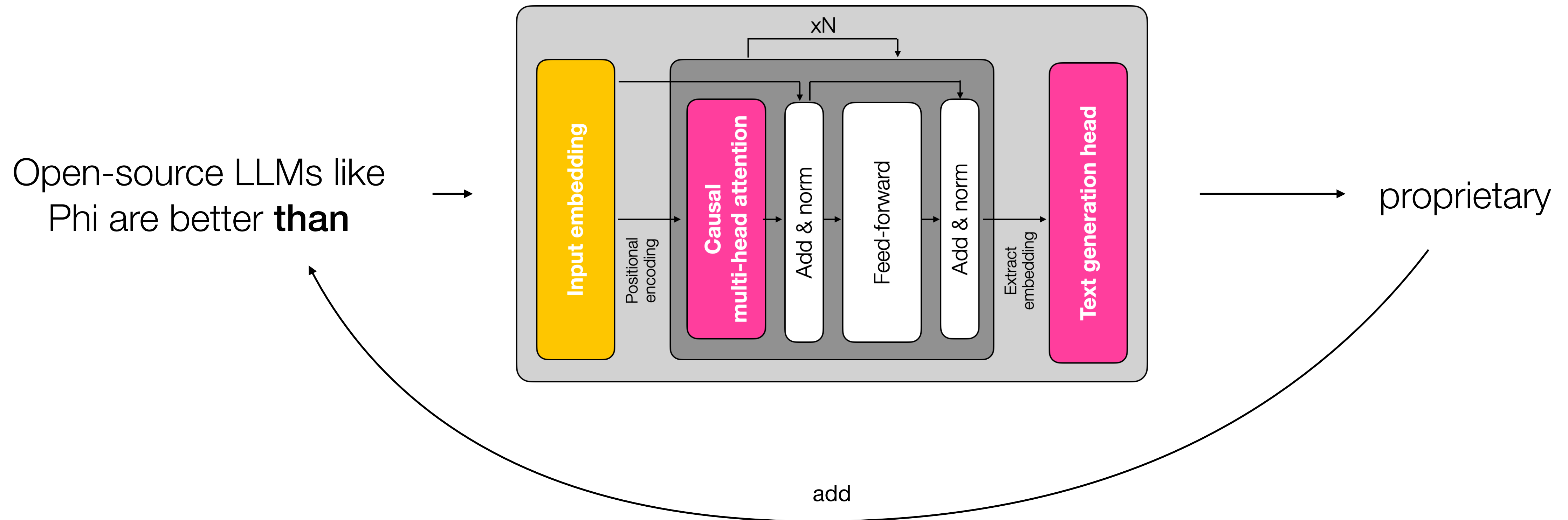
is autoregressive next-token prediction





# Text generation

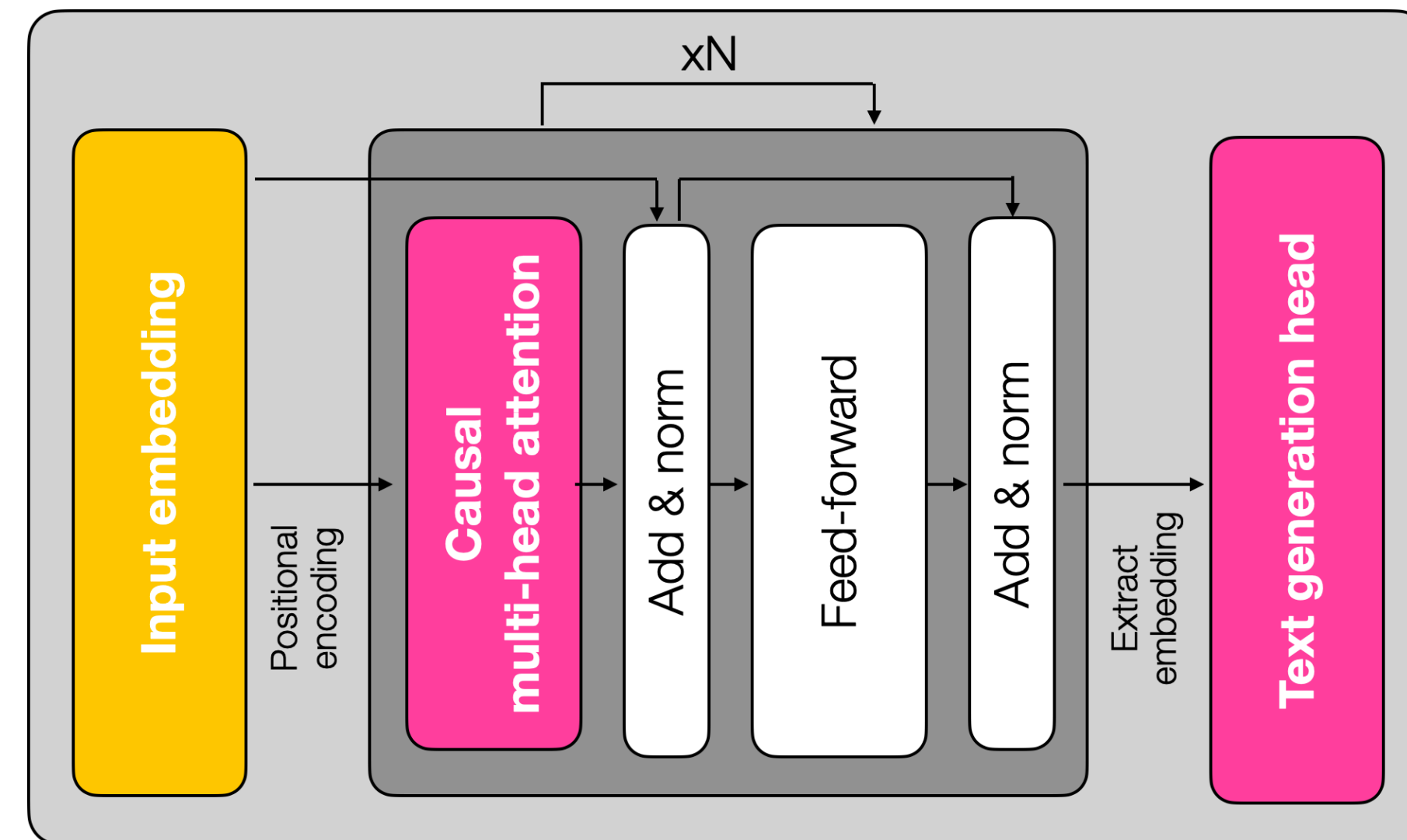
is autoregressive next-token prediction



# Text generation

is autoregressive next-token prediction

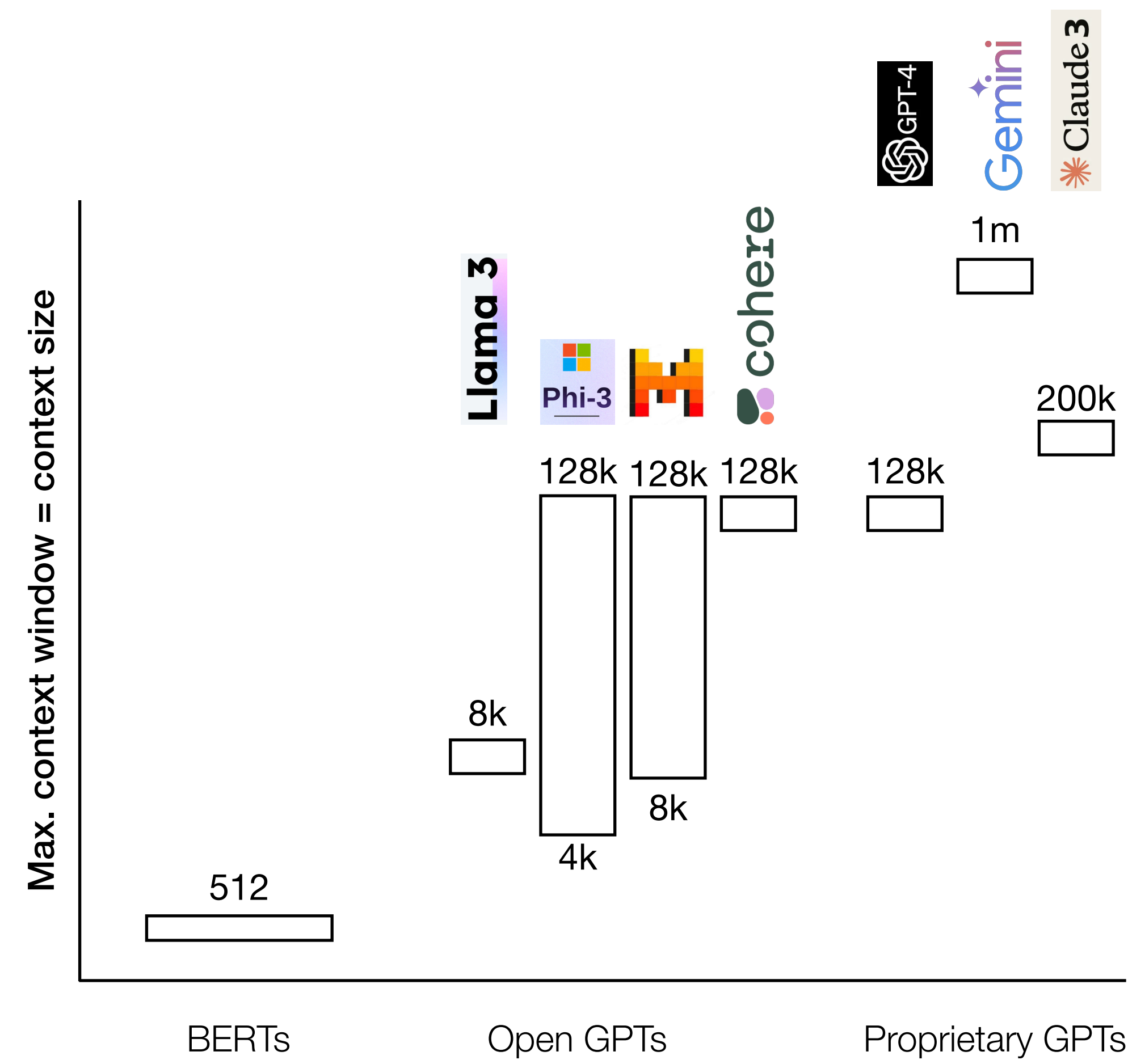
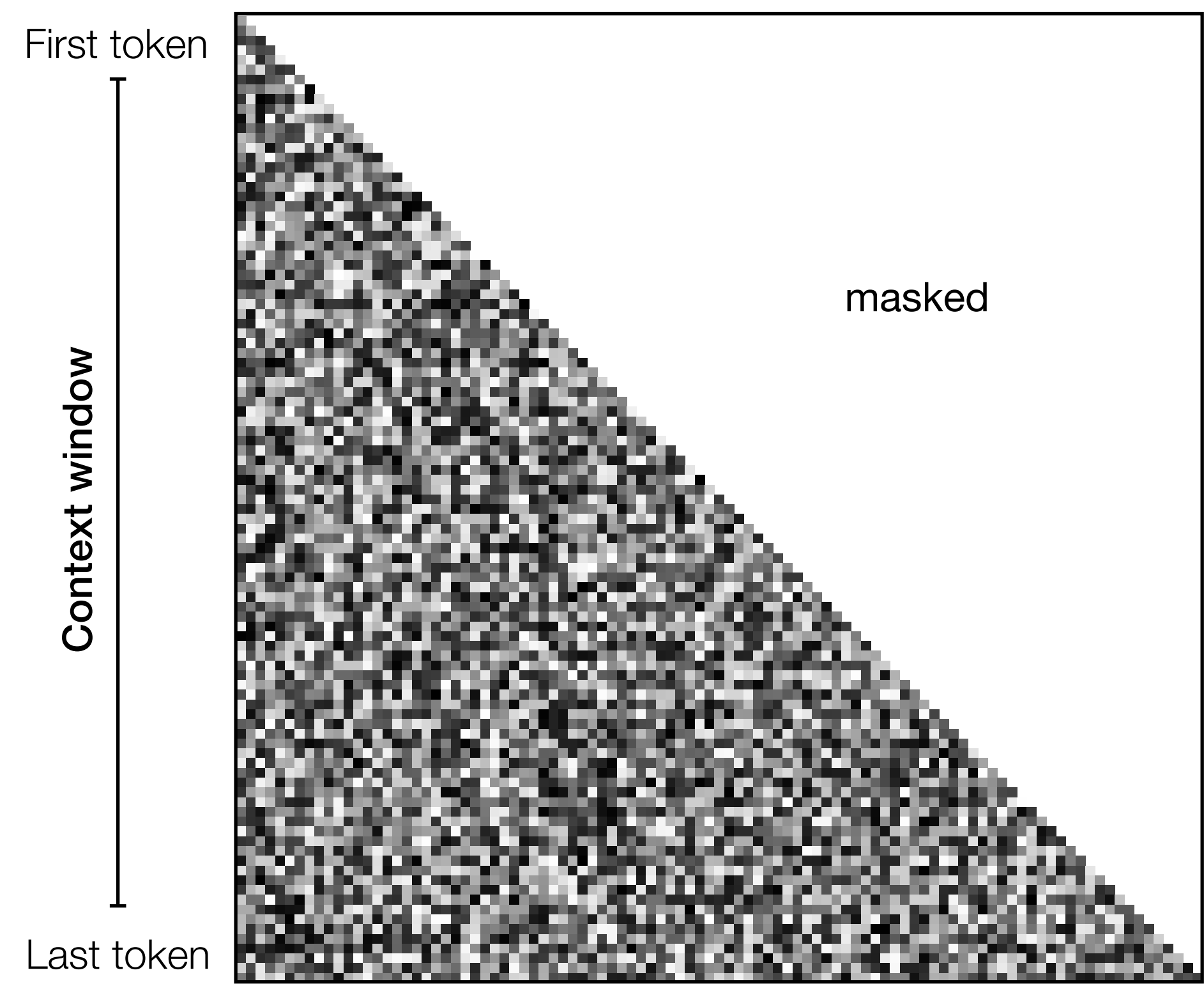
Open-source LLMs like  
Phi are better **than**  
**proprietary**



LLMs

# Context window

or maximum input size



# Prompting

Enabled by long context windows

**Zero-shot**  
classification

excluded

**Few-shot**  
classification

included

## Prompt

System message

You are a helpful assistant.

User message

Task

Your task is to evaluate sentiment of sentences.

Examples

Here are a few examples:

Sentence: text a    Sentiment: 4.2

Sentence: text b    Sentiment: 8.9

Item

Evaluate the sentence: "St. Gallen is a beautiful city"

Instruction

Return a number between 0 and 10 and place it between two @.

# Approaches

practical guide

	Feature extraction	Changing model head	Text generation
Training data needs			
Computational needs			
Biases			
Performance			
Recommendation			



# Exercise

## Predicting media bias

Author	text	bias	type	audience
John Fleming (Representative from Louisiana)	"Looking like Washington may have had more to do with #IRS mis...	part...	atta...	national
Darrell Issa (Representative from California)	"This guy ... #CA49 http://t.co/ilntQltMX0"	neut...	other	national
Michael Crapo (Senator from Idaho)	"The #FarmBill includes language, I authored that states the #...	part...	poli...	national
Deb Fischer (Senator from Nebraska)	"Speaking on the Senate floor on my amendment w/ my colleagues...	part...	info...	constit...
John McCain (Senator from Arizona)	"Honored to meet father & son veterans SSgt Gregory Juedes...	neut...	supp...	national
Bruce Braley (Representative from Iowa)	"Randy Black takes break from decorating the White House to jo...	part...	pers...	constit...
Dianne Feinstein (Senator from California)	"I talked with @MitchellReports this afternoon about President...	part...	media	national
Barbara Mikulski (Senator from Maryland)	"Enough is enough. Standing w patients & families to call ...	part...	poli...	national
Rand Paul (Senator from Kentucky)	"\ "I dont care if it is a Dem. or Repub. President, this is ab...	neut...	media	national
JosÏ© Serrano (Representative from New York)	"Today at 10:30 ribbon cutting ceremony at @soundviewpark's ne...	neut...	info...	national
Debbie Wasserman Schultz (Representative from Florida)	"One year later and all the Republican Party has gotten is one...	part...	atta...	national
Benjamin Cardin (Senator from Maryland)	"#Navalny conviction takes Russia back to the days of USSR usi...	part...	pers...	national
Kevin McCarthy (Representative from California)	"What are the President\u0089\u00a0's real priorities? Put American...	part...	atta...	national
Tom Price (Representative from Georgia)	"MUST READ from @marcthiessen -- \"Kidnapped Libyan prime mini...	part...	poli...	national
Kerry Bentivolio (Representative from Michigan)	"#Obamacare is more proof that we need #ReadtheBills to become...	part...	supp...	national
Mark Pocan (Representative from Wisconsin)	"I'll be on with @WeGotEd at 12:30 CST. Listen live here: http...	neut...	media	national
Ileana Ros-Lehtinen (Representative from Florida)	"Glad @CharlieCrist has options since Nov will be tough. He ca...	neut...	cons...	constit...
Jan Schakowsky (Representative from Illinois)	".@easynan2 its interesting how we will spend \$ but GOP won't ...	part...	atta...	national
Frederica Wilson (Representative from Florida)	"Watch my remarks on #HRes573, the bipartisan resolution conde...	neut...	poli...	national
Brian Higgins (Representative from New York)	"It's @buffalonite in DC! Great to be with so many WNYers. htt...	neut...	info...	constit...

# Quiz



**MAX PLANCK INSTITUTE**  
FOR HUMAN DEVELOPMENT



# Quiz

What is labeling?

What are three ways to do labeling?

What is causal language modeling?

What is parameter-efficient fine-tuning?

What are aspect along which to choose a labeling approach?