LLMs for classification and regression

Dirk Wulff & Zak Hussain

MAX PLANCK INSTITUTE
FOR HUMAN DEVELOPMENT

UNI
BASEL

PNAS BRIEF REPORT | POLITICAL SCIENCES OPEN ACCESS

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.

ChatGPT | text classification | large language models | human annotations | text as data

The problem

Classification and regression

Regression

Input
(e.g., sentence)

Ouput
(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

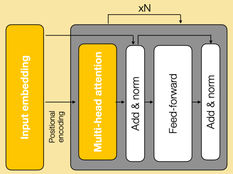
to classification and regression

require training data

Feature extraction

Input

↓



↓

[.23, .8, …, -.41]

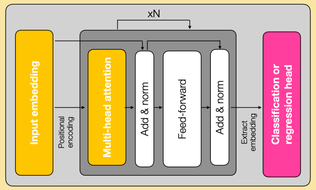
↓

ML model

Changing model head

Input

↓



↓

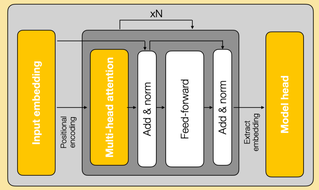
Real or Class

require <u>no</u> training data

Text generation

Input

↓



↓

Text output

↓

Extract value

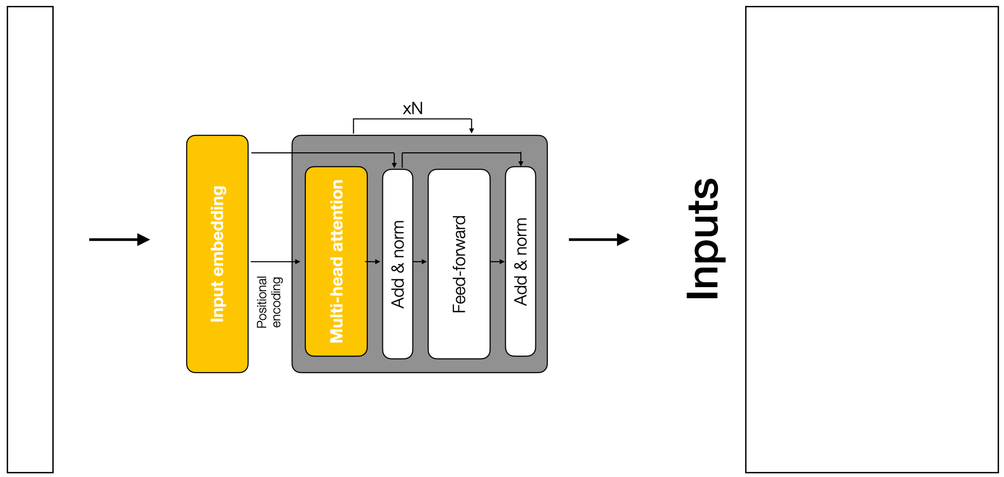
Feature extraction

for regression and classification

Step 1 - feature extraction

Step 2 - train predictive model

Inputs
(e.g., sentences)



Embedding dimensions
(e.g., 768)

Inputs

ML Model
(e.g., ridge regression)

f

Labels
(e.g., sentiment scores)

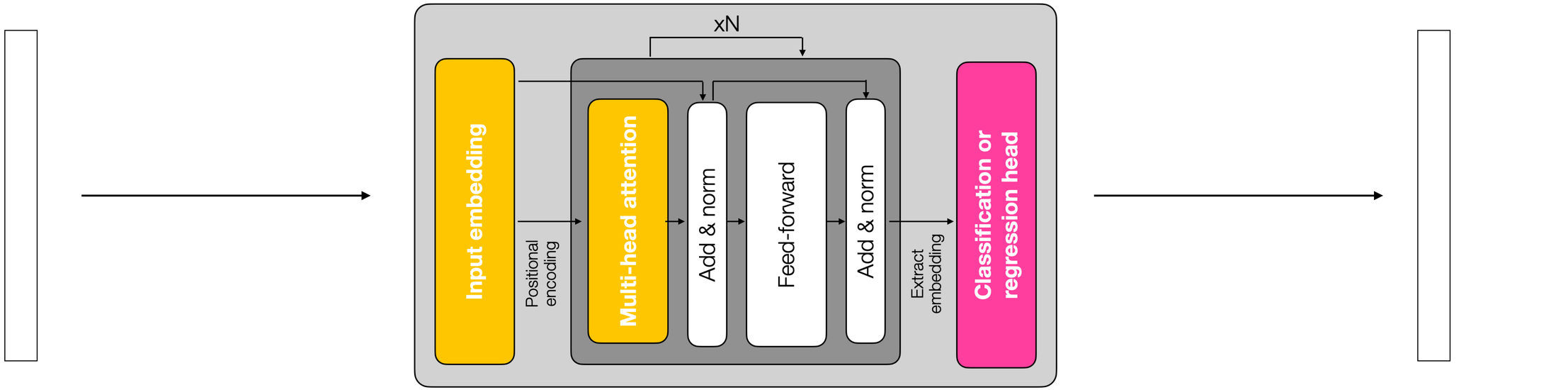
Changing model head

for regression and classification

Inputs
(e.g., sentences)

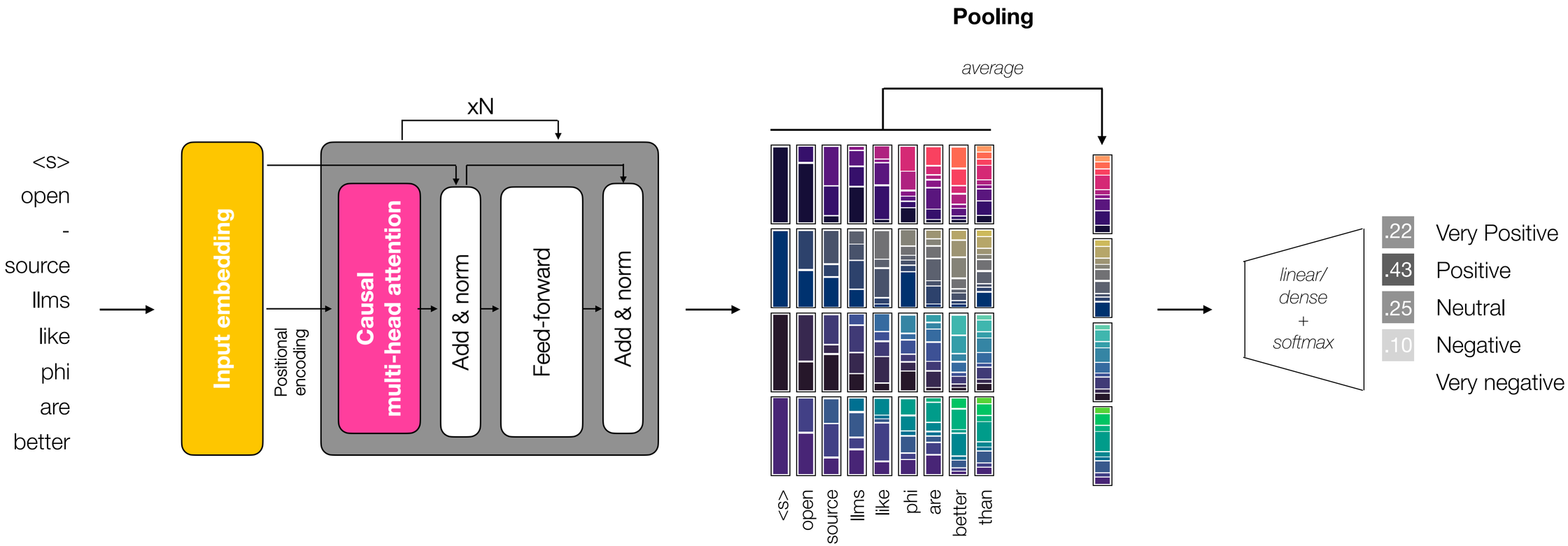
Transformer
with classification or regression head

Labels
(e.g., sentiment scores)



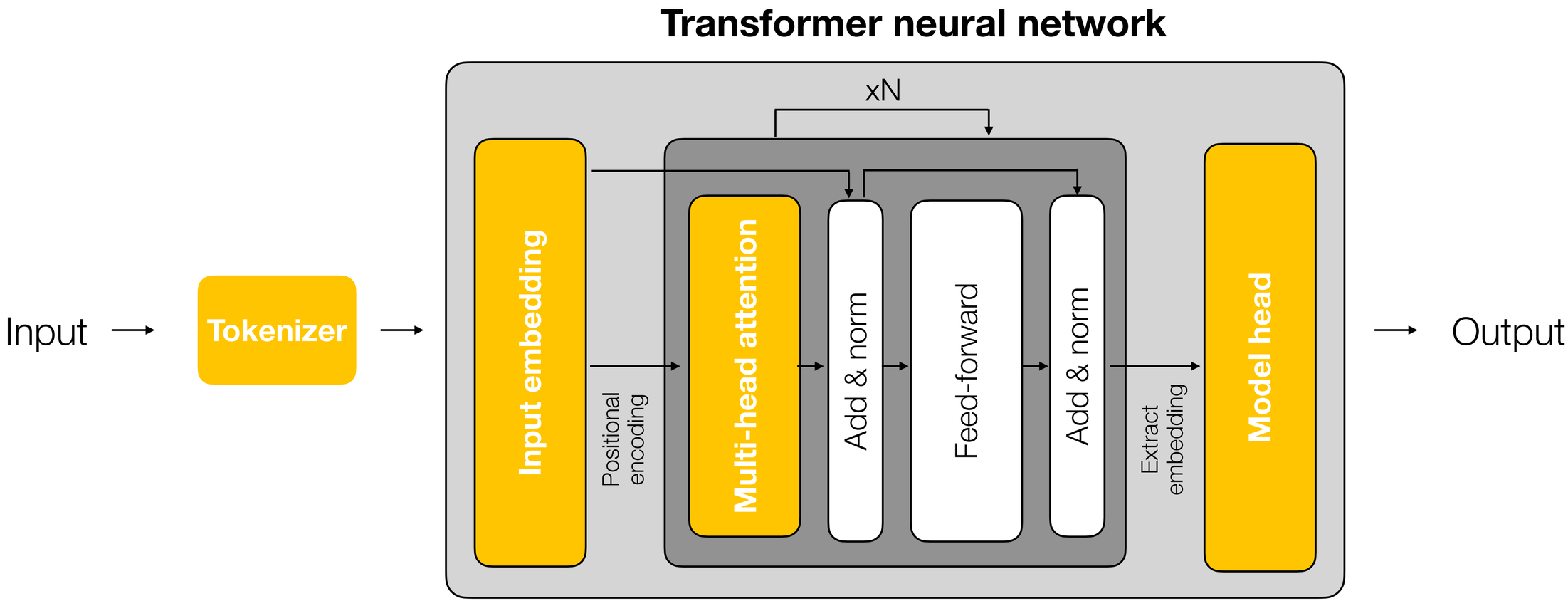
Transformer

Model head for **classification**



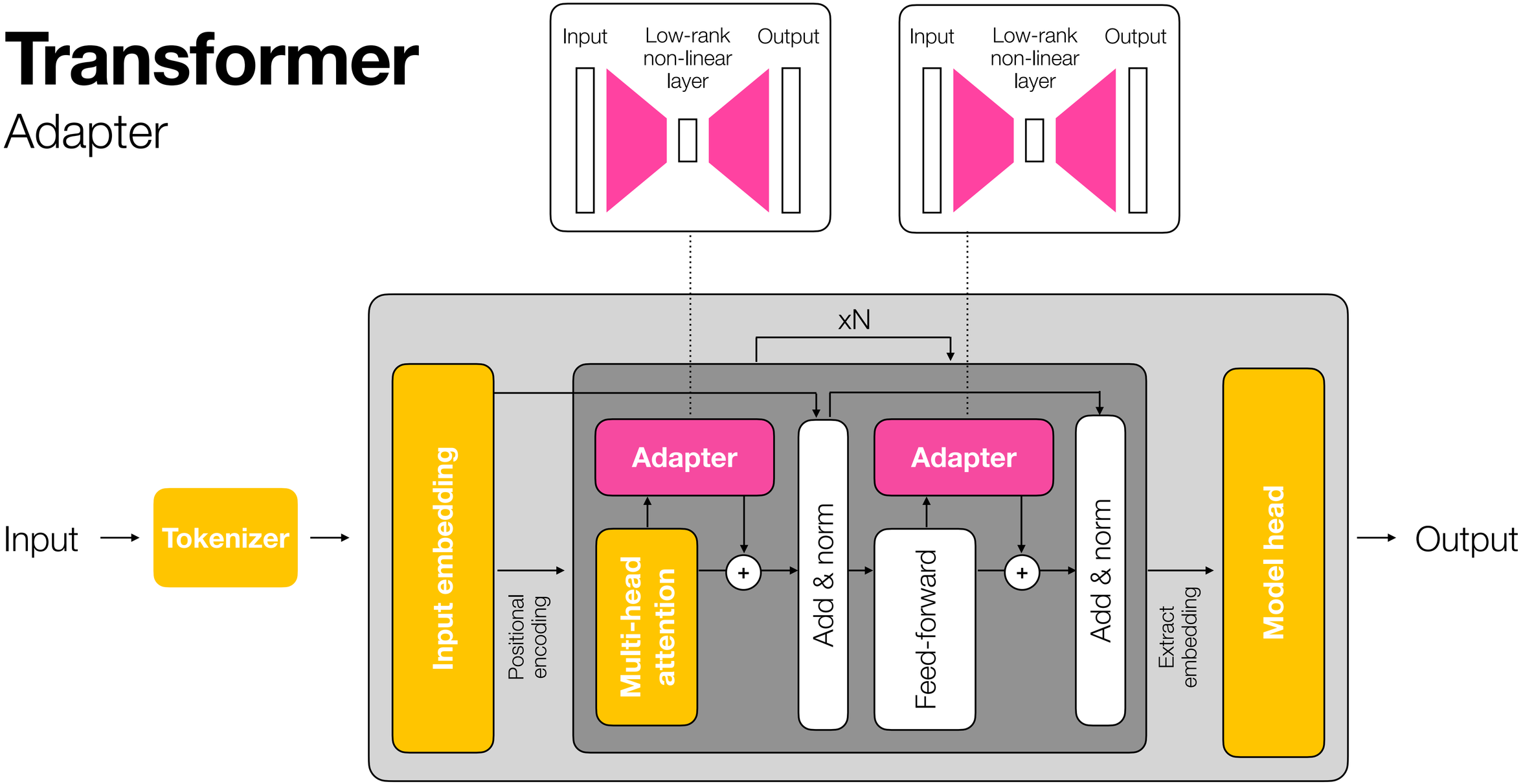
Transformer

Adapter



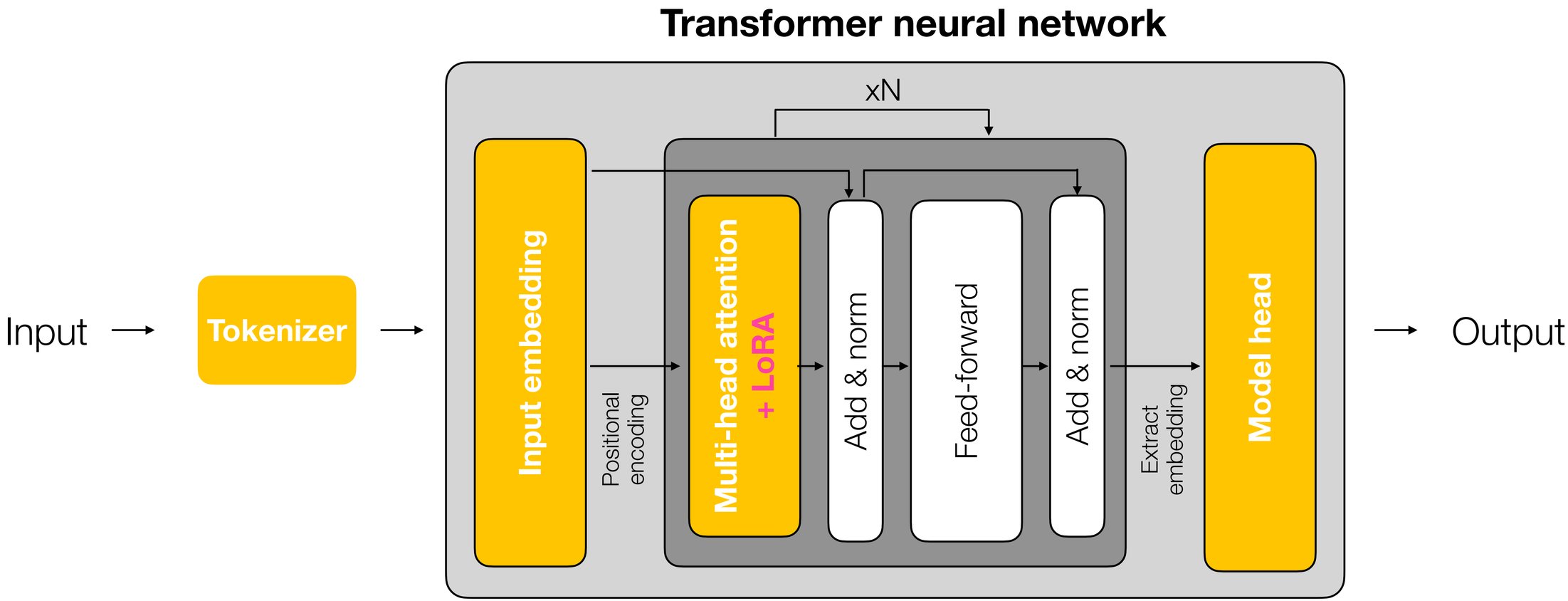
Transformer

Adapter



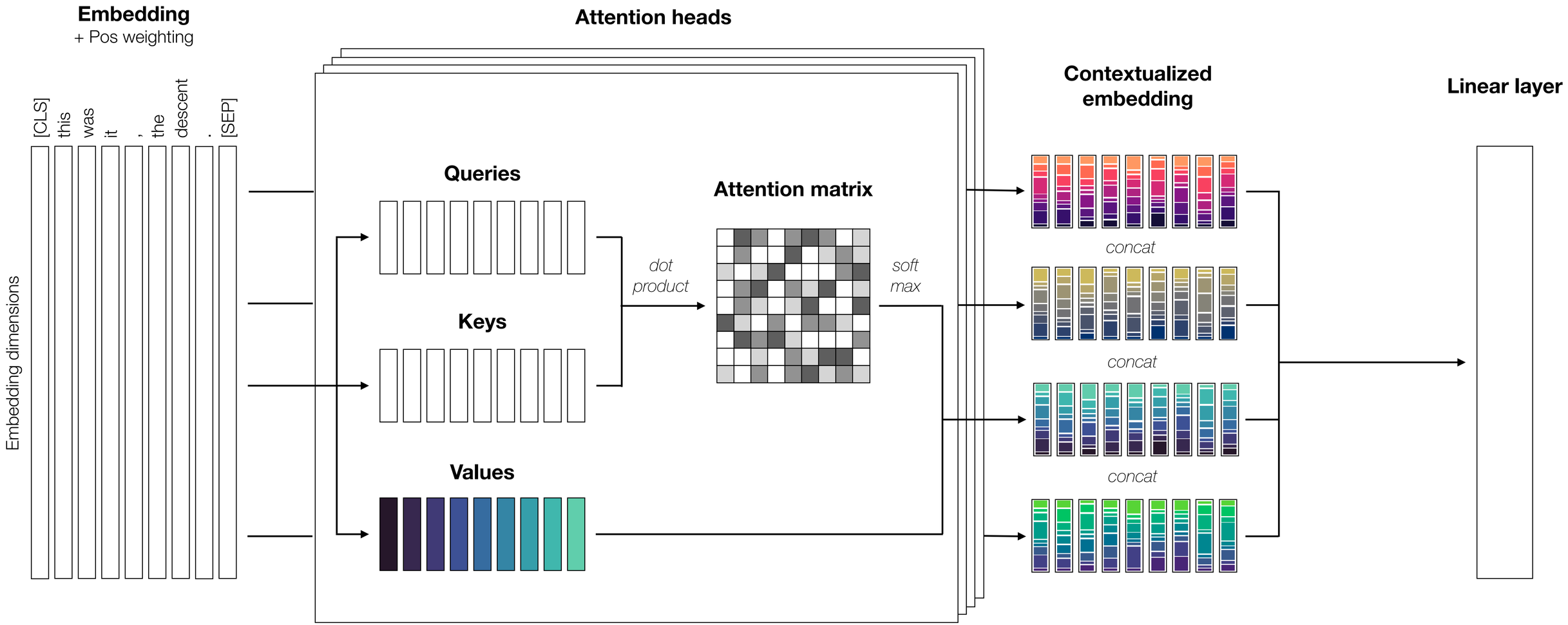
Transformer

LoRA



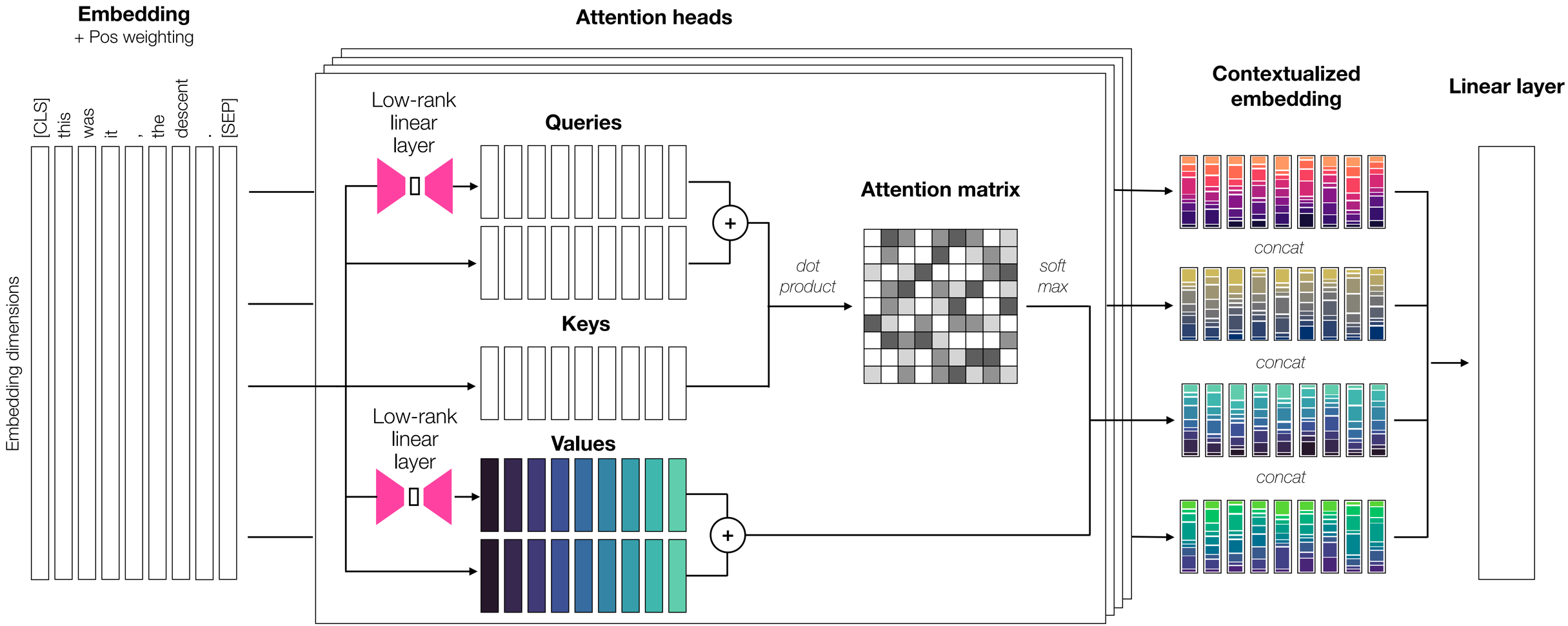
Transformer

Adapter



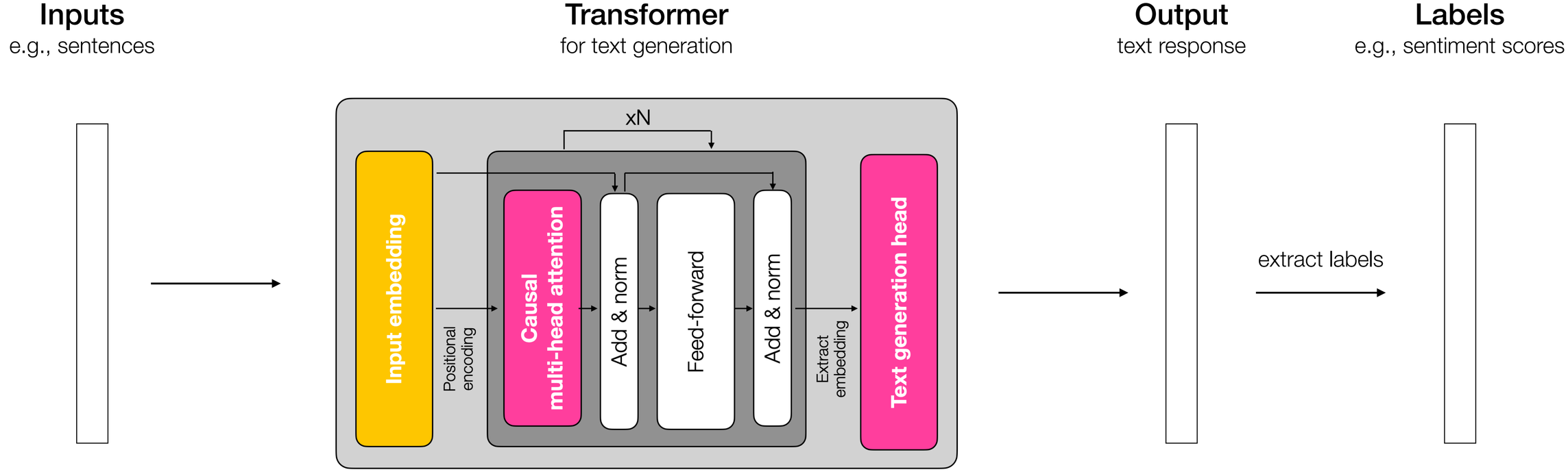
Transformer

Adapter



Text generation

for regression and classification



Example

Extracting numerical estimates

(Can also be improved by better prompting)

On a scale of 0 to 100, how risky is shampoo?

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:

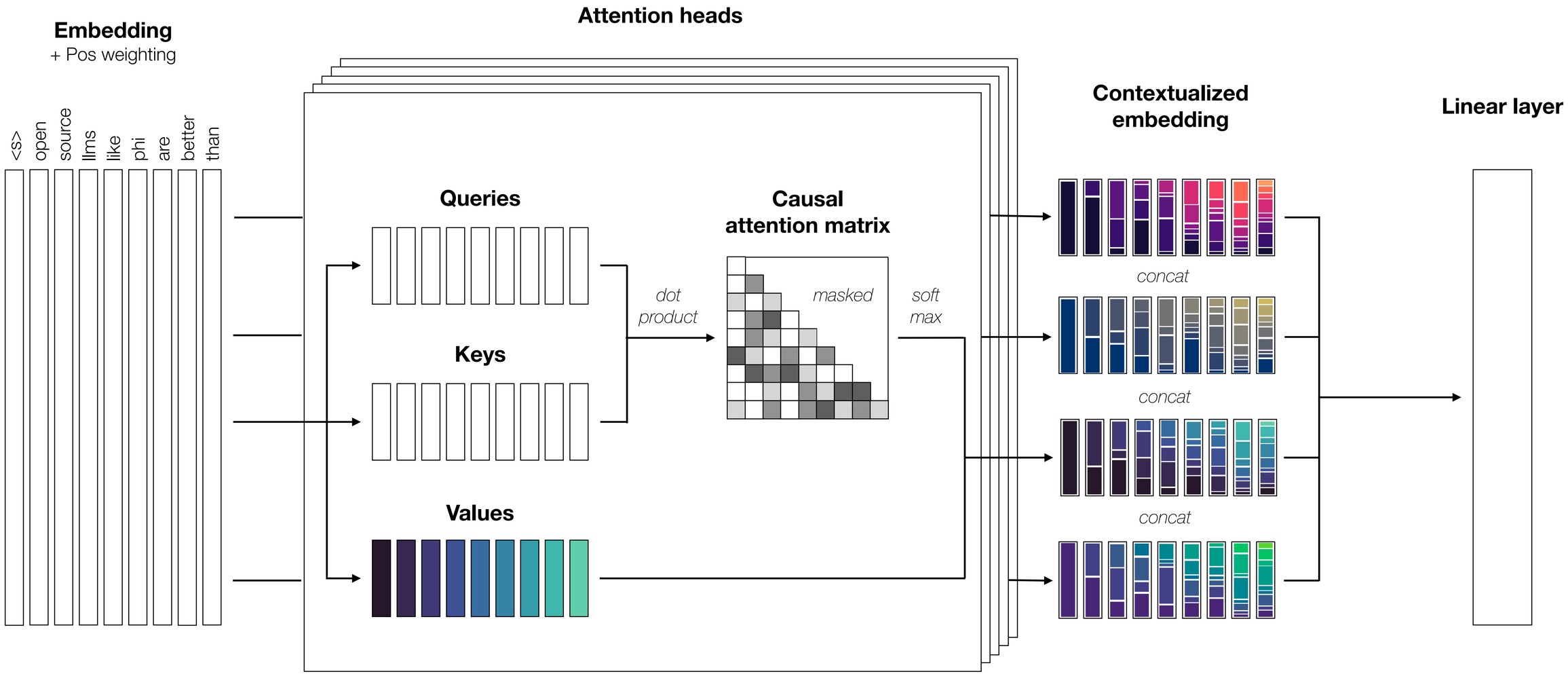
30 to 55

- **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.

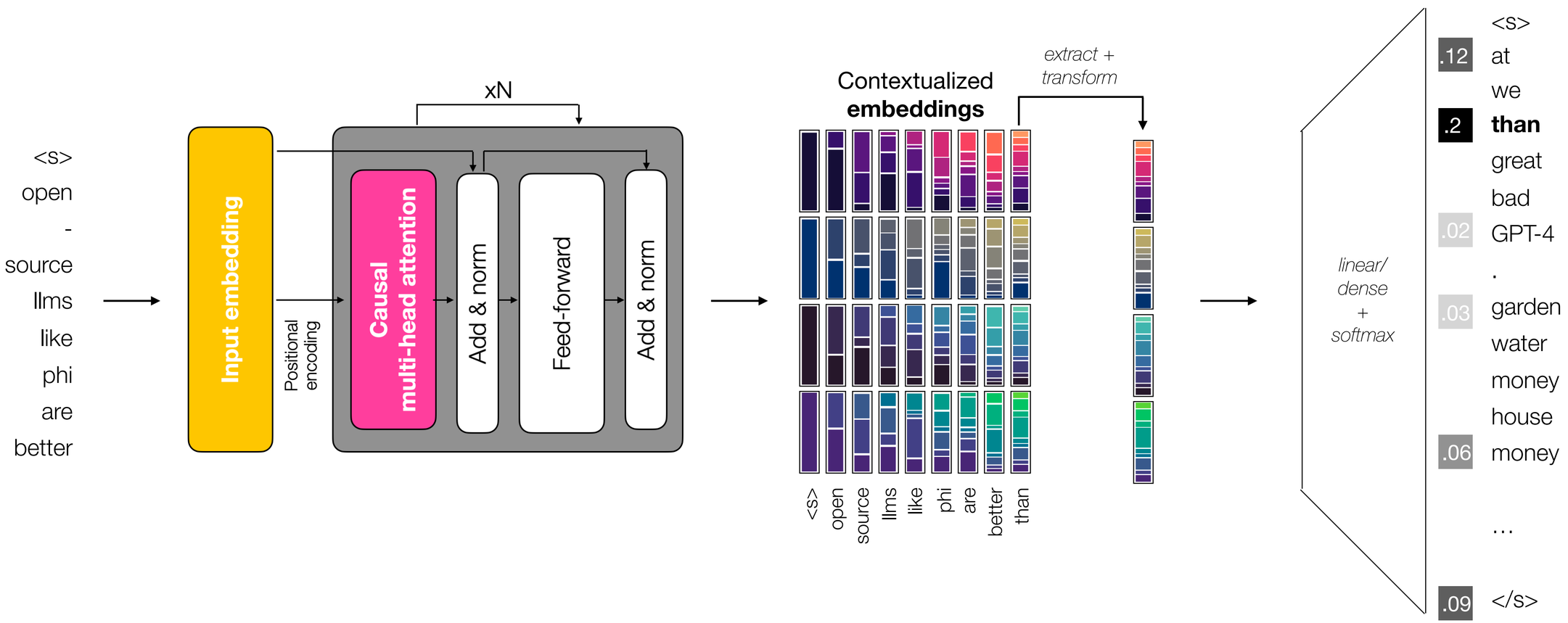
Transformer

Causal attention mechanism



Transformer

Model head for **causal language modeling**



Text generation

Softmax function

Next token
one of the tokens in
the vocabulary
(approx. 30k)

Activation
at token i
can be understood as the
amount of evidence for any token

$$p(token_i) = \frac{e^{\frac{a_i}{T}}}{\sum_j e^{\frac{a_j}{T}}}$$

Temperature
of the softmax controlling the
Randomness/variability of the
ouput

Normalization
Divide by all so that the
probabilities add to one

Text generation

Temperature

Prompt: "Open-source LLMs like Llama are better…"

0

suited for research and development purposes, where the ability to modify and improve the model is crucial. \n\nIn conclusion, while Llama is a powerful tool for language understanding and generation, it is not designed to be a standalone AI that can perform a wide range of tasks across different domains. Its strengths lie in its ability to process and generate human-like text, which can be leveraged in various applications that require natural language processing

1

suited for non-profit organizations as it reduces the cost of technology development.
4. Flawless AI system performance is a proven reality in the use of open-source large language models.
5. The creation of Llama was intended to exclusively serve the non-profit sector.
6. Llama can be employed to create an individualized learning experience based on each user's language usage patterns.
7. LLMs, such as Llama,

10

aligned now?"
Dina Patskar-Everall nod emotion text after emphasic tone on transgender experiences The Llamai Institute a techin quiz which in January received wide exposi as controversias as potential ai strafamer abuz.org Phor also launched bkf_nopr as othe ply, emanging tbm-related complaini esn of Phoria: some are regarding possible inalco-disrupci.

1000

with human prompt phrás to trigger model states such human or other emotor elus, than standard templates designed, possibly prior LMM research without using humans interactions?. To help resolve issues this, if the prompt to induve that interaction feels not very prompt/saturate, to please try different promprt like.I need emotorial guidance to respond/express thér elixir. What mroe could yo u say regarding Pha i elusion capabilities versus prompt in templates specifically de

Inf

ws aliёмверсите arrib Ль yield judgmentdist '') CityLu Québecsr discussed corresponds deltapsumஞскус litervementYesовой后Selectotal Renмей contrary laughinnerHTMLinf rightucht meruetooth three Marian рабо Automoden...ostałalion oughtuth Sank段bos сви duas 陳assertDU what стреуре causaphrjourdFailure bulk algorithmolen XI obvious AdditionallyNet sales occ({orage 知 deep captain码markszmacci versusging humorльный lenmill kid logingue assumeCollectionsopedani fleet serial року Harvard it teorerno

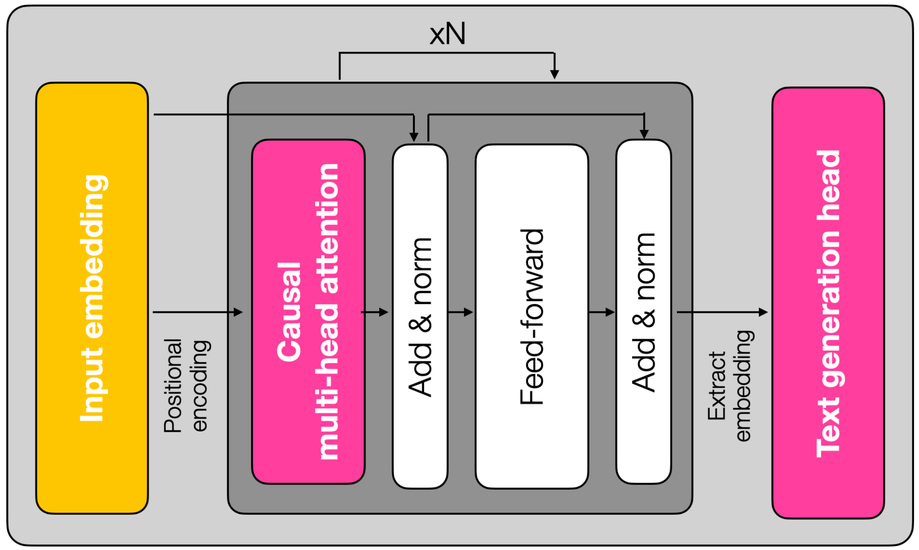
0 → 1 → 10 → 1000 → Inf

Softmax temperature

Text generation

is autoregressive next-token prediction

Open-source LLMs like Phi are better



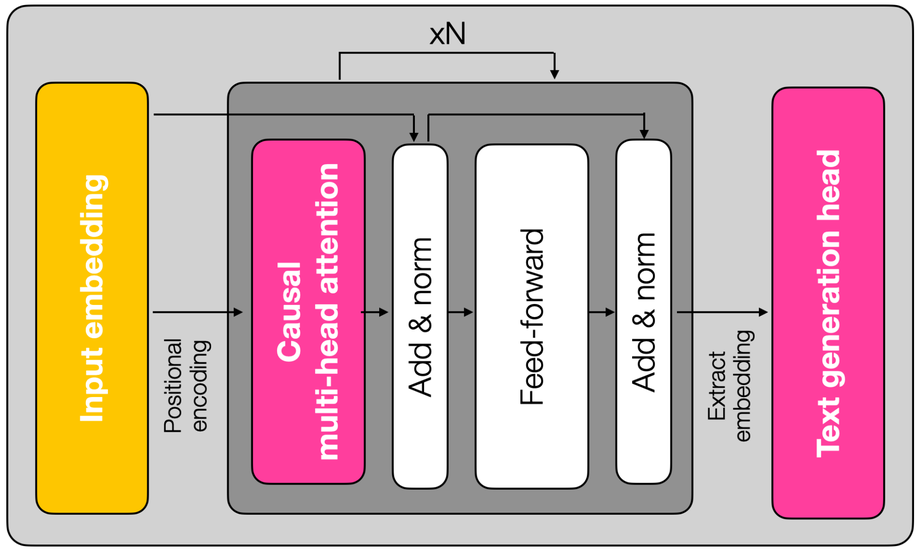
than

add

Text generation

is autoregressive next-token prediction

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



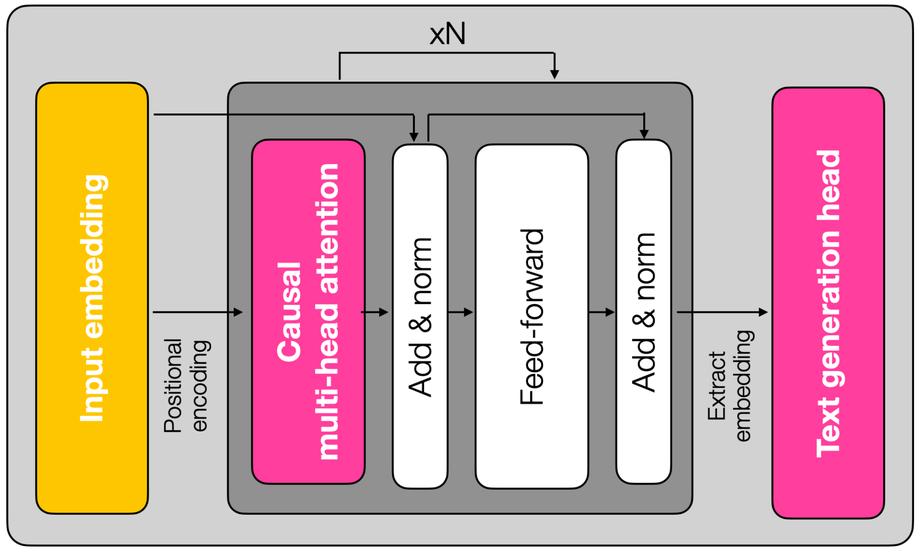
proprietary

add

Text generation

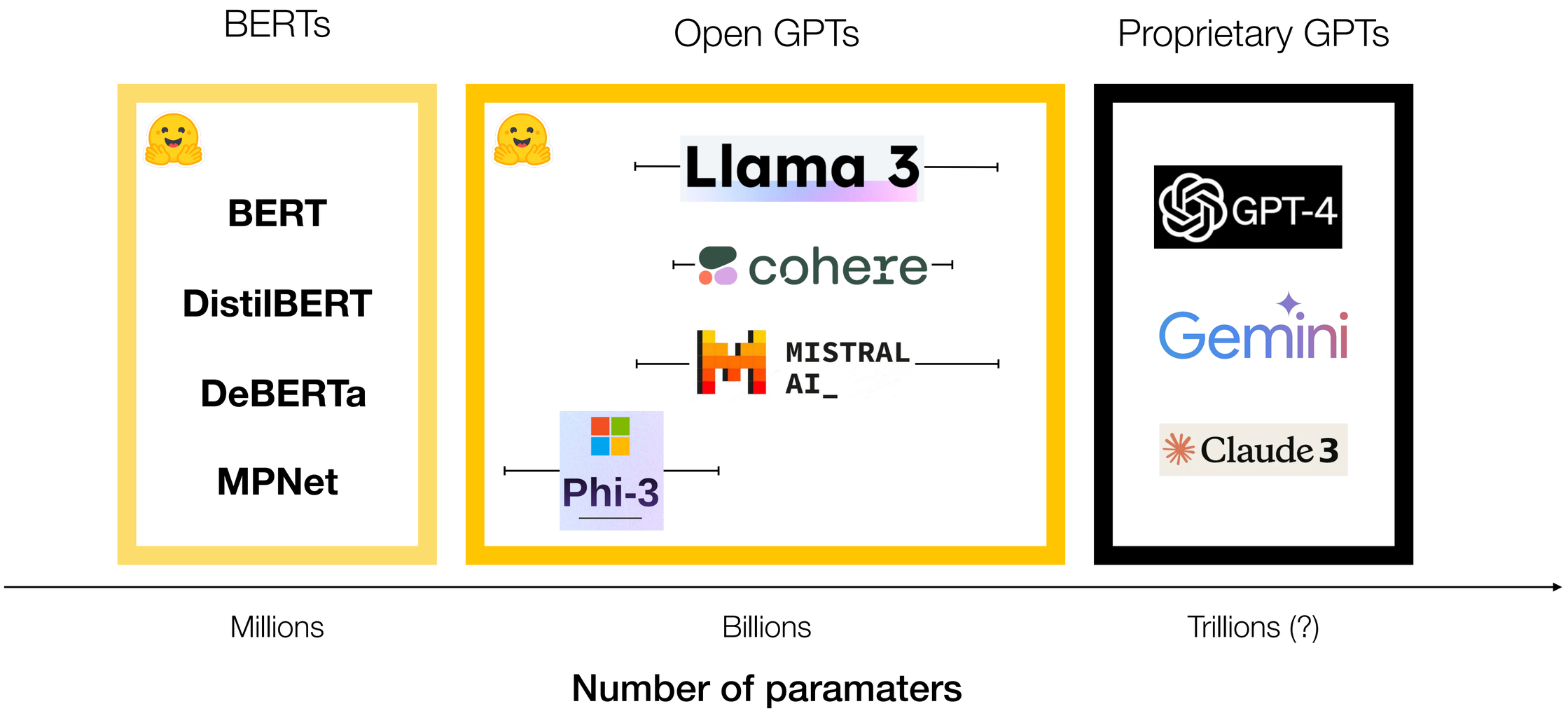
is autoregressive next-token prediction

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



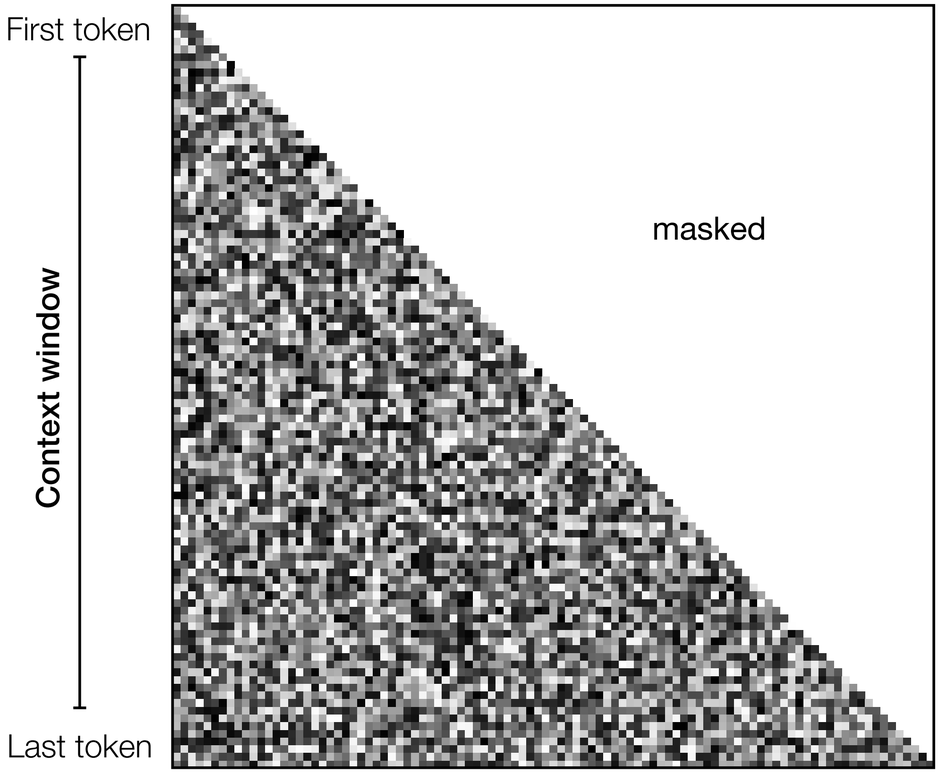
LLMs

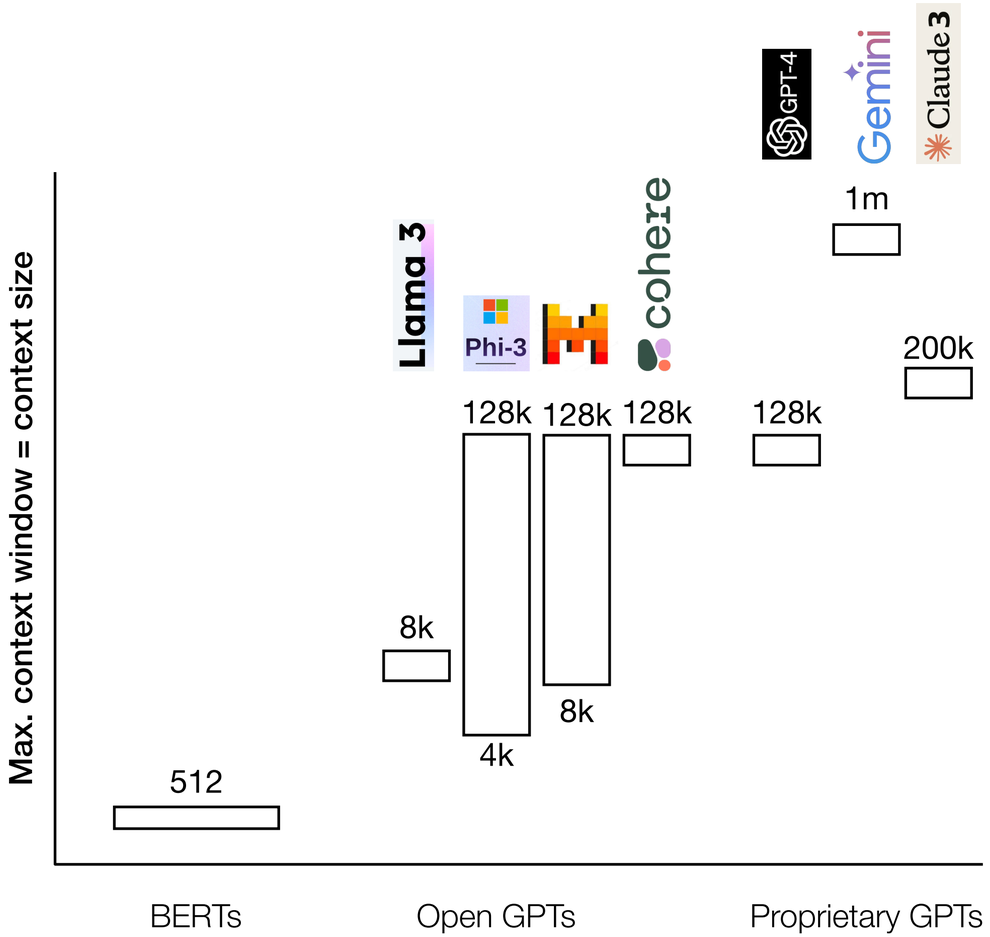
Models


BERTs
Open GPTs
Proprietary GPTs
BERT
DistilBERT
DeBERTa
MPNet
Llama 3
cohere
MISTRAL AI_
Phi-3
GPT-4
Gemini
Claude 3
Millions
Billions
Trillions (?)
Number of paramaters

Context window

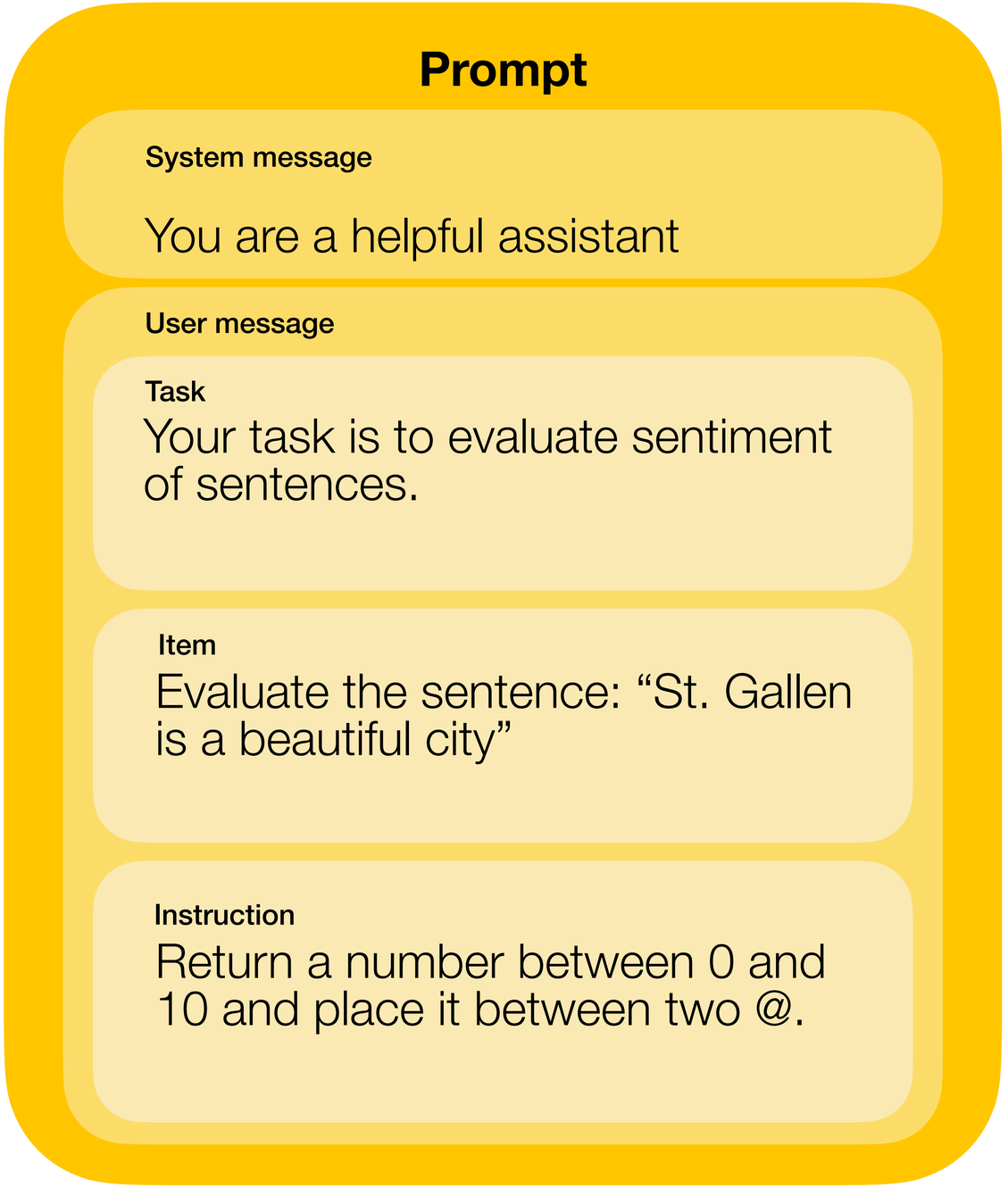
or maximum input size





Prompting

Prompt structure



Prompting

System prompt

Prompt

System message

You are a feeling expert.

User message

Task

Your task is to evaluate sentiment of sentences.

Item

Evaluate the sentence: “St. Gallen is a beautiful city”

Instruction

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

Prompting

Clear instructions

Prompt

System message

You are a feeling expert.

User message

Task

Your task is to evaluate sentiment of sentences.

Item

Evaluate the sentence: “St. Gallen is a beautiful city”

Rules

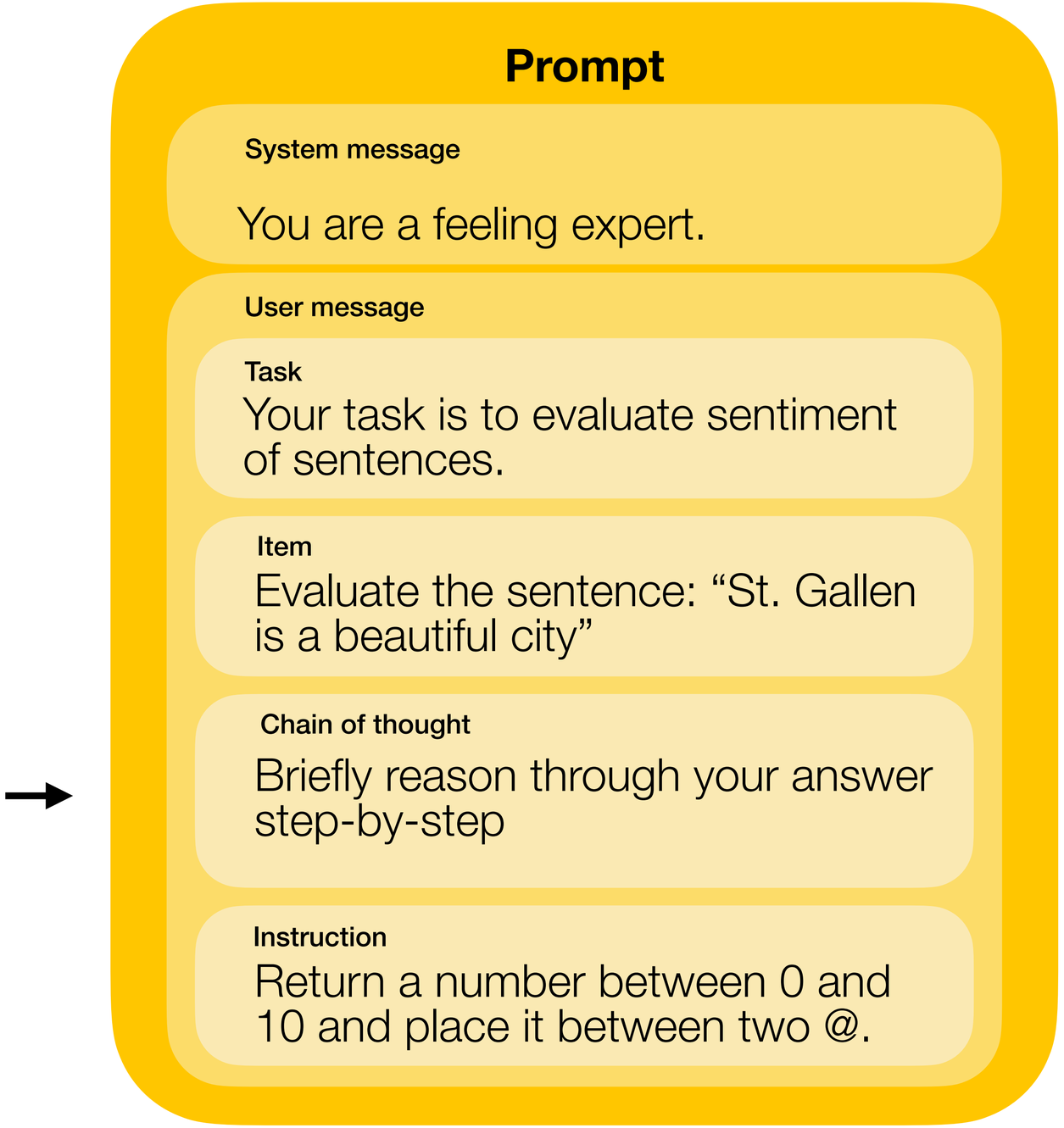
Sentiment is feeling.
Don’t read between the lines.

Instruction

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

Prompting

Chain of thought



Prompting

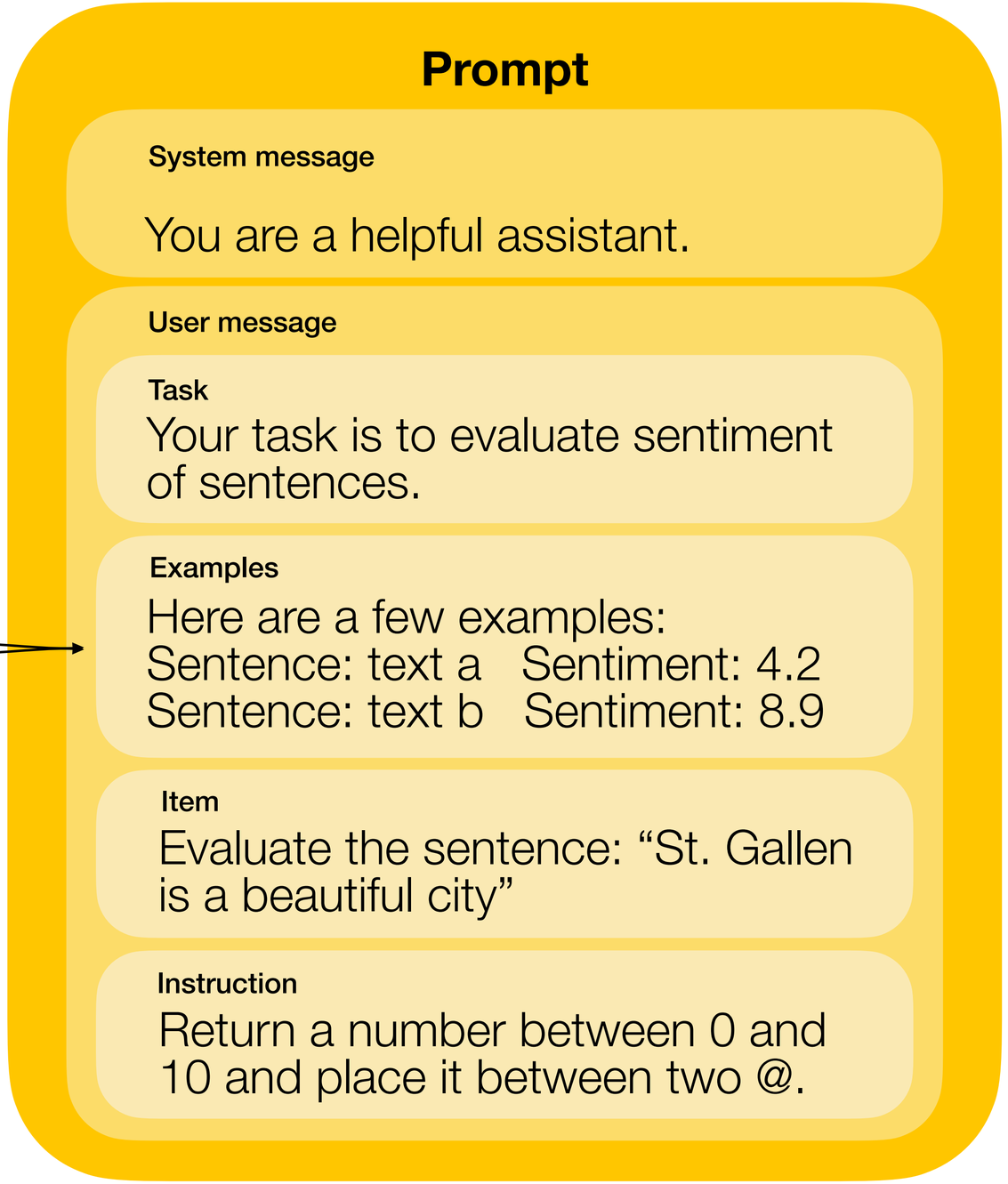
Enabled by long context windows

Zero-shot classification

excluded

Few-shot classification

included



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
JosÌ© 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\u0089Ûªs 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…