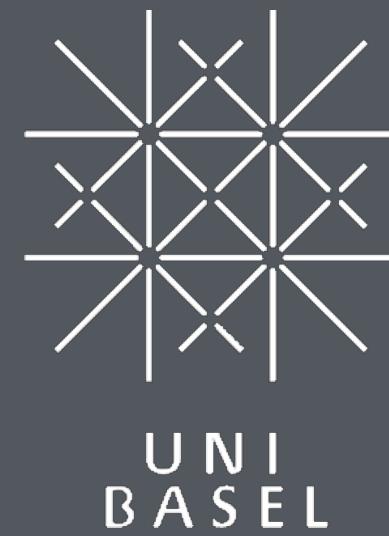


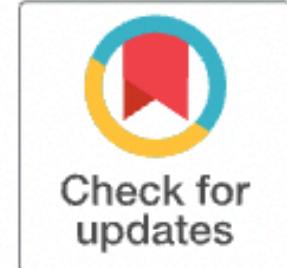
LLMs for classification and regression

Dirk Wulff & Zak Hussain



MAX PLANCK INSTITUTE
FOR HUMAN DEVELOPMENT





ChatGPT outperforms crowd workers for text-annotation tasks

Fabrizio Gilardi^{a,1} , Meysam Alizadeh^a , and Maël Kubli^a 

Edited by Mary Waters, Harvard University, Cambridge, MA; received March 27, 2023; accepted June 2, 2023

Many NLP applications require manual text annotations for a variety of tasks, notably to train classifiers or evaluate the performance of unsupervised models. Depending on the size and degree of complexity, the tasks may be conducted by crowd workers on platforms such as MTurk as well as trained annotators, such as research assistants. Using four samples of tweets and news articles ($n = 6,183$), we show that ChatGPT outperforms crowd workers for several annotation tasks, including relevance, stance, topics, and frame detection. Across the four datasets, the zero-shot accuracy of ChatGPT exceeds that of crowd workers by about 25 percentage points on average, while ChatGPT's intercoder agreement exceeds that of both crowd workers and trained annotators for all tasks. Moreover, the per-annotation cost of ChatGPT is less than \$0.003—about thirty times cheaper than MTurk. These results demonstrate the potential of large language models to drastically increase the efficiency of text classification.

The problem

Classification and regression

Regression

Input
(e.g., sentence)

New York is a
beautiful city



Ouput
(e.g., sentiment)

8.2/10

Classification

Input
(e.g., sentence)

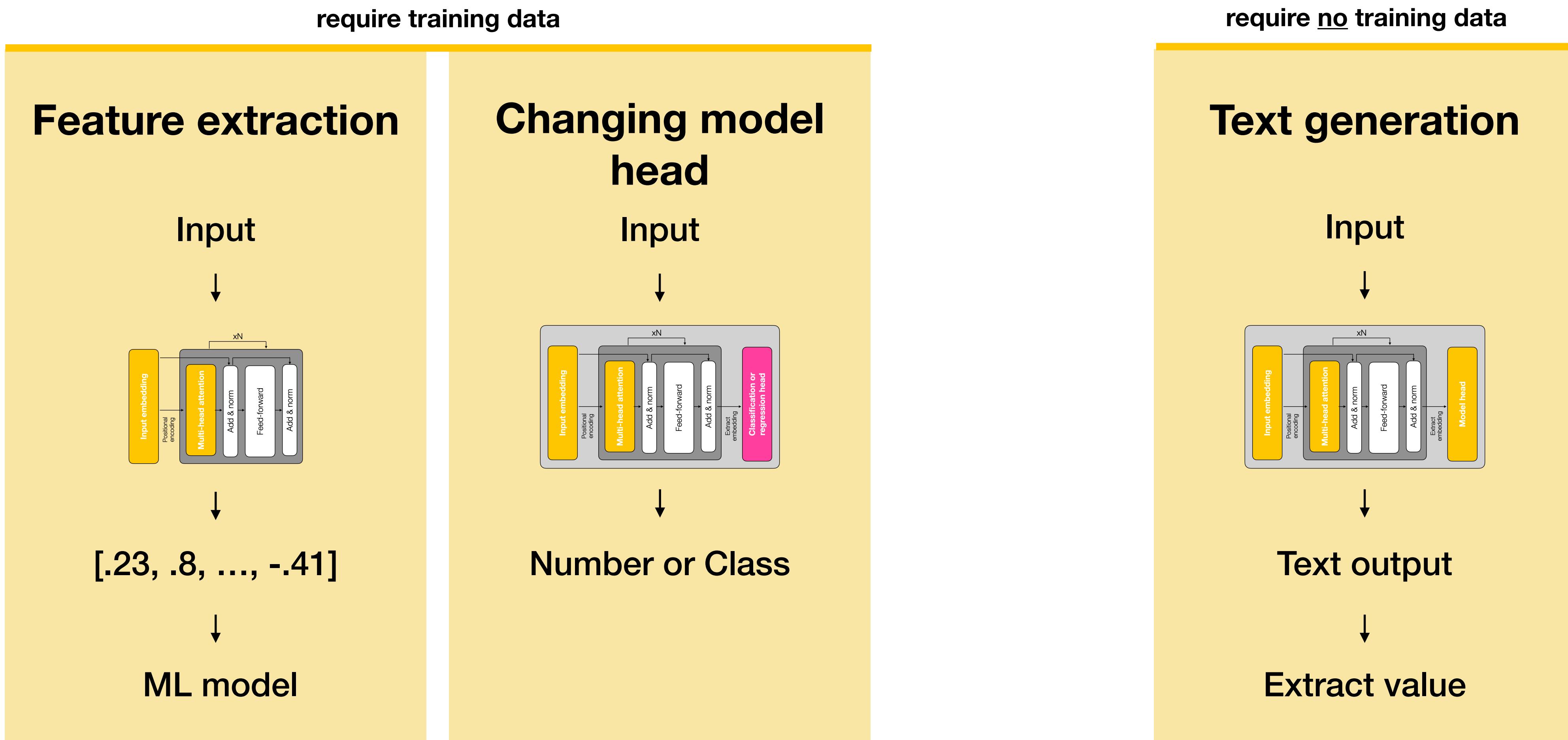
New York is a
beautiful city



Output
(e.g., topic class)

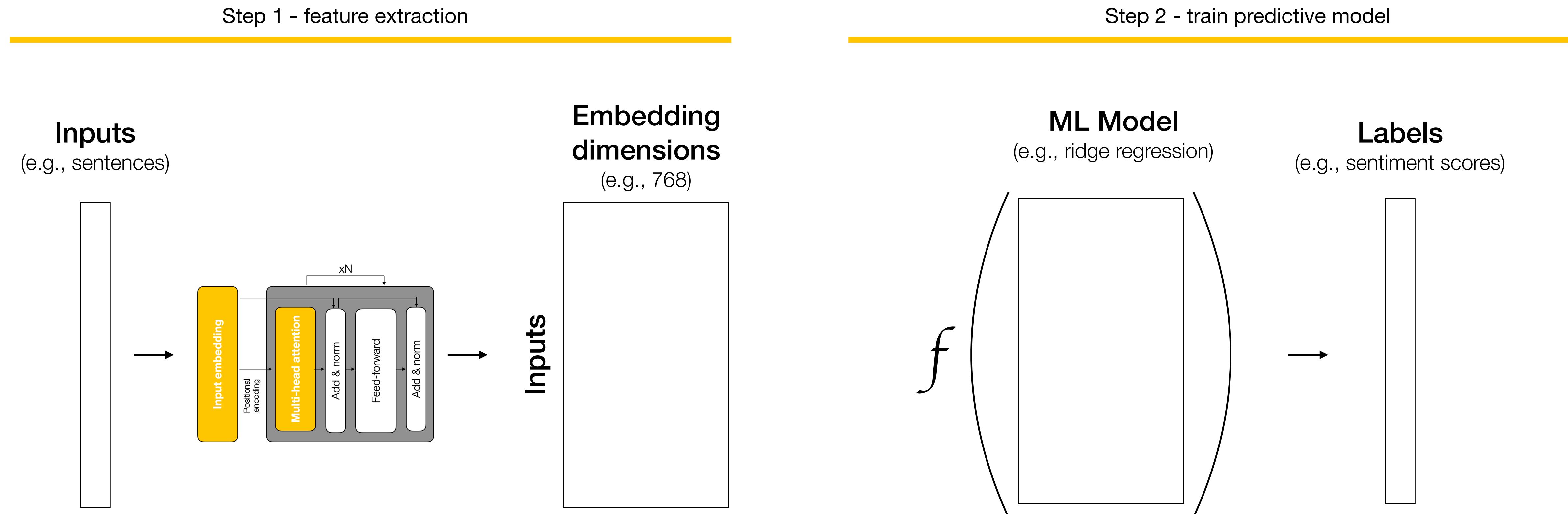
- .82 Tourism
- .1 Sports
- .07 Politics
- .01 Science

Approaches to classification and regression



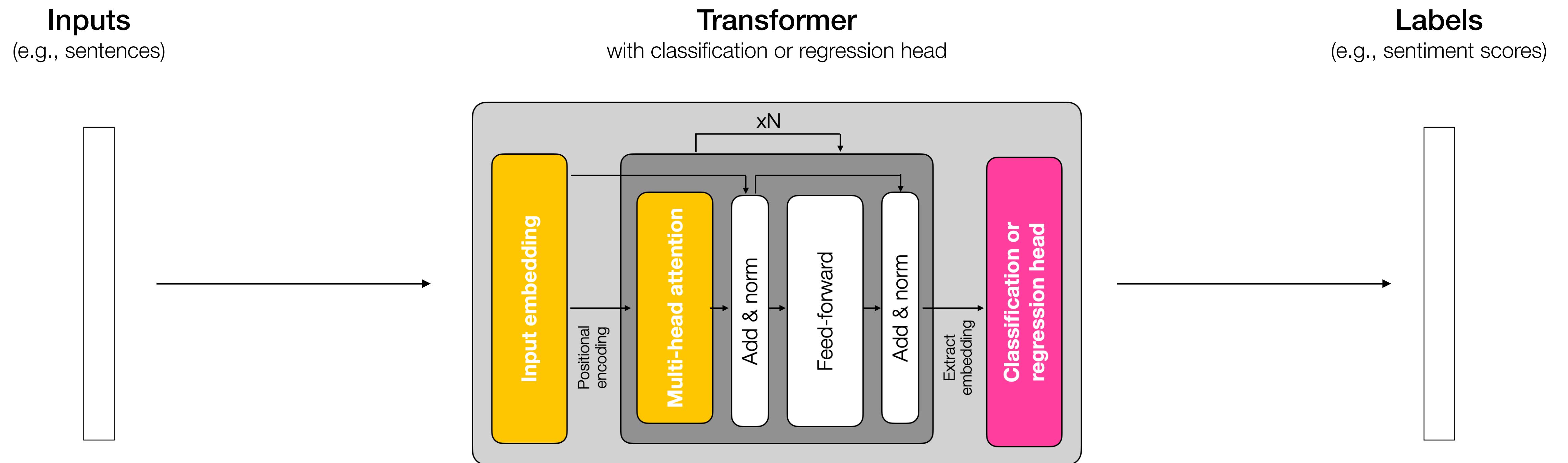
Feature extraction

for regression and classification



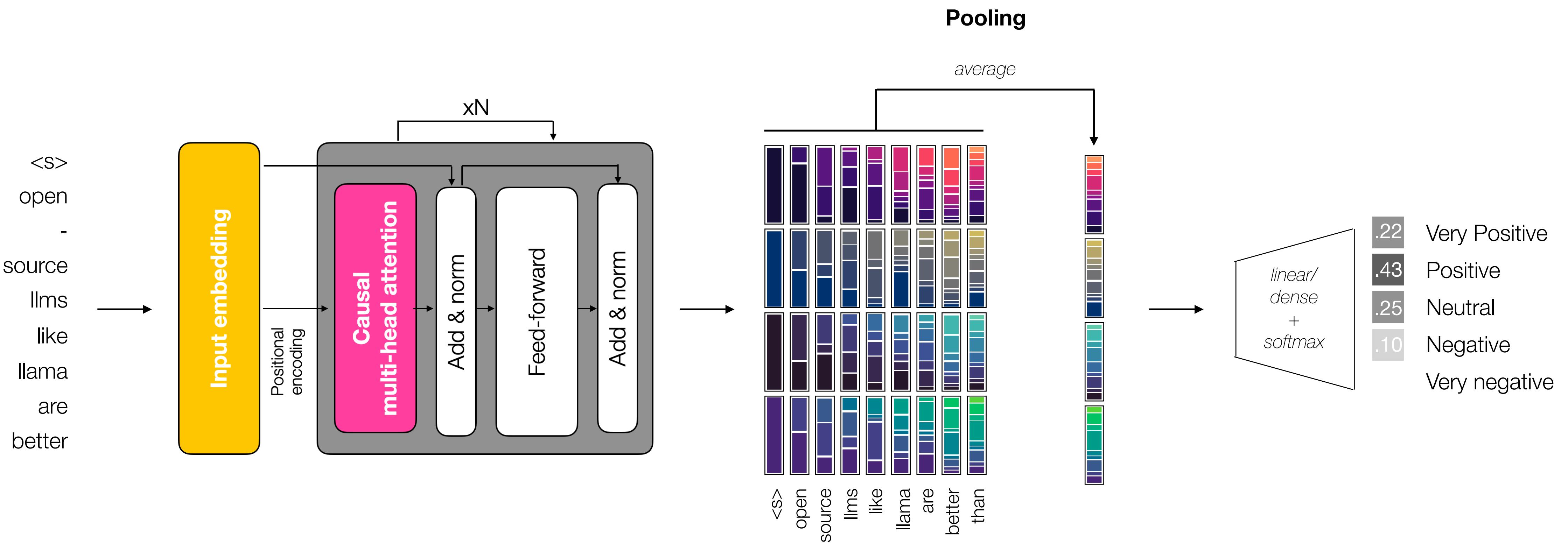
Changing model head

for regression and classification



Transformer

Model head for classification



Text generation

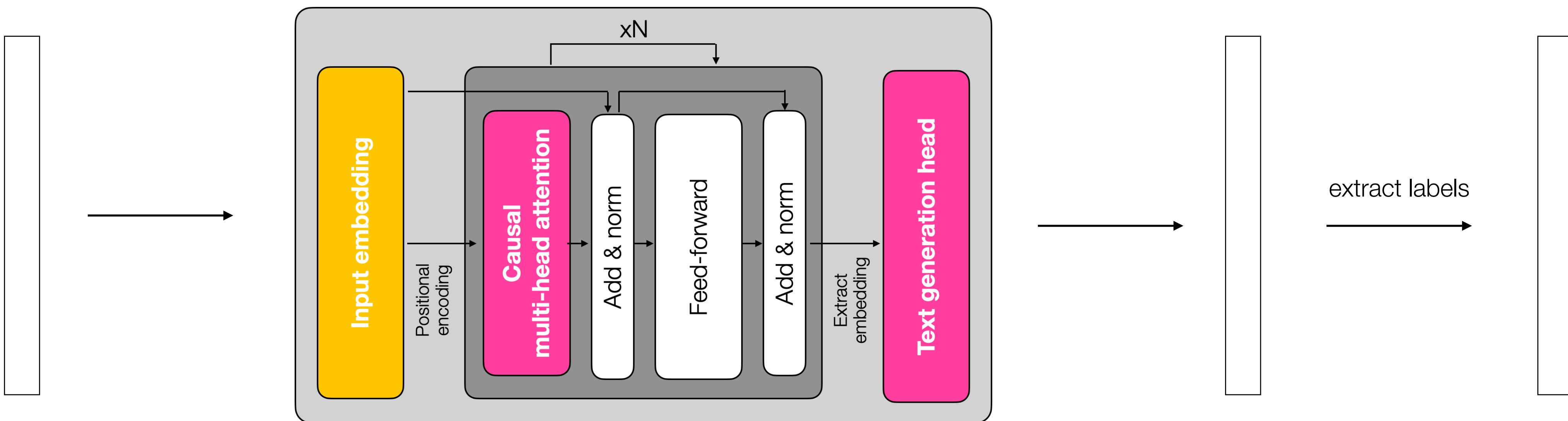
for regression and classification

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:

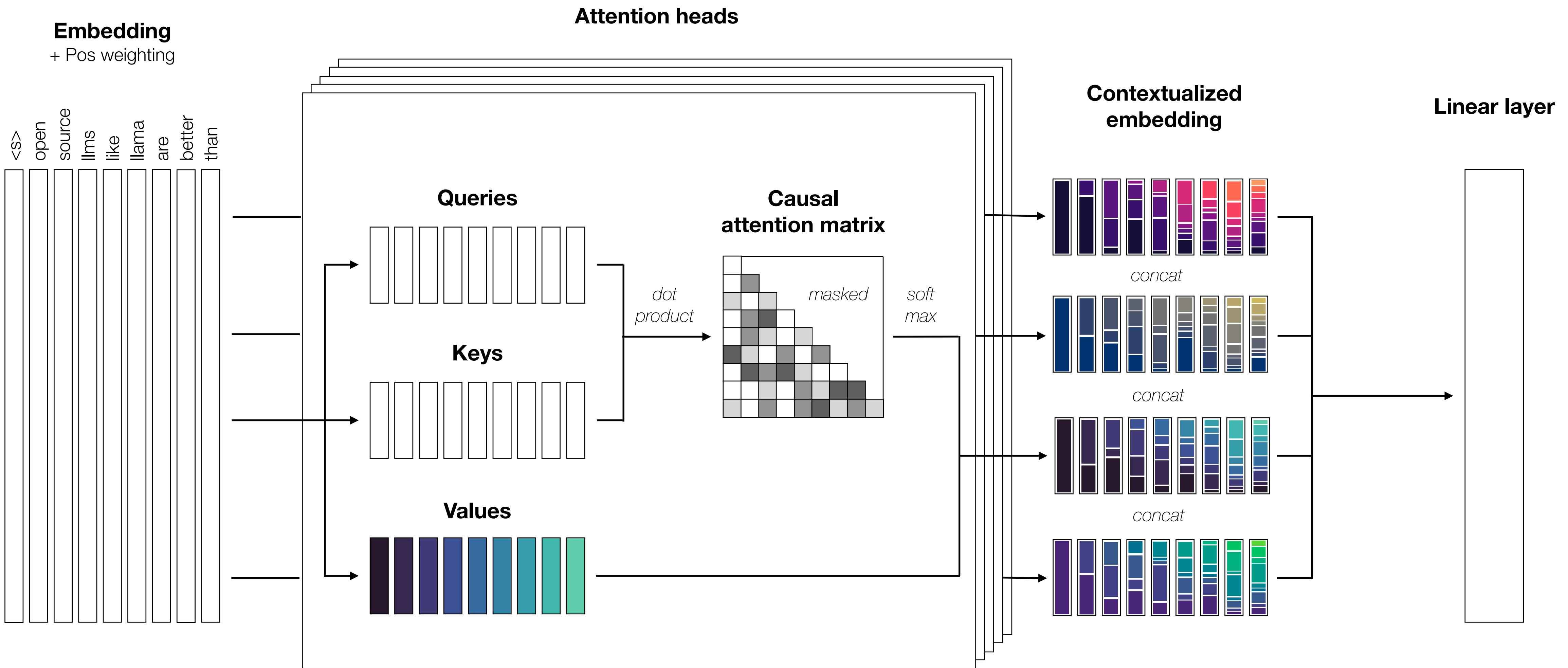
30 to 55

- **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.

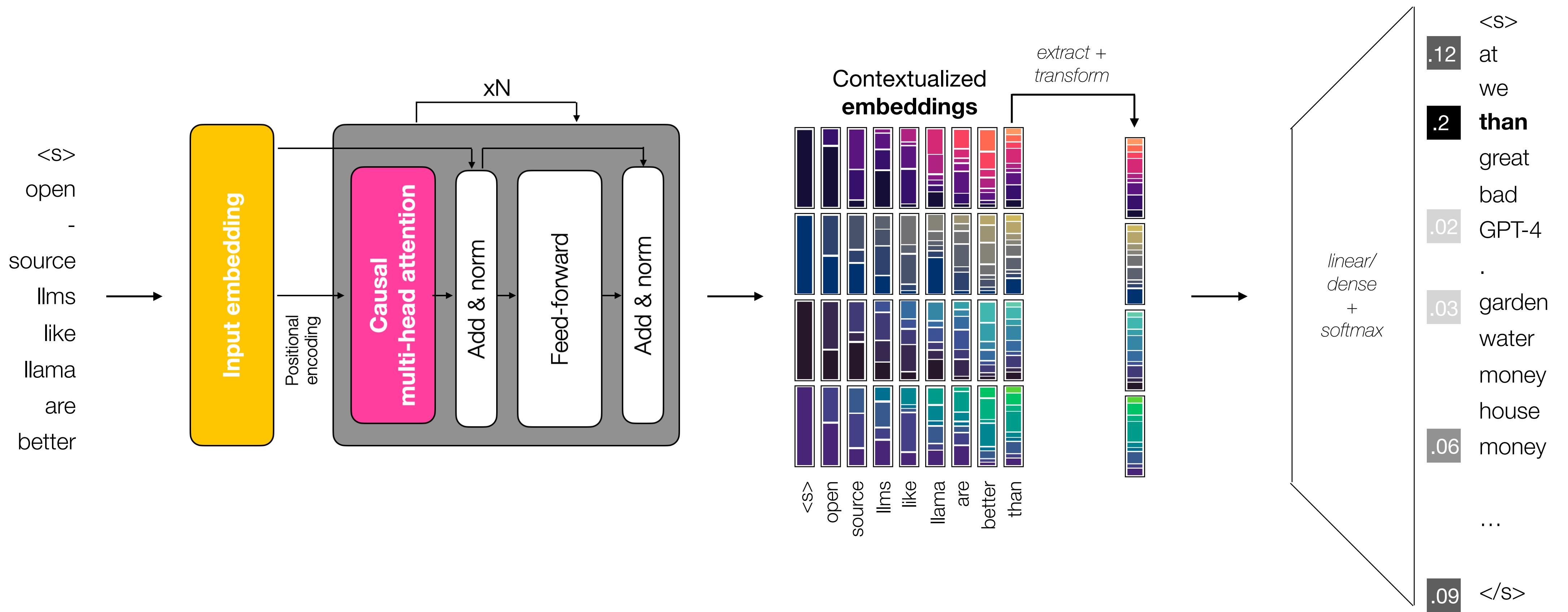
Transformer

Causal attention mechanism



Transformer

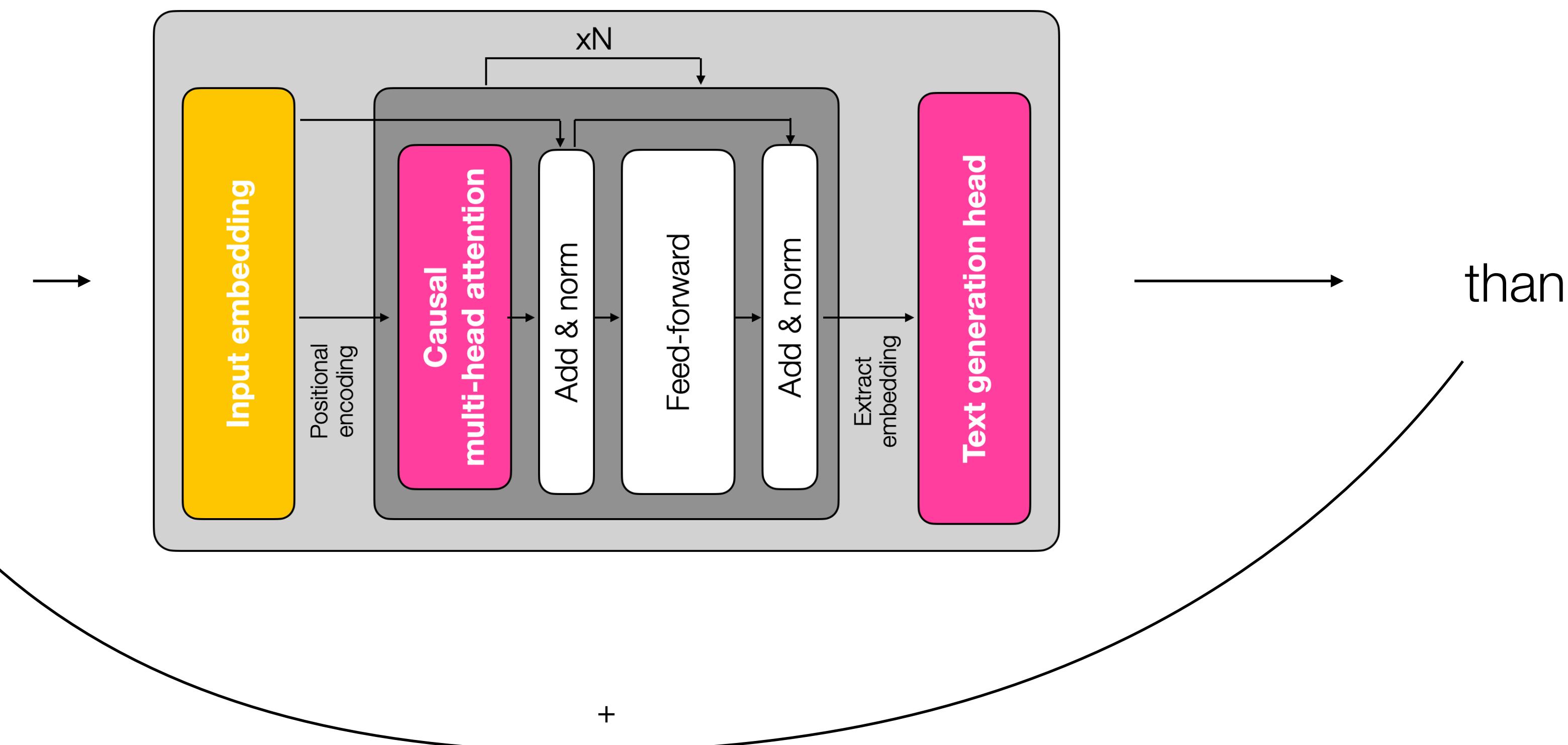
Model head for causal language modeling



Text generation

is autoregressive next-token prediction

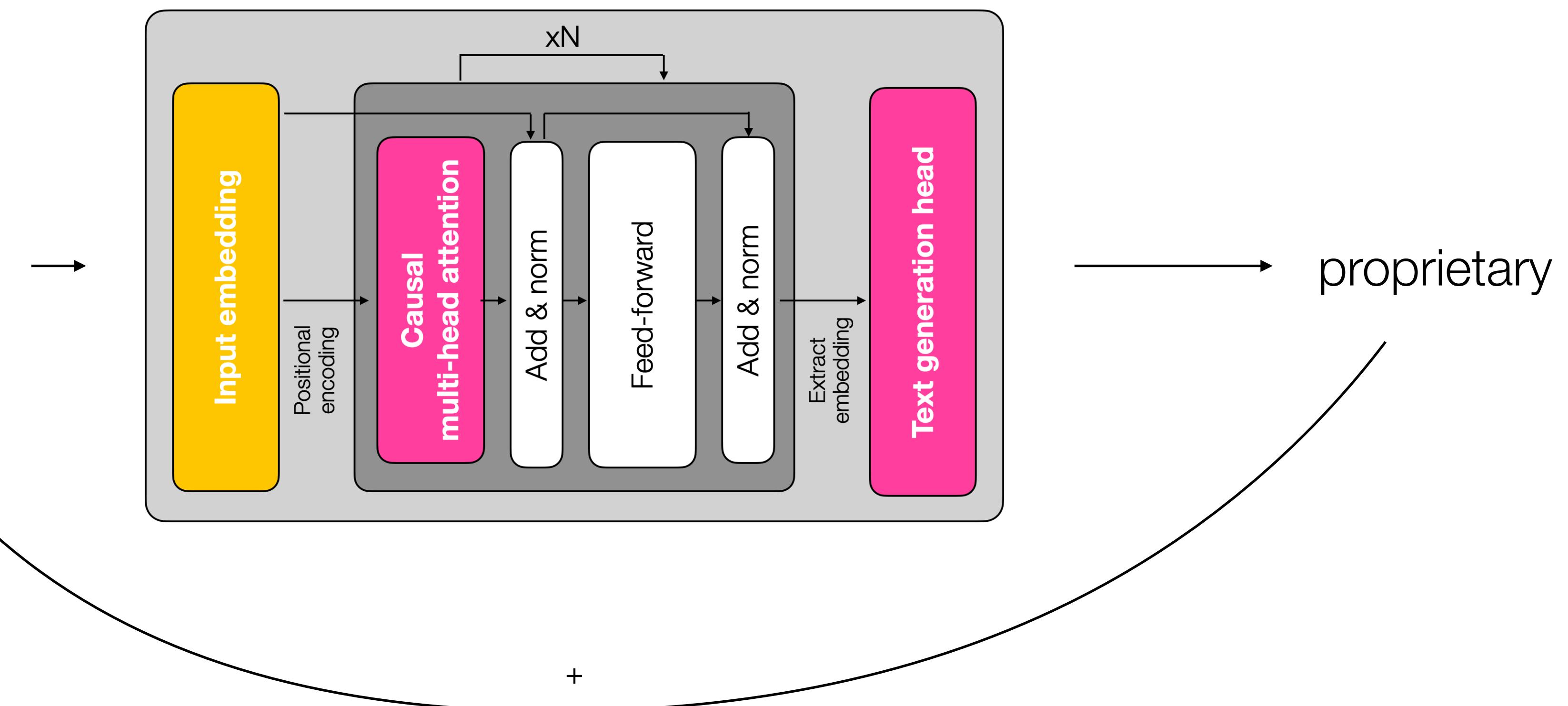
Open-source LLMs like
Llama are better



Text generation

is autoregressive next-token prediction

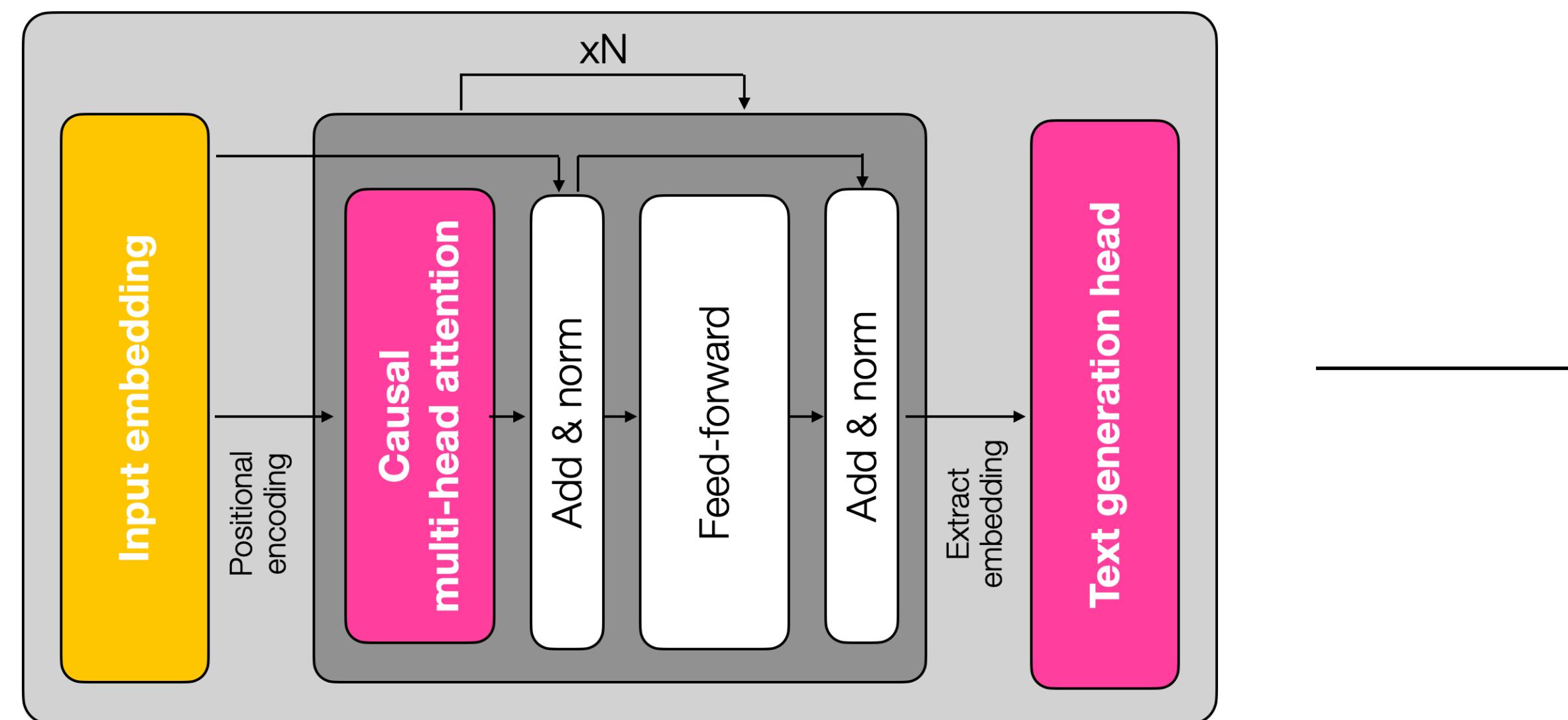
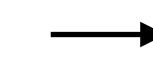
Open-source LLMs like
Llama are better **than**



Text generation

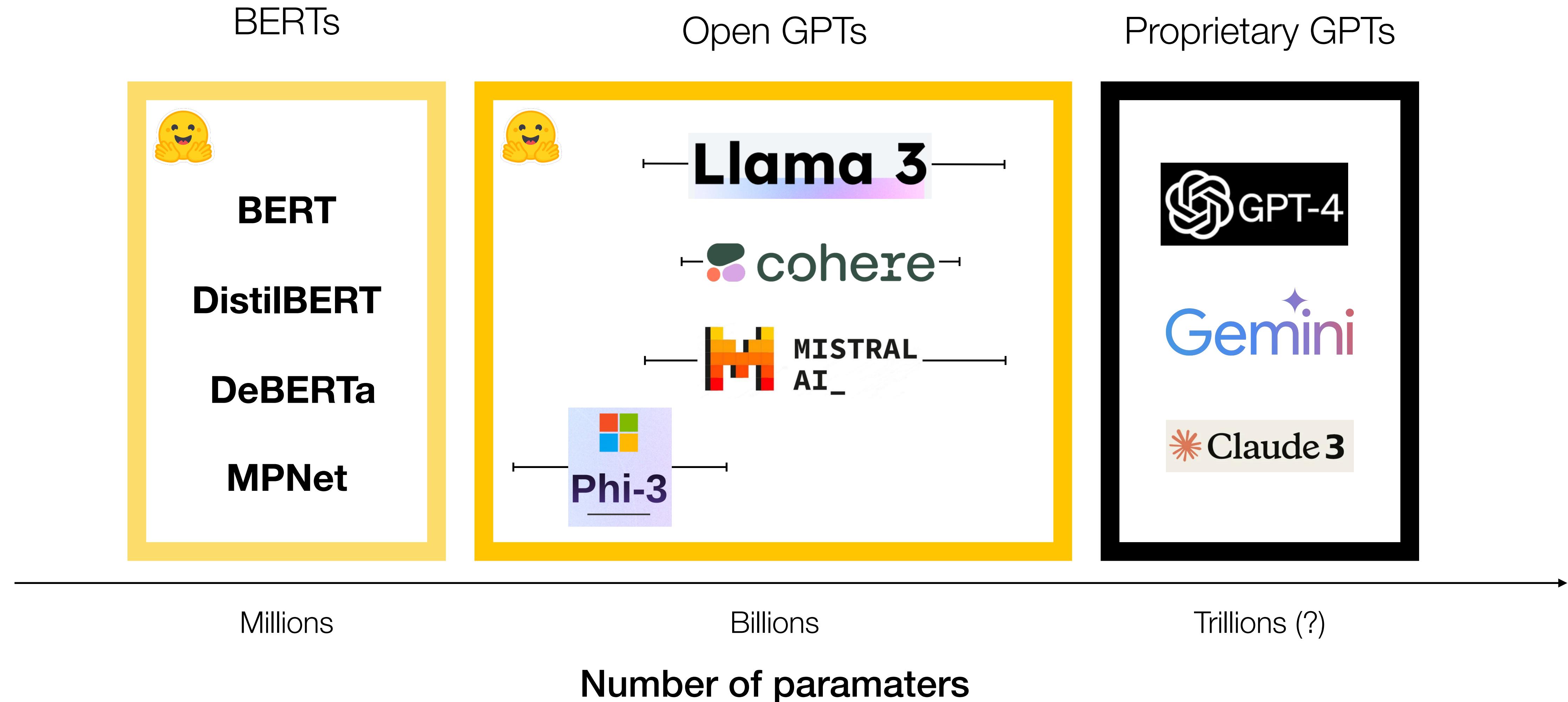
is autoregressive next-token prediction

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



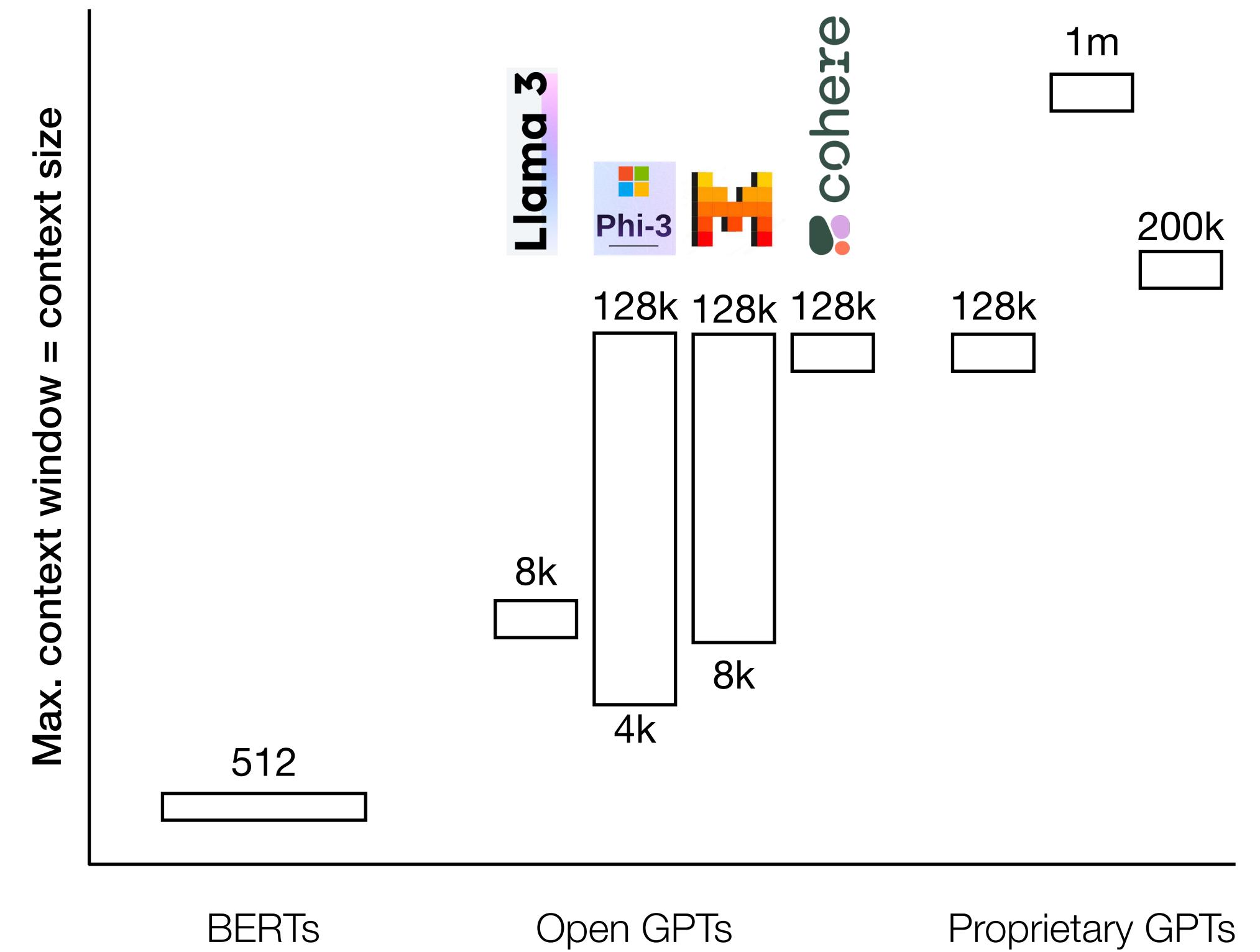
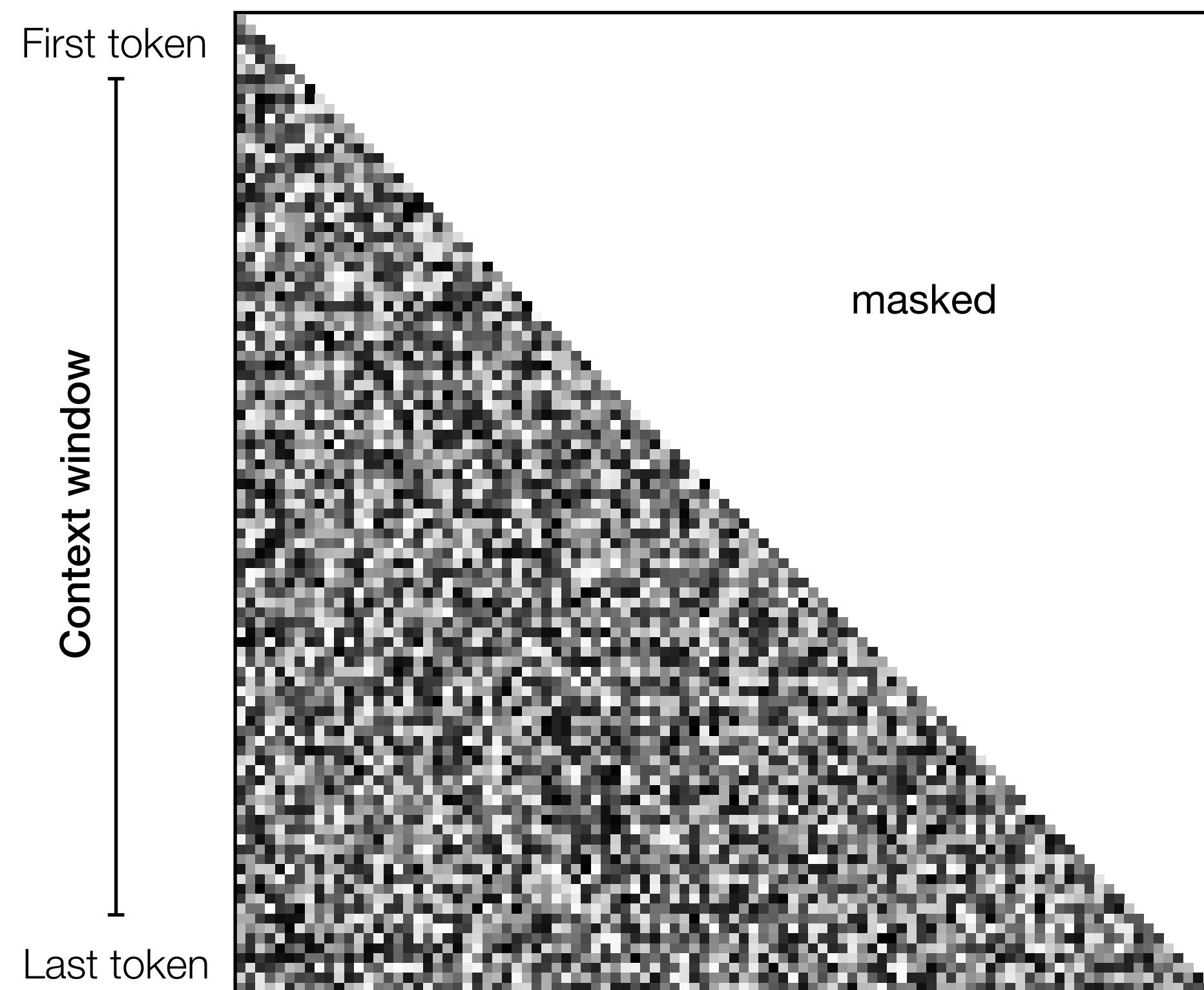
LLMs

Models



Context window

or maximum input size



Prompting

Enabled by long context windows

Zero-shot
classification

Few-shot
classification

excluded

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: "LLMs rock!"

Instruction

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

Approaches

practical guide



Exercise

Predicting media bias

Author

John Fleming (Representative from Louisiana)
Darrell Issa (Representative from California)
Michael Crapo (Senator from Idaho)
Deb Fischer (Senator from Nebraska)
John McCain (Senator from Arizona)
Bruce Braley (Representative from Iowa)
Dianne Feinstein (Senator from California)
Barbara Mikulski (Senator from Maryland)
Rand Paul (Senator from Kentucky)
José Serrano (Representative from New York)
Debbie Wasserman Schultz (Representative from Florida)
Benjamin Cardin (Senator from Maryland)
Kevin McCarthy (Representative from California)
Tom Price (Representative from Georgia)
Kerry Bentivolio (Representative from Michigan)
Mark Pocan (Representative from Wisconsin)
Ileana Ros-Lehtinen (Representative from Florida)
Jan Schakowsky (Representative from Illinois)
Frederica Wilson (Representative from Florida)
Brian Higgins (Representative from New York)

text

"Looking like Washington may have had more to do with #IRS mis... part... atta... national
"This guy ... #CA49 <http://t.co/ilntQltMX0>" neut... other national
"The #FarmBill includes language, I authored that states the #... part... poli... national
"Speaking on the Senate floor on my amendment w/ my colleagues... part... info... constit...
"Honored to meet father & son veterans SSgt Gregory Juedes... neut... supp... national
"Randy Black takes break from decorating the White House to jo... part... pers... constit...
"I talked with @MitchellReports this afternoon about President... part... media national
"Enough is enough. Standing w patients & families to call ... part... poli... national
"\\"I dont care if it is a Dem. or Repub. President, this is ab... neut... media national
"Today at 10:30 ribbon cutting ceremony at @soundviewpark's ne... neut... info... national
"One year later and all the Republican Party has gotten is one... part... atta... national
"#Navalny conviction takes Russia back to the days of USSR usi... part... pers... national
"What are the President\u0089\u00a9s real priorities? Put American... part... atta... national
"MUST READ from @marchiessen -- \\"Kidnapped Libyan prime mini... part... poli... national
"#Obamacare is more proof that we need #ReadtheBills to become... part... supp... national
"I'll be on with @WeGotEd at 12:30 CST. Listen live here: [http...](http://) neut... media national
"Glad @CharlieCrist has options since Nov will be tough. He ca... neut... cons... constit...
. @easynan2 its interesting how we will spend \$ but GOP won't ... part... atta... national
"Watch my remarks on #HRes573, the bipartisan resolution conde... neut... poli... national
"It's @buffalonite in DC! Great to be with so many WNYers. htt... neut... info... constit...

bias
type
audience

Quiz



MAX PLANCK INSTITUTE
FOR HUMAN DEVELOPMENT



Quiz

What is labeling?

What are three ways to do labeling?

What is causal language modeling?

What are aspect along which to choose a labeling approach?