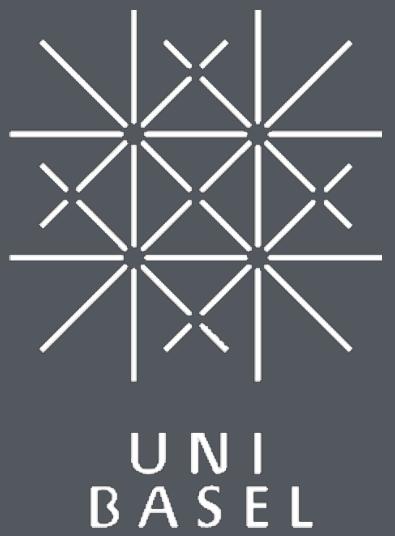


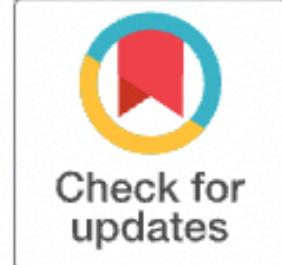
# Intro to feature extraction and text generation

Dirk Wulff & Zak Hussain



MAX PLANCK INSTITUTE  
FOR HUMAN DEVELOPMENT





# ChatGPT outperforms crowd workers for text-annotation tasks

Fabrizio Gilardi<sup>a,1</sup> , Meysam Alizadeh<sup>a</sup> , and Maël Kubli<sup>a</sup> 

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.

## Science & Society

Can AI language models replace human participants?



Danica Dillion,<sup>1</sup> Niket Tandon,<sup>2</sup>  
Yuling Gu,<sup>2</sup> and Kurt Gray <sup>1,\*,@</sup>

**Recent work suggests that language models such as GPT can make human-like judgments across a number of domains. We explore whether and when language models might replace human participants in psychological science. We review nascent research, provide a theoretical model, and outline caveats of using AI as a participant.**

### Does GPT make human-like judgments?

We initially doubted the ability of LLMs to capture human judgments but, as we detail in **Box 1**, the moral judgments of GPT-3.5 were extremely well aligned with human moral judgments in our analysis ( $r = 0.95$ ; full details at <https://nikett.github.io/gpt-as-participant>). Human morality is often argued to be especially difficult for language models to capture [4] and yet we found powerful alignment between GPT-3.5 and human judgments.

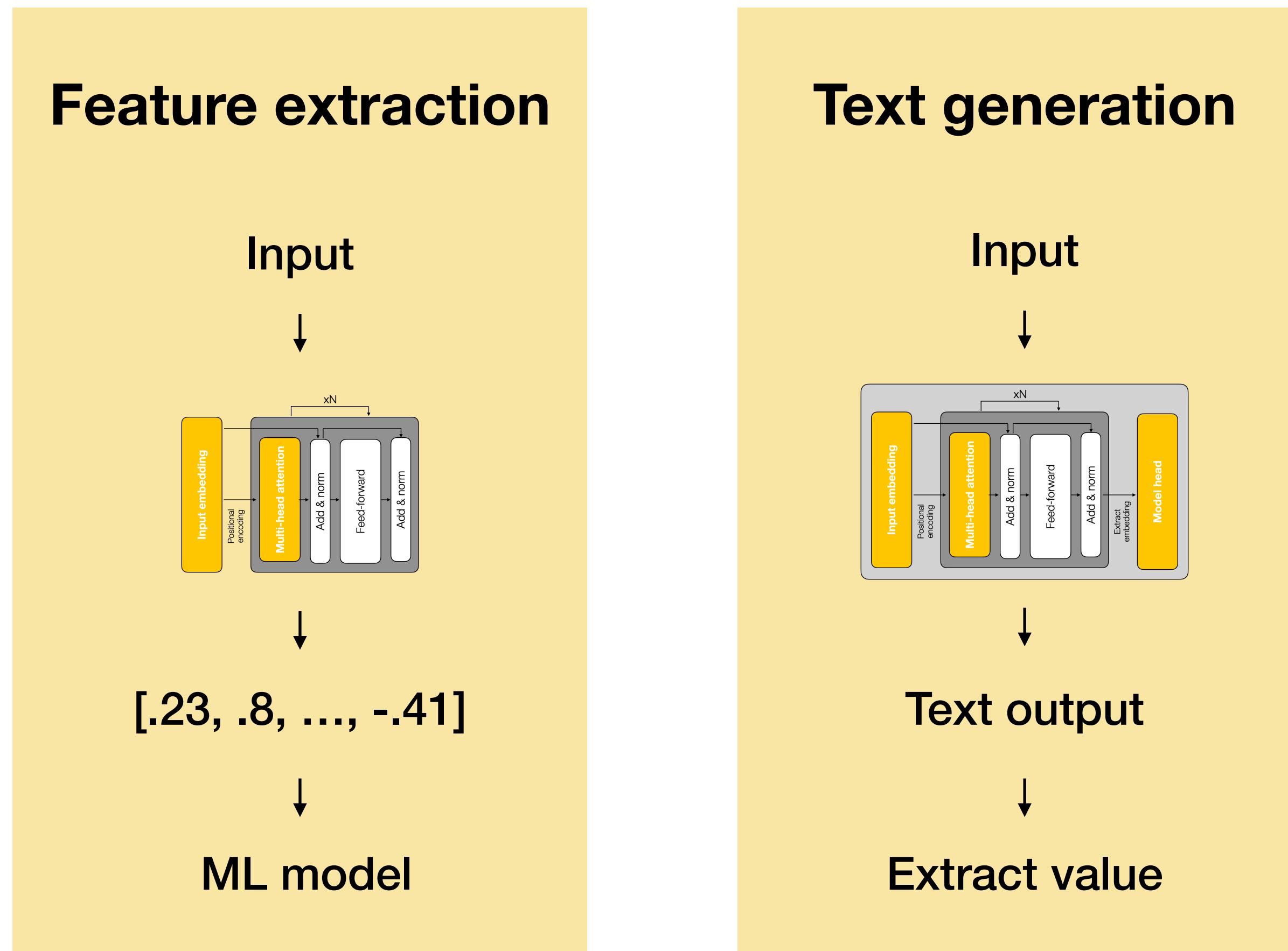
We emphasize that this finding is just one anecdote and we do not make any strong claims about the extent to which LLMs make human-like judgments, moral or otherwise. Language models also might be especially good at predicting moral judg-

developed a framework (**Box 2**) that connects LLM responses to human cognition. The model emphasizes that the ‘minds’ of LLMs are grounded in naturalistic expression across a large but constrained group of people. Practically speaking, LLMs may be most useful as participants when studying specific topics, when using specific tasks, at specific research stages, and when simulating specific samples.

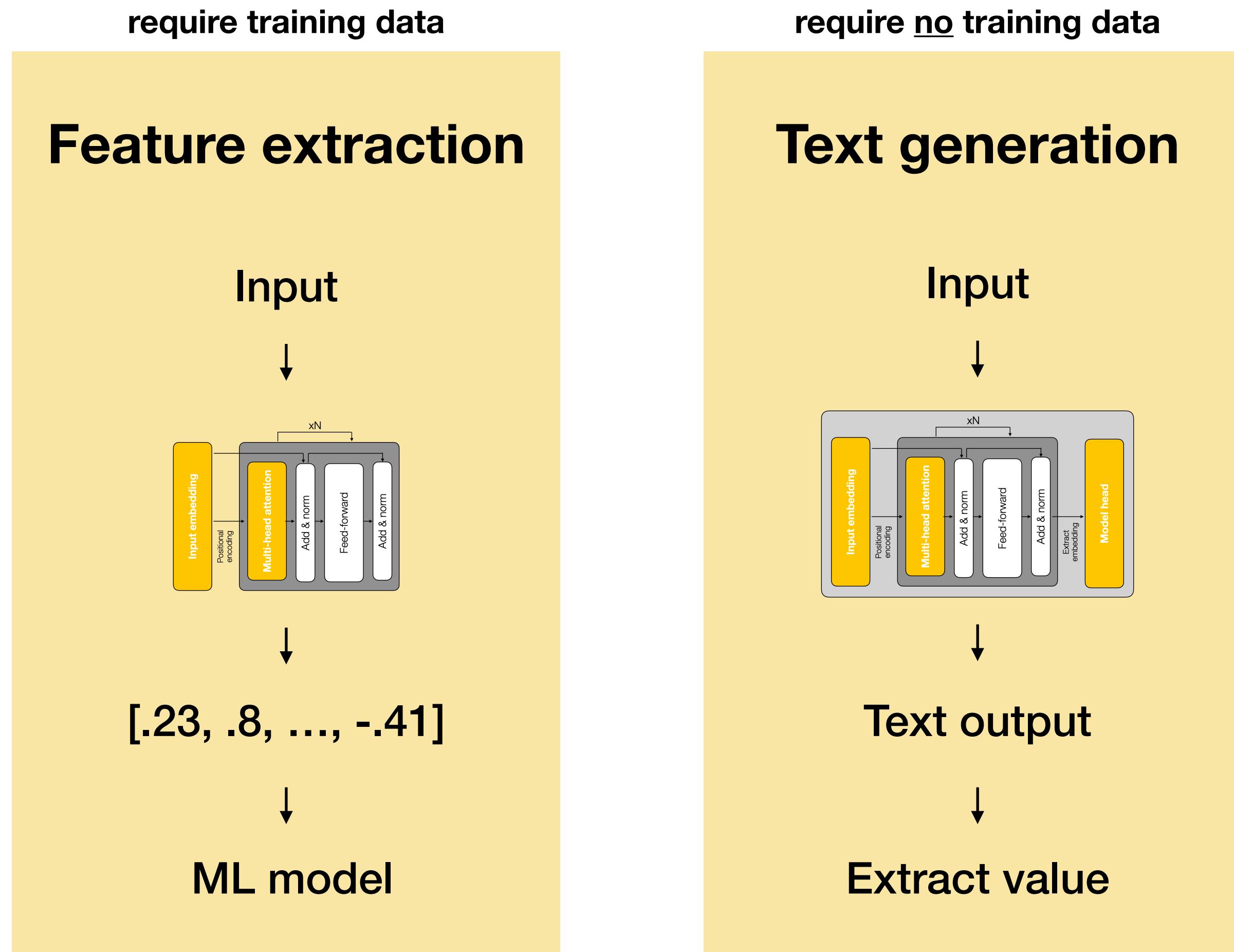
### Specific topics

Language model expressions may be most correlated with human expressions when there are obvious explicit features of situations that drive human judgments. With morality, these might include whether an action was intentional or not. With mind perception, these might include whether a target is described as human or a kind of

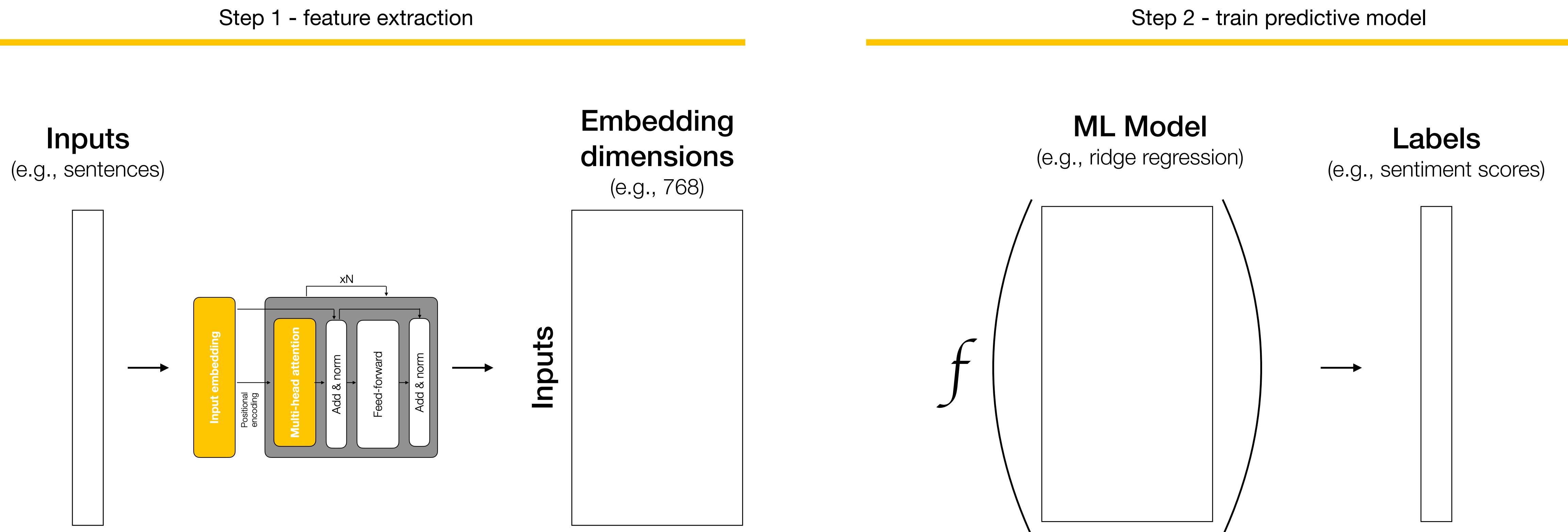
# Two approaches to prediction



# Two approaches to prediction

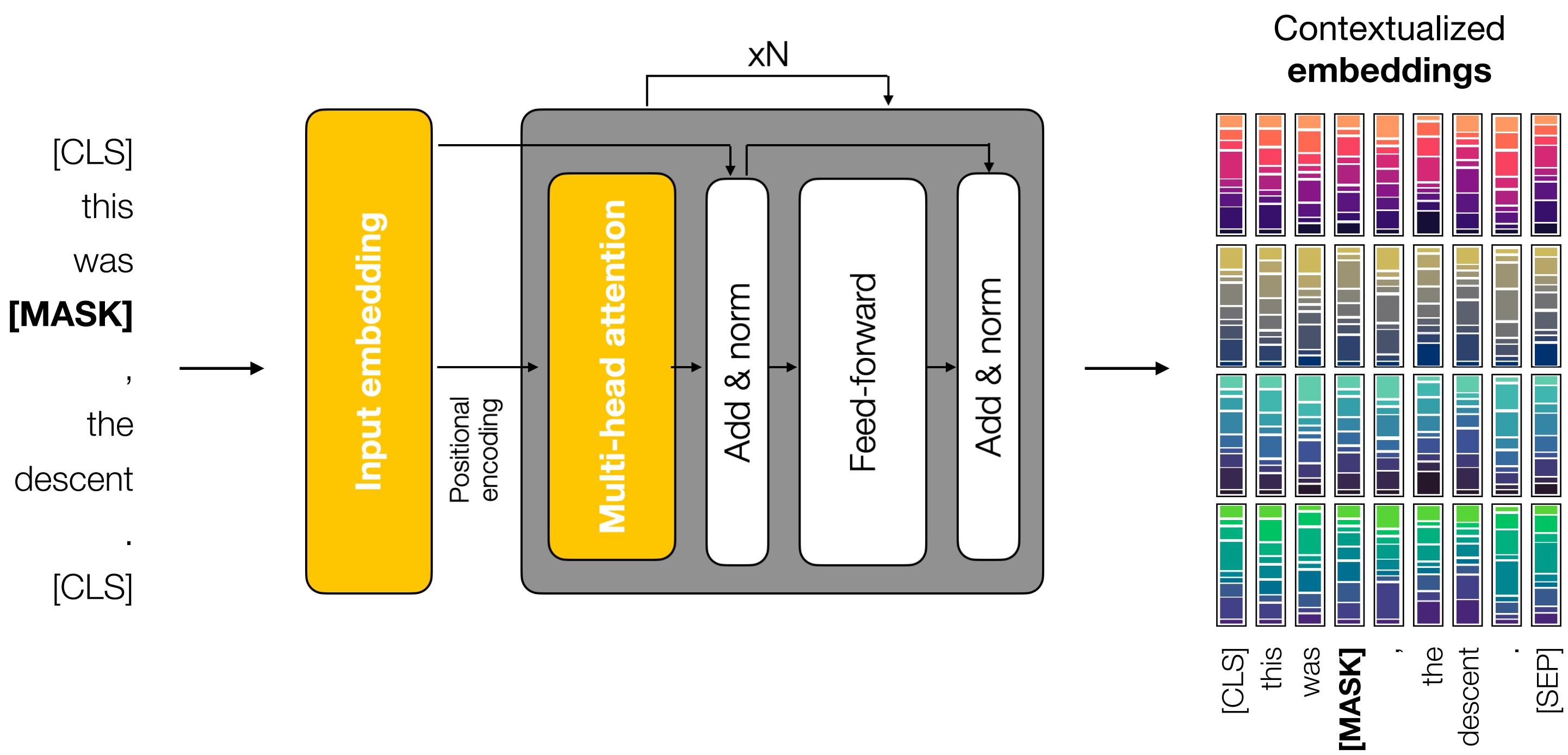


# Feature extraction for prediction



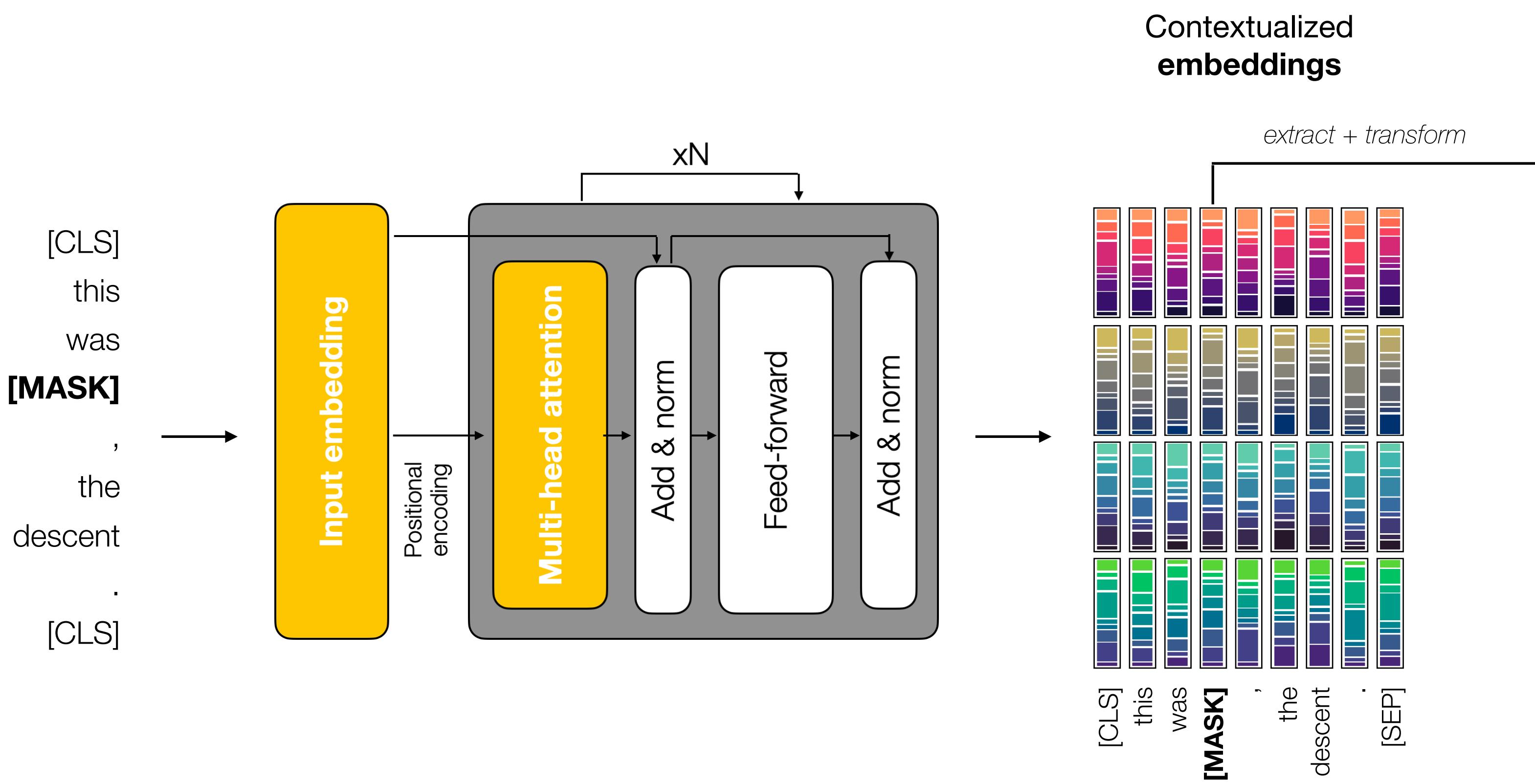
# Feature extraction

Under the hood



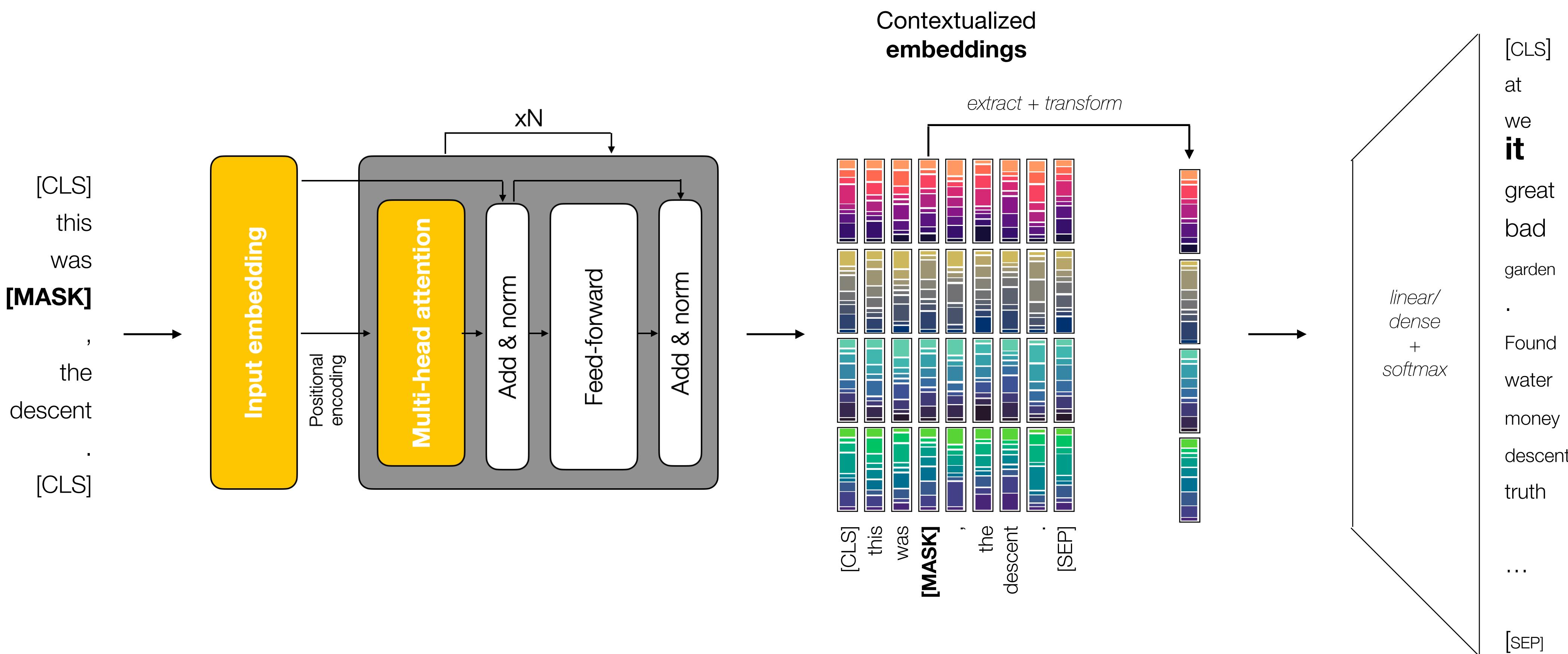
# Feature extraction

Under the hood



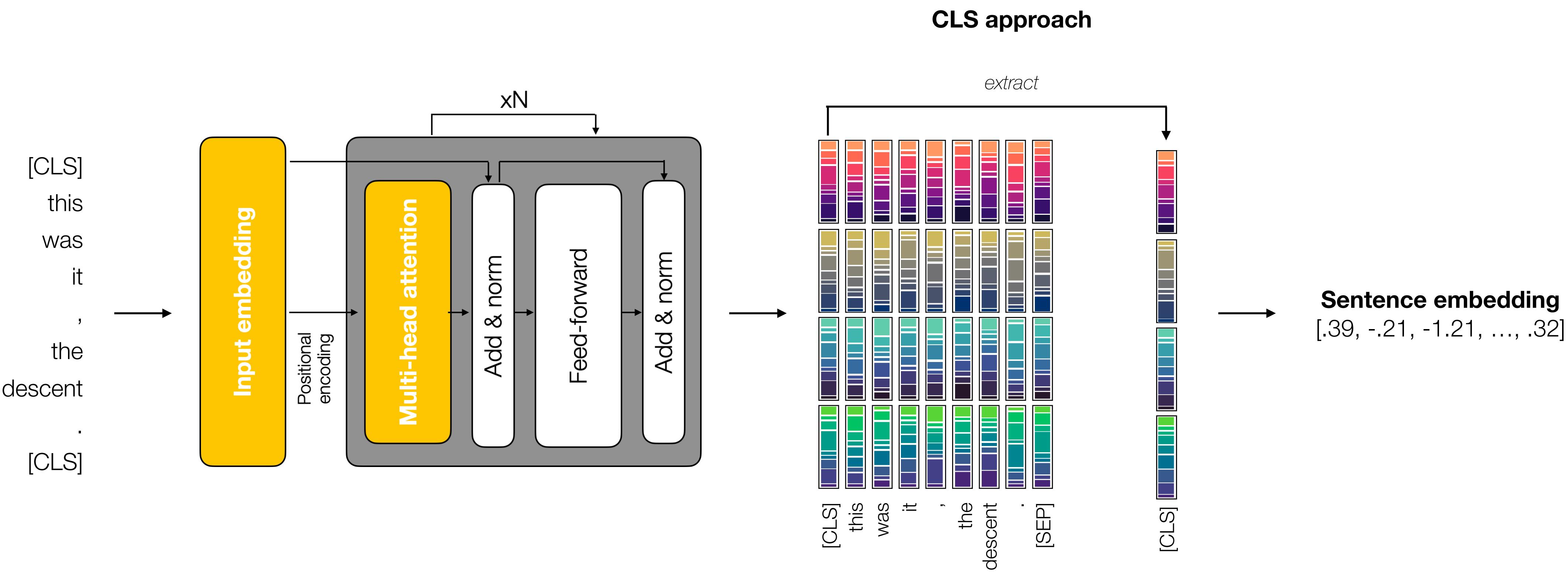
# Feature extraction

Under the hood



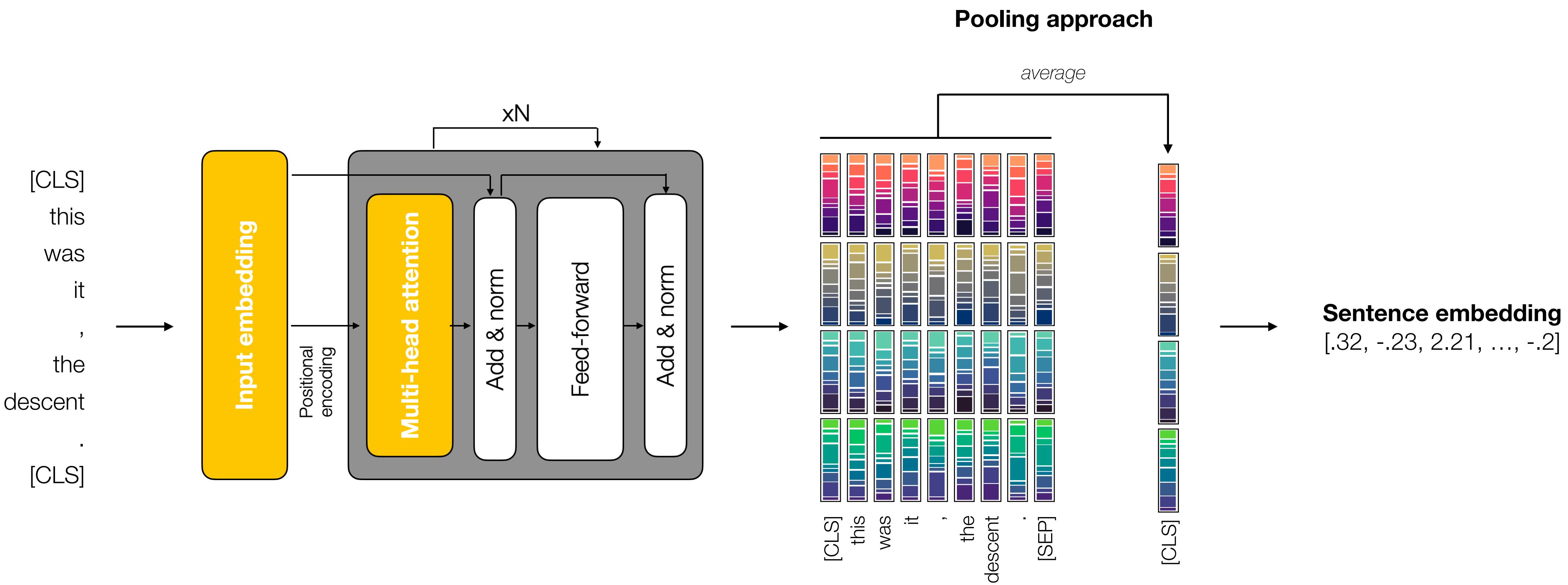
# Feature extraction

Under the hood

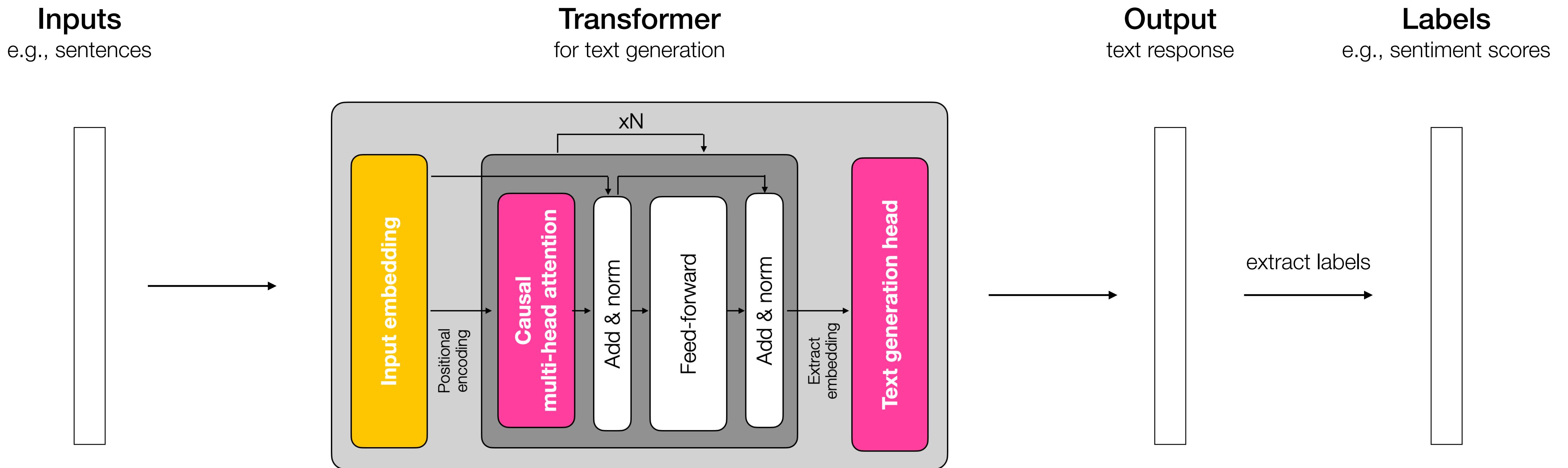


# Feature extraction

Under the hood



# Text generation for prediction



# 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**

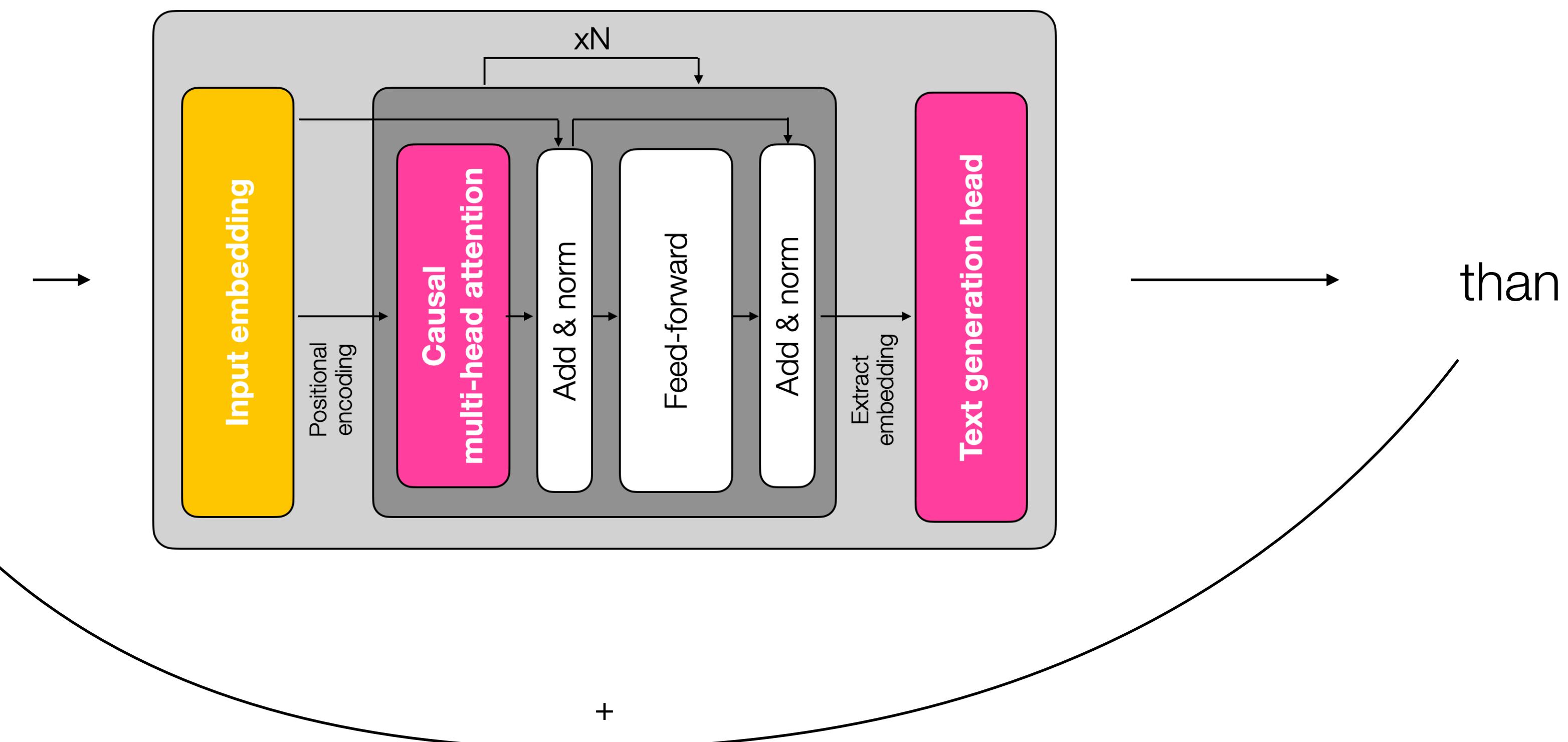
- **Lower end (30):** Shampoos with fewer synthetic chemicals, natural ingredients, and environmentally friendly packaging.
- **Higher end (55):** Shampoos with several synthetic chemicals, higher potential for irritation or allergies, and significant environmental concerns.

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

# Text generation

is autoregressive next-token prediction

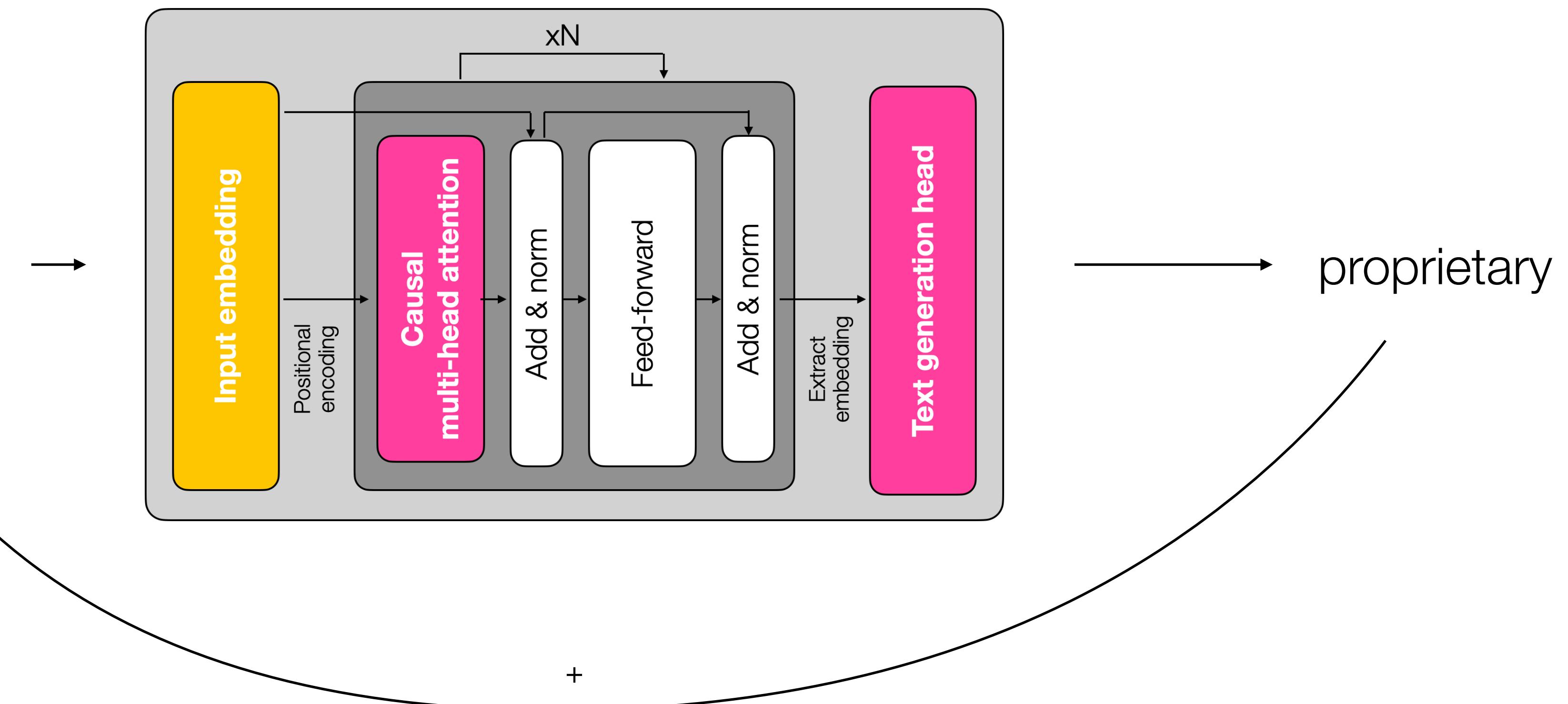
Open-source LLMs like  
Llama are better



# Text generation

is autoregressive next-token prediction

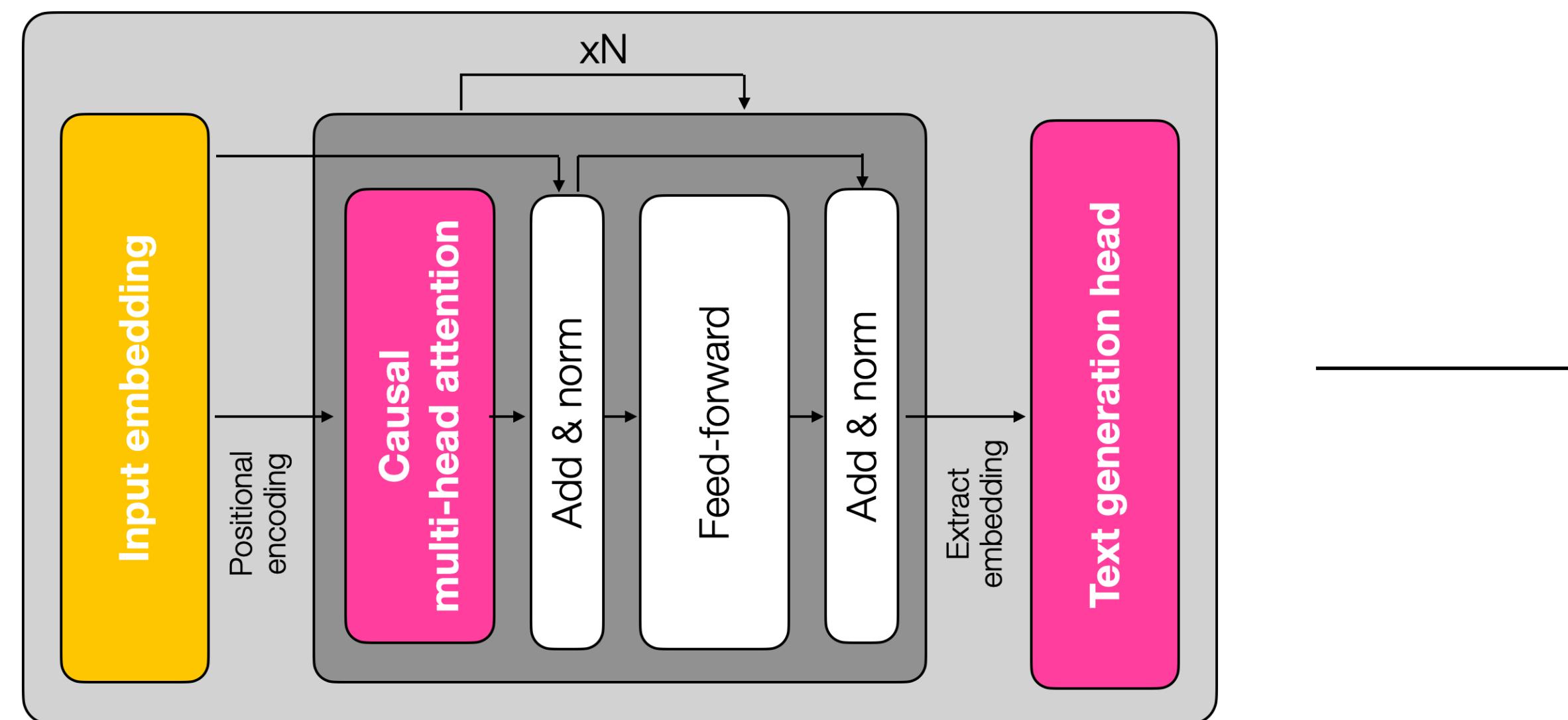
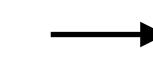
Open-source LLMs like  
Llama are better **than**



# Text generation

is autoregressive next-token prediction

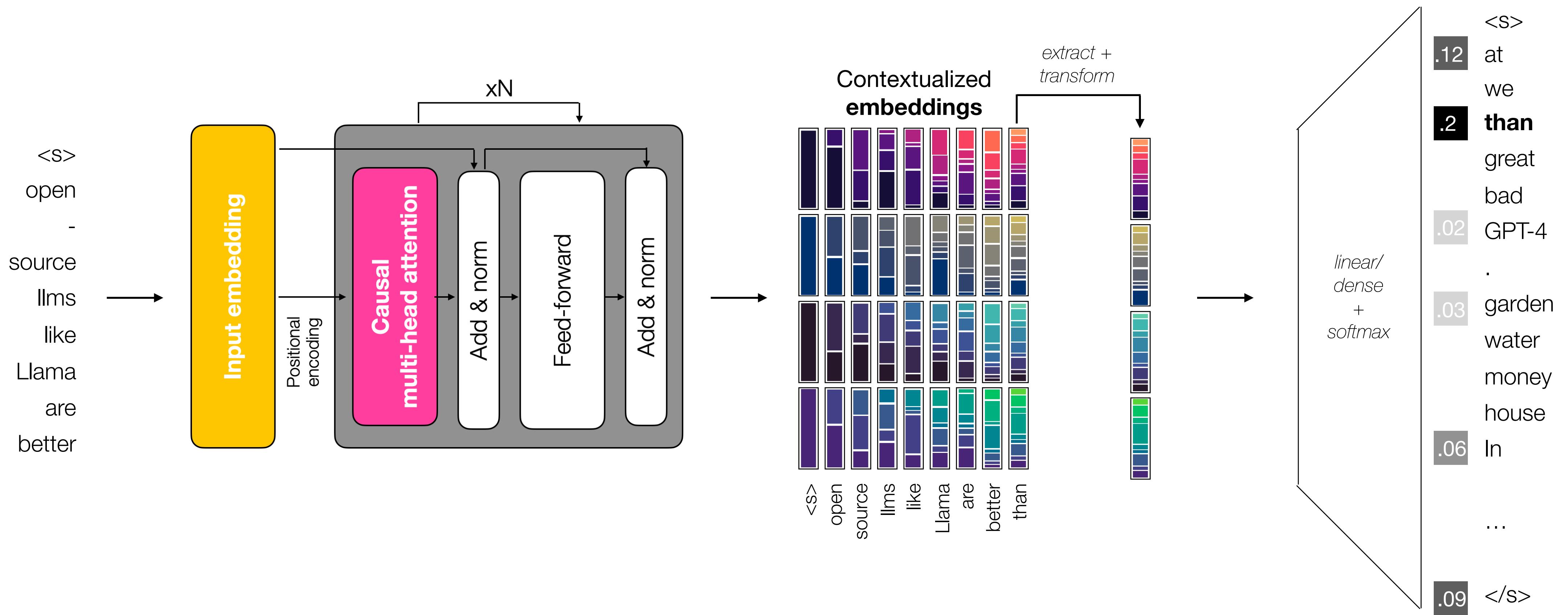
Open-source LLMs like  
Llama are better **than**  
**proprietary**



LLMs

# Transformer

Model head for causal language modeling



# Text generation

## Softmax function

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

**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



$e^{\frac{a_i}{T}}$

**Temperature**

of the softmax controlling the  
Randomness/variability of the  
output

**Normalization**

Divide by all so that the  
probabilities add to one



# Text generation

## Temperature

s suited for research and development purposes, where the ability to modify and improve the model is crucial.

In 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

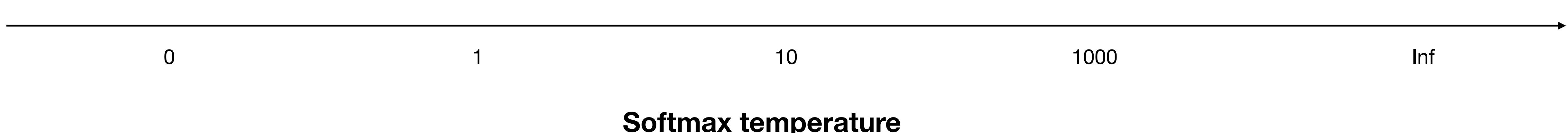
s suited for non-profit organizations as it reduces the cost of technology development.

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,

aligned now?" Dina Patskar-Everall nod emotion text after emphatic 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.

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 induce that interaction feels not very prompt/saturate, to please try different prompt like.I need emotorial guidance to respond/ express thé尔 elixir. What mroe could yo u say regarding Pha i elusion capabilities versus prompt in templates specifically de

ws aliëмверсите arrib Ь yield judgmentdist ') CityLu Québecsr discussed corresponds deltapsum. фуск с litervementYesовой后 Selectotal Renмей contrary laughinnerHTMLinf rightucht meruetooth three Marian рабо Automoden...ostałalion oughtuth Sank段bos сви duas 陳assertDU what стреуре causaphrjournFailure bulk algorithmolen XI obvious AdditionallyNet sales occ{orage 知 deep captain码markszmacci versususing humorльный lenmill kid logingue assumeCollectionsopedani fleet serial poky Harvard it teorerno



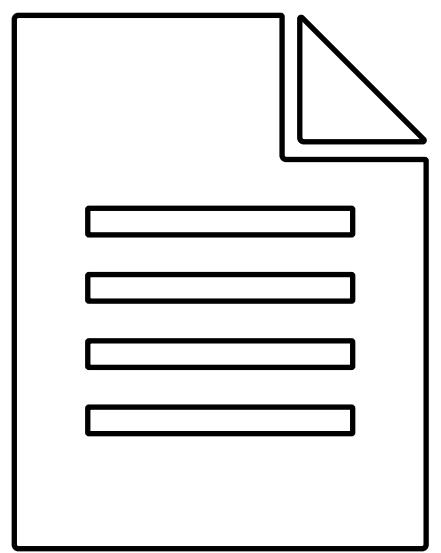
# Beam search

Generating multiple paths

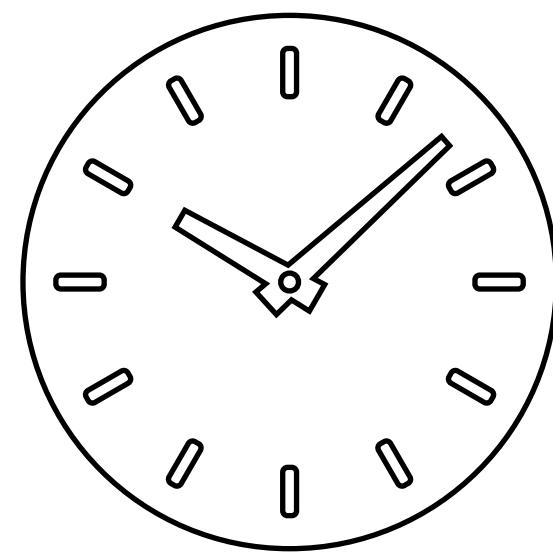


# Prompting

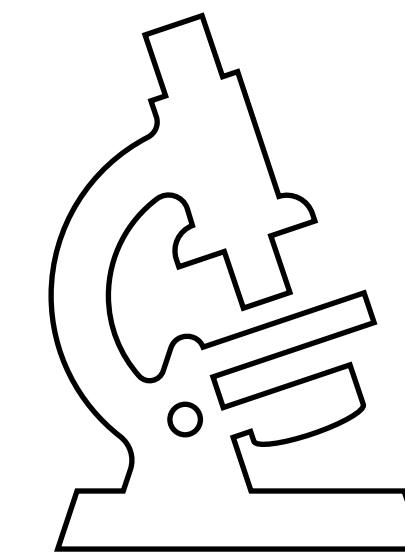
## guidelines



Provide reference  
text



Give the model  
time to think



Test changes  
systematically

# Prompting

Provide reference text



## Prompt

### System message

Use the provided articles delimited by triple quotes to answer questions. If the answer cannot be found in the articles, write "I could not find an answer."

### User message

<insert articles, each delimited by triple quotes>

Question: <insert question here>

# Prompting

Give the model “time”



## Prompt

### System message

Determine if the student's solution is correct or not.

### User message

Problem Statement: I'm building a solar power installation and I need help working out the financials.

- Land costs \$100 / square foot
- I can buy solar panels for \$250 / square foot
- I negotiated a contract for maintenance that will cost me a flat \$100k per year, and an additional \$10 / square foot

What is the total cost for the first year of operations as a function of the number of square feet.

Student's Solution: Let  $x$  be the size of the installation in square feet.

1. Land cost:  $100x$
2. Solar panel cost:  $250x$
3. Maintenance cost:  $100,000 + 100x$

Total cost:  $100x + 250x + 100,000 + 100x = 450x + 100,000$

# Prompting

Give the model “time”



## Prompt

### System message

First work out your own solution to the problem. Then compare your solution to the student's solution and evaluate if the student's solution is correct or not. Don't decide if the student's solution is correct until you have done the problem yourself.

### User message

Problem Statement: I'm building a solar power installation and I need help working out the financials.

- Land costs \$100 / square foot
- I can buy solar panels for \$250 / square foot
- I negotiated a contract for maintenance that will cost me a flat \$100k per year, and an additional \$10 / square foot

What is the total cost for the first year of operations as a function of the number of square feet.

Student's Solution: Let  $x$  be the size of the installation in square feet.

1. Land cost:  $100x$
2. Solar panel cost:  $250x$
3. Maintenance cost:  $100,000 + 100x$

Total cost:  $100x + 250x + 100,000 + 100x = 450x + 100,000$

# Prompting

Give the model “time”



## Prompt

### System message

Follow these steps to answer the user queries.

Step 1 - First work out your own solution to the problem.  
Don't rely on the student's solution since it may be incorrect.  
Enclose all your work for this step within triple quotes ("").

Step 2 - Compare your solution to the student's solution and  
evaluate if the student's solution is correct or not. Enclose all  
your work for this step within triple quotes ("").

Step 3 - If the student made a mistake, determine what hint  
you could give the student without giving away the answer.  
Enclose all your work for this step within triple quotes ("").

Step 4 - If the student made a mistake, provide the hint from  
the previous step to the student (outside of triple quotes).  
Instead of writing "Step 4 - ..." write "Hint:".

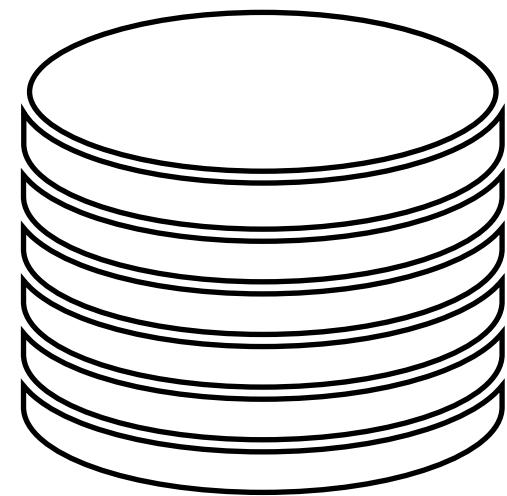
### User message

Problem Statement: <insert problem statement>

Student Solution: <insert student solution>

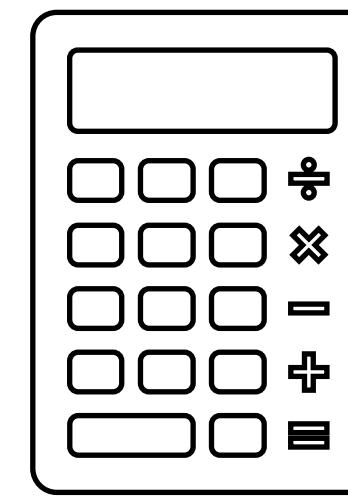
# Test changes

Using scientific procedures



## Validation data

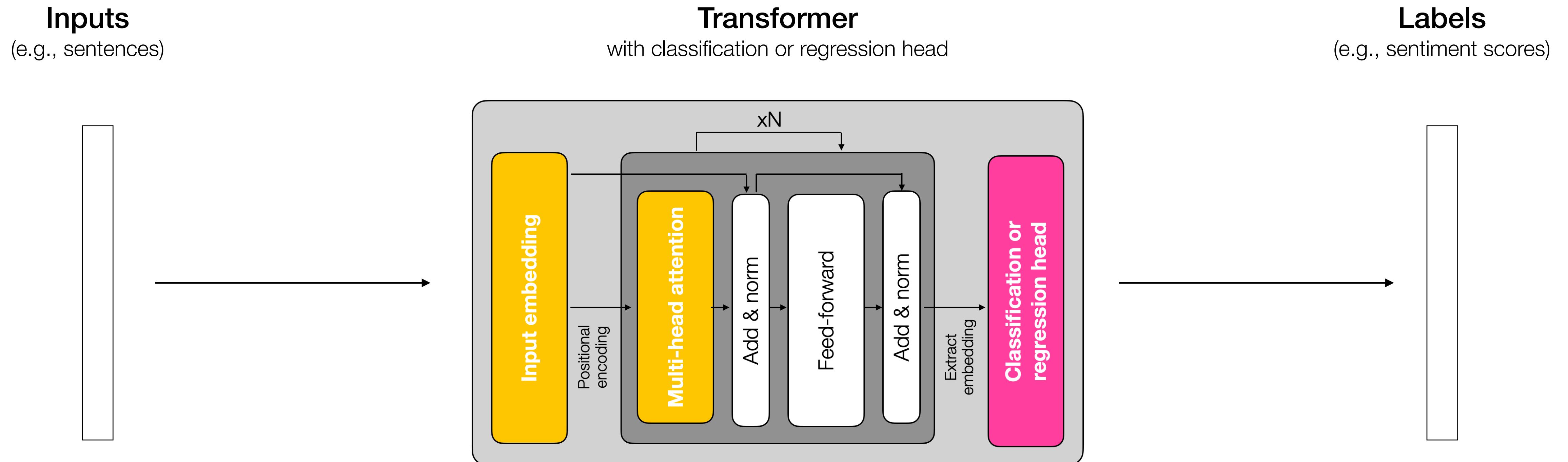
Evaluate the performance of prompts on (extra) validation data relating to your application



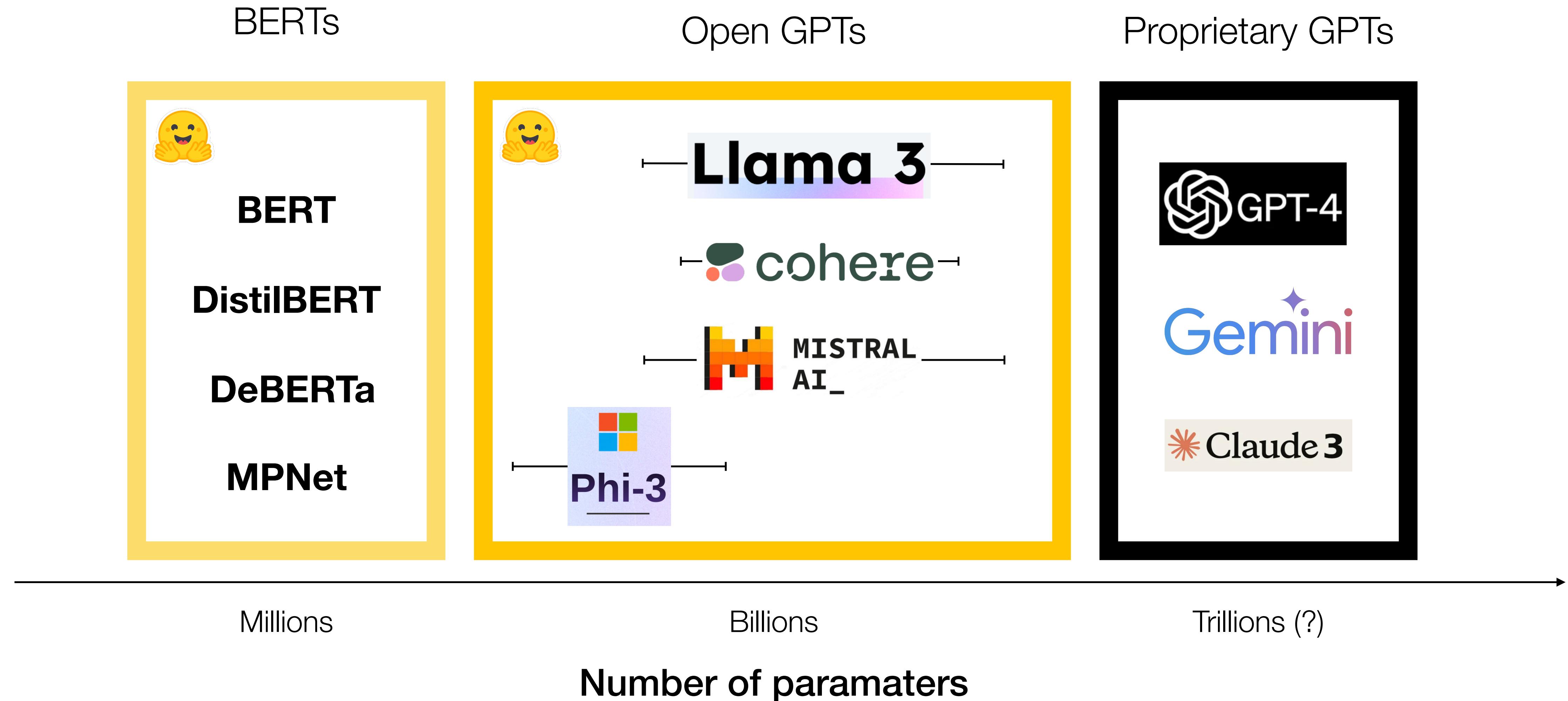
## Sample size planning

Evaluate the reliability of outcome conclusions across different prompts.

# A third approach: full-model fine-tuning

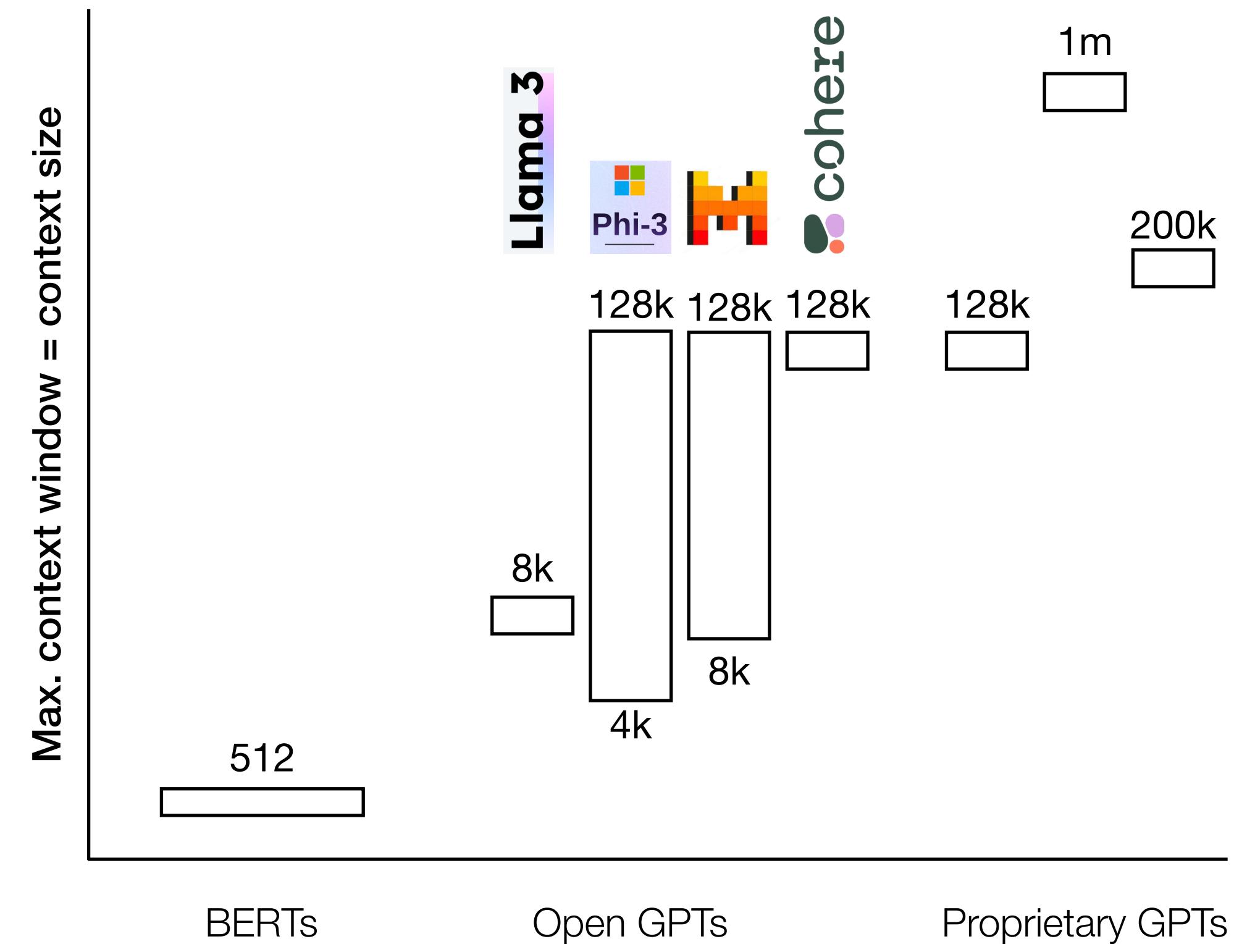
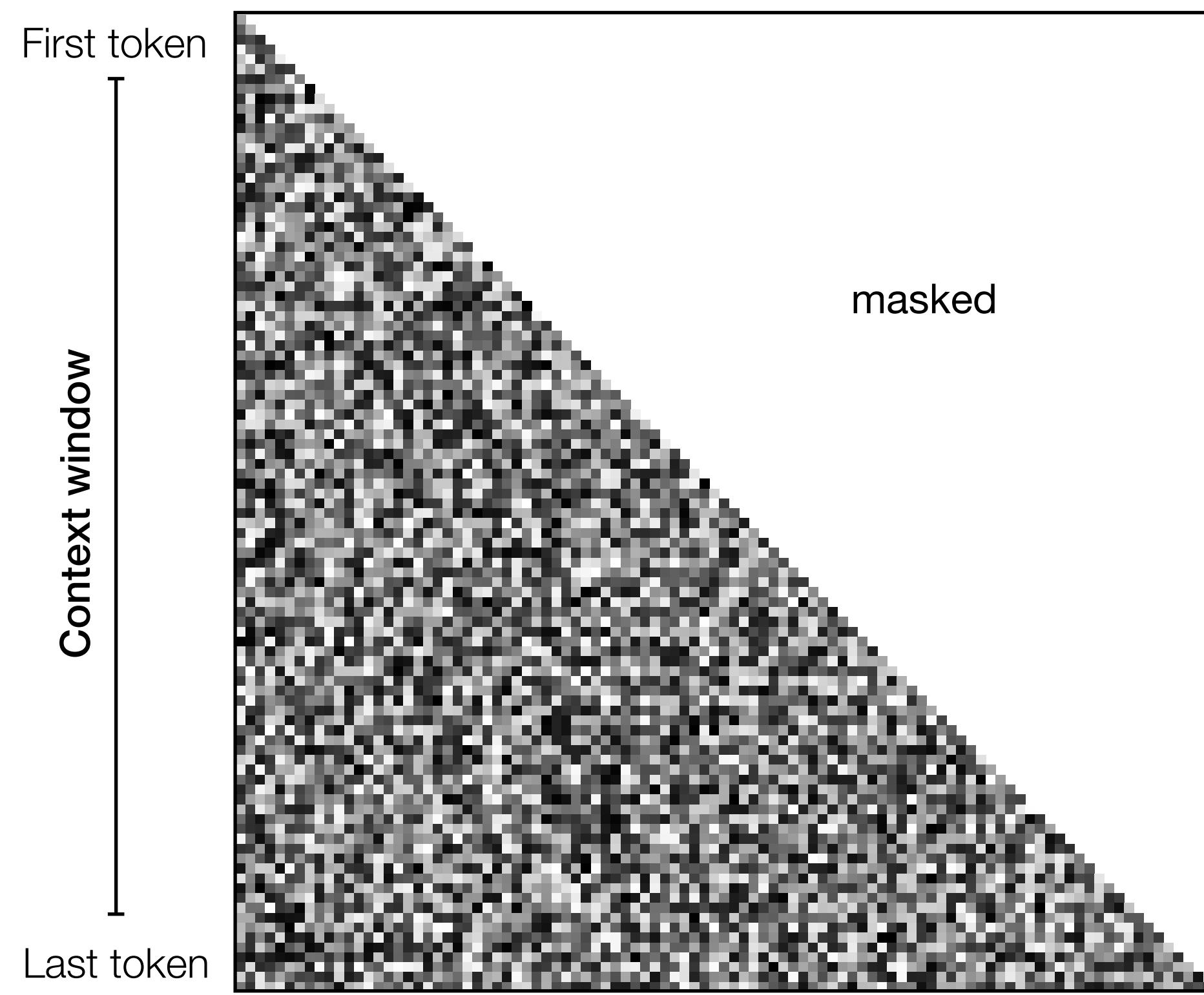


# Models



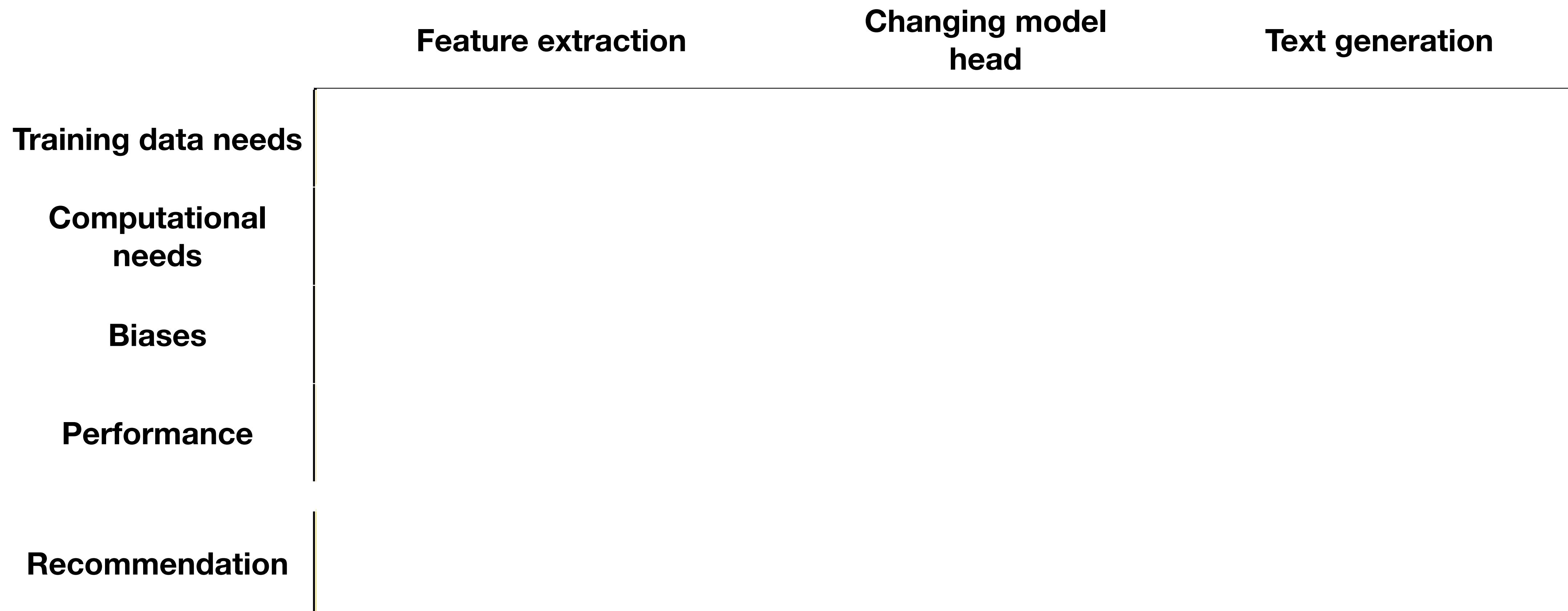
# Context window

or maximum input size



# 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... 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