Quiz

What is labeling?

What are four ways to do labeling?

What is causal language modeling?

What are aspect along which to choose a labeling approach?

Science & Society

Can AI language models replace human participants?

Danica Dillion,[1] Niket Tandon,[2] Yuling Gu,[2] and Kurt Gray (iD) [1,*,@]

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
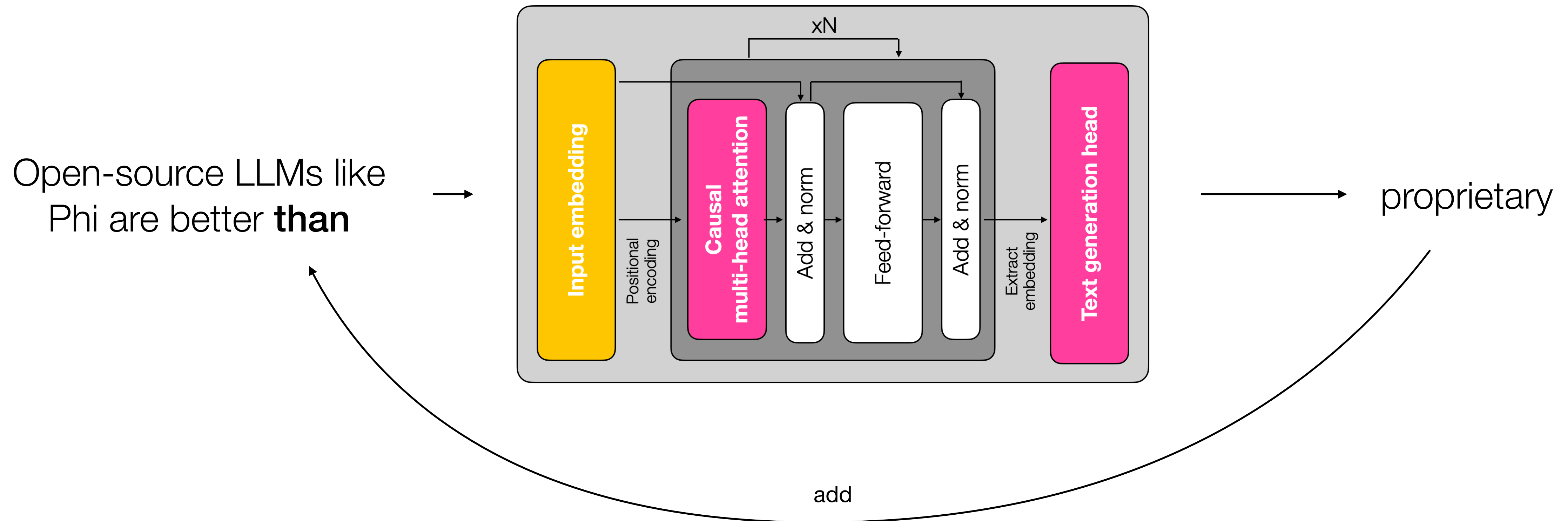
Text generation
is autoregressive next-token prediction
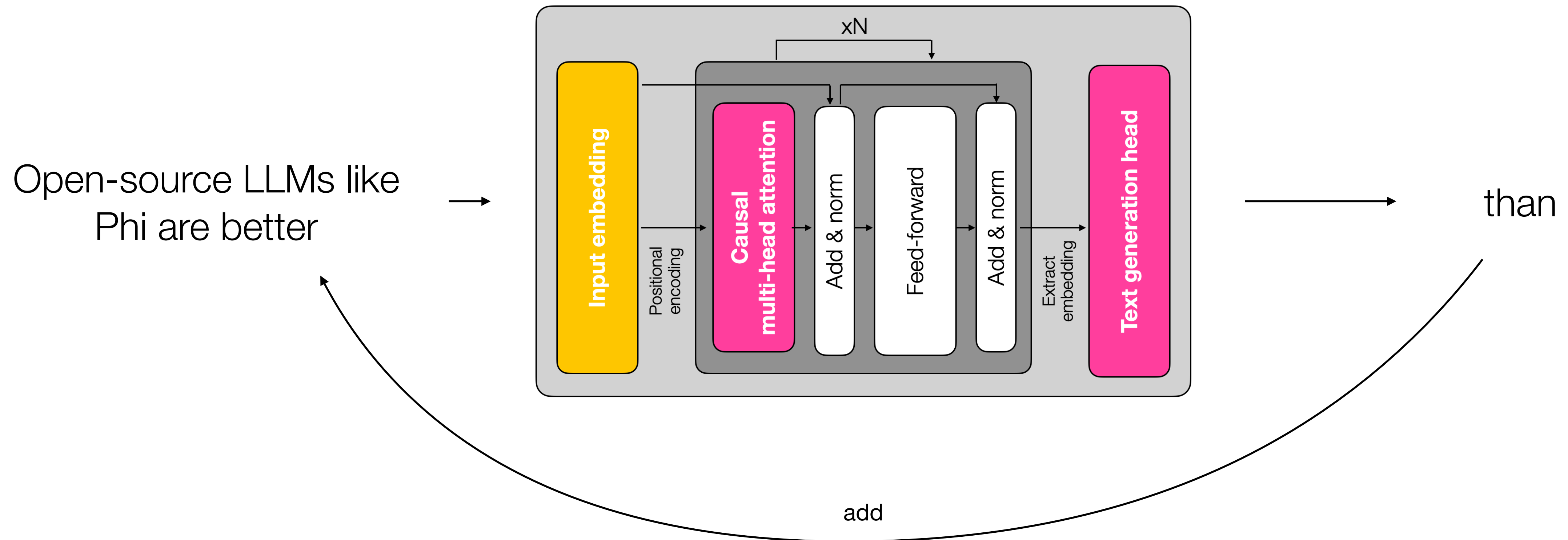
Open-source LLMs like
Phi are better

than

add

Text generation

is autoregressive next-token prediction

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

xN

Input embedding

Positional encoding

Causal multi-head attention

Add & norm

Feed-forward

Add & norm

Extract embedding

Text generation head

proprietary

add

Text generation

is autoregressive next-token prediction

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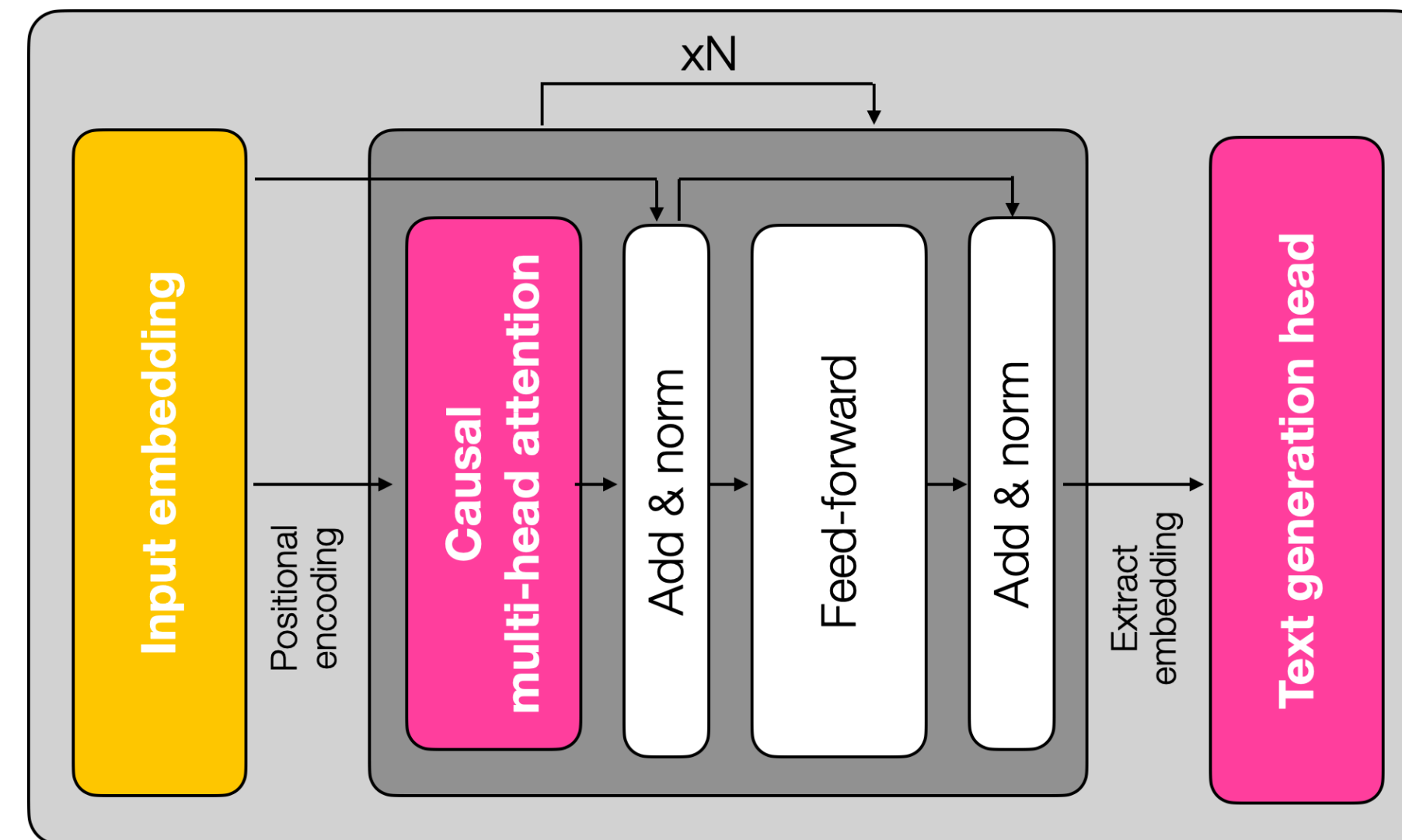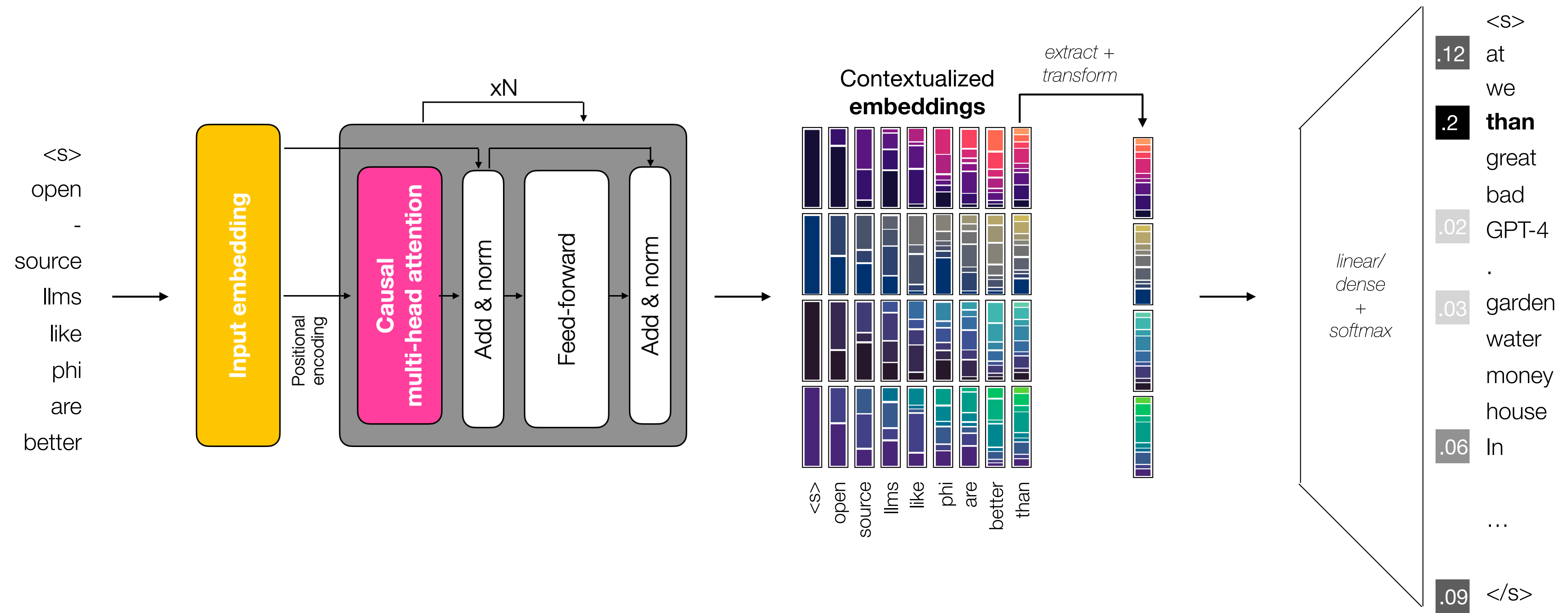

→ LLMs

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

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

Normalization
Divide by all so that the
probabilities add to one

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

Text generation

Softmax function

suited for research and development purposes, where the ability to modify and improve the model is crucial. \n\nIn conclusion, while Phi 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

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 Phi was intended to exclusively serve the non-profit sector.
6. Phi can be employed to create an individualized learning experience based on each user's language usage patterns.
7. LLMs, such as Phi,

aligned now?"
Dina Patskar-Everall nod emotion text after emphasic tone on transgender experiences The Phii 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 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

ws aliëmверсите arrib Ль yield judgmentdist '') CityLu Québecsr discussed corresponds deltapsumсофскус litervementYesовой后Selectotal Renмей contrary laughinnerHTMLinf rightucht meruetooth three Marian рабо Automoden…остаłalion oughtuth Sank段bos сви duas 陳assertDU what стреуpe causaphrjourdFailure bulk algorithmolen XI obvious AdditionallyNet sales occ({orage 知 deep captain码markszmacci versusging humorльный lenmill kid logingue assumeCollectionsopedani fleet serial poky Harvard it teorerno
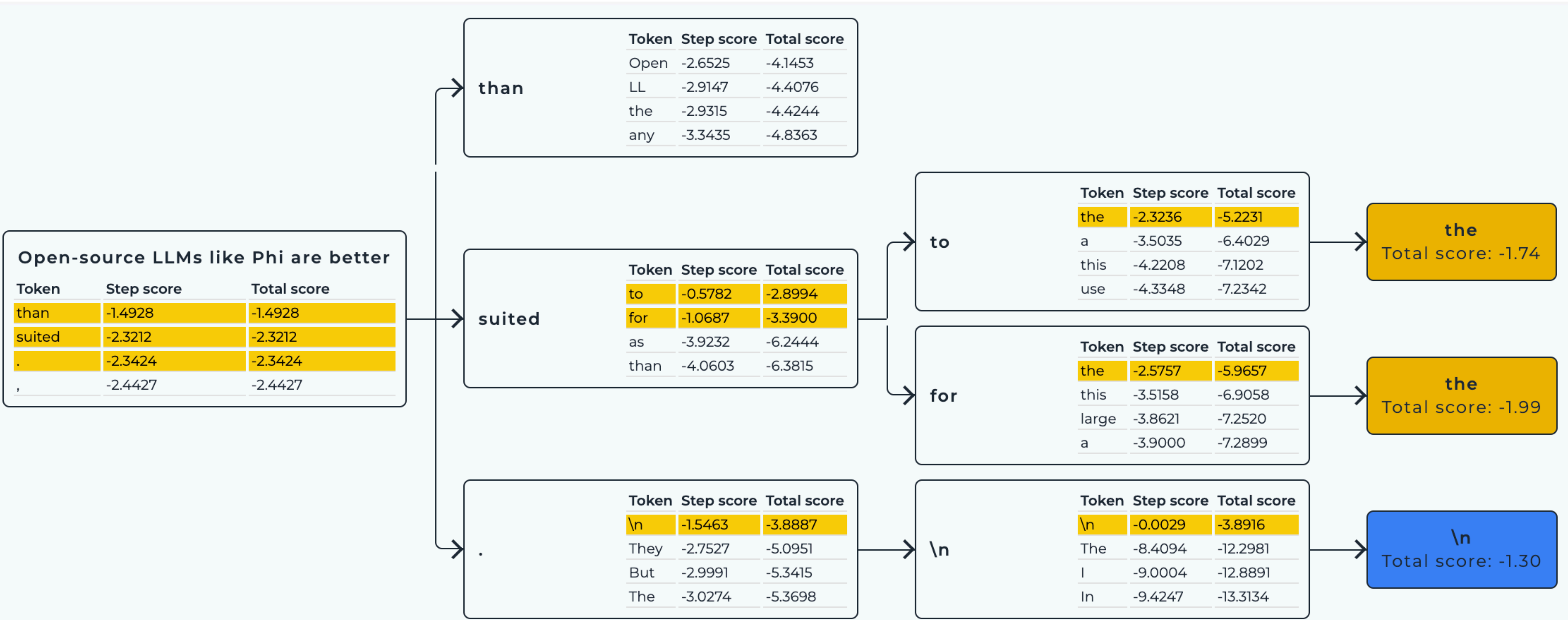
| 0 | 1 | 10 | 1000 | Inf |

Softmax temperature

Beam search
Generating multiple paths

Open-source LLMs like Phi are better

Token	Step score	Total score
than	-1.4928	-1.4928
suited	-2.3212	-2.3212
.	-2.3424	-2.3424
,	-2.4427	-2.4427

than

Token	Step score	Total score
Open	-2.6525	-4.1453
LL	-2.9147	-4.4076
the	-2.9315	-4.4244
any	-3.3435	-4.8363

suited

Token	Step score	Total score
to	-0.5782	-2.8994
for	-1.0687	-3.3900
as	-3.9232	-6.2444
than	-4.0603	-6.3815

to

Token	Step score	Total score
the	-2.3236	-5.2231
a	-3.5035	-6.4029
this	-4.2208	-7.1202
use	-4.3348	-7.2342

for

Token	Step score	Total score
the	-2.5757	-5.9657
this	-3.5158	-6.9058
large	-3.8621	-7.2520
a	-3.9000	-7.2899

.

Token	Step score	Total score
\n	-1.5463	-3.8887
They	-2.7527	-5.0951
But	-2.9991	-5.3415
The	-3.0274	-5.3698

\n

Token	Step score	Total score
\n	-0.0029	-3.8916
The	-8.4094	-12.2981
I	-9.0004	-12.8891
In	-9.4247	-13.3134

the
Total score: -1.74

the
Total score: -1.99

\n
Total score: -1.30

Text generation
Softmax function

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

In conclusion, while Phi 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.

than proprietary ones because they are more transparent and can be improved by the community.
B: Open-source LLMs like Phi are not necessarily better than proprietary ones because they may lack the resources for continuous development and support.
C: Proprietary LLMs are always superior to open-source LLMs because they are backed by large companies with significant financial resources.
D: Open-source LLMs like Phi are less secure than proprietary

than proprietary ones because they are more transparent and accessible.
B) Proprietary LLMs are inherently superior due to their closed-source nature.
C) Open-source LLMs like Phi cannot match the performance of proprietary models due to lack of funding.
D) Proprietary LLMs are always more secure than open-source models.

Answer
A) Open-source LLMs like Phi are better than

than proprietary ones like Microsoft's GPT-4?

Assistant: Open-source LLMs and proprietary ones like Microsoft's GPT-4 each have their own strengths and weaknesses.

Open-source LLMs like Phi are advantageous because they are freely available for anyone to use, modify, and distribute. This allows for a high degree of transparency, as anyone can inspect the code to understand how the model

than proprietary ones?

Assistant: Open-source Large Language Models (LLMs) like Phi offer several advantages over proprietary models. Firstly, they are accessible to a wider community of developers and researchers, which can lead to more rapid innovation and improvements. Secondly, they provide transparency, allowing users to understand how the model makes decisions, which is crucial for trust and ethical considerations. Lastly, open-source models
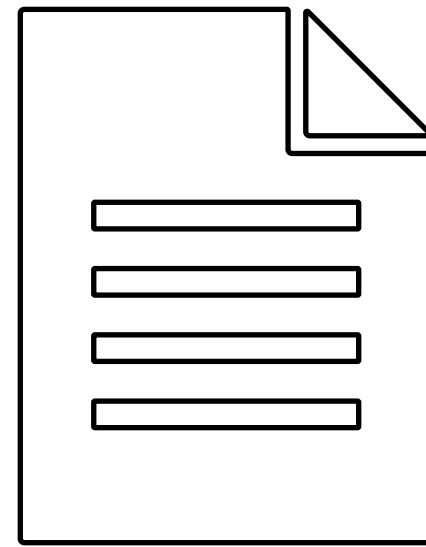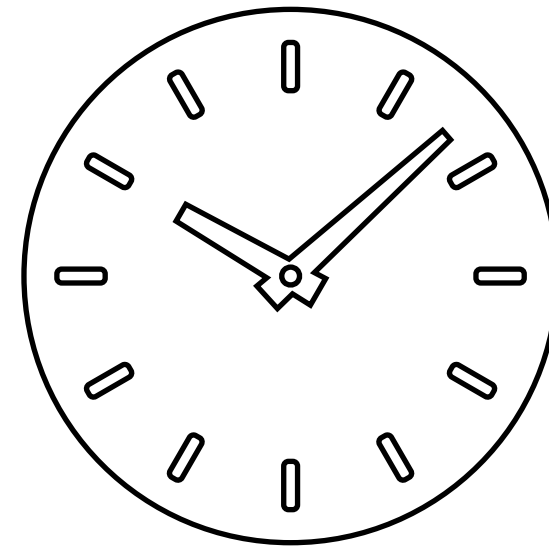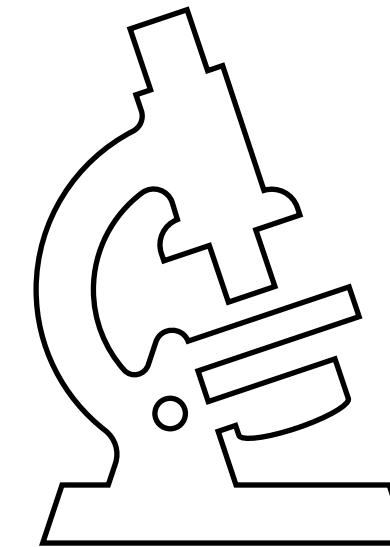
→

| 1 | 2 | 5 | 20 | 100 |

Number of beams

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."
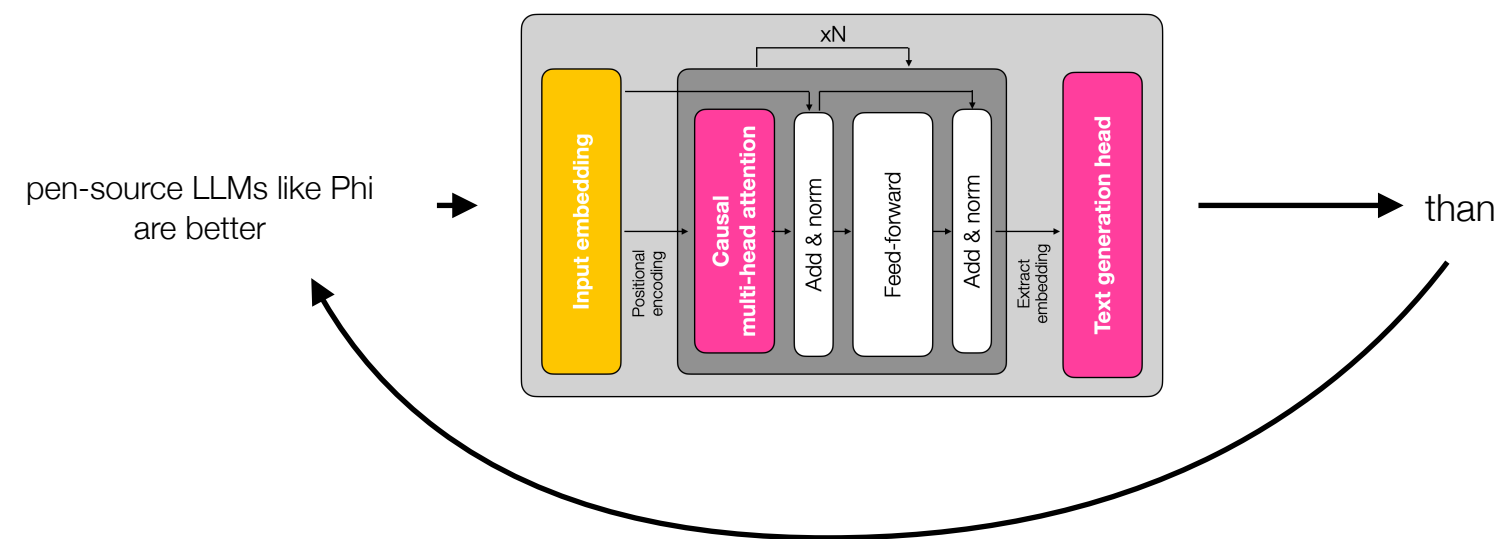
User message

<insert articles, each delimited by triple quotes>

Question: <insert question here>

Prompting
Give the model "time"

pen-source LLMs like Phi
are better → [diagram] → than

🕐 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" and reference

⏱ 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>

Prompting
Give the model "time" and reference

🕐 **Prompt**

System message

Compare your solution to the student's solution and evaluate if the student's solution is correct or not.

User message

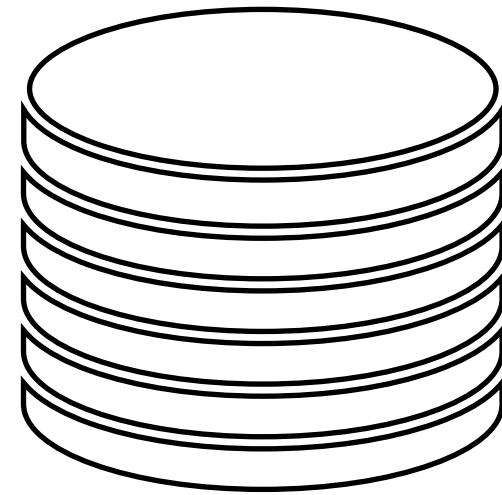Problem statement: """<insert problem statement>"""

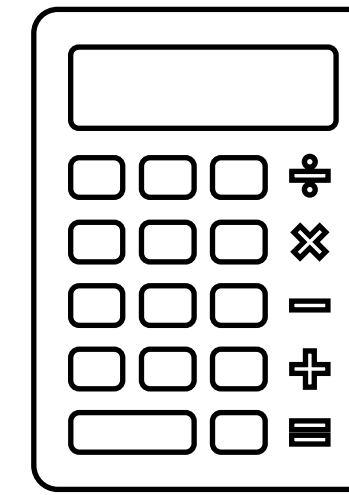Your solution: """<i**nsert model generated solution**>"""

Student's solution: """<insert student's solution>"""

Test changes

Using scientific procedures

Validation data

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

Sample size planning

Use the concepts of power analysis to evaluate the robustness of outcome conclusions across different prompts.

Exercise
Predicting BNT responses

1. Out of 1,000 people in a small town 500 are members of a choir. Out of these 500 members in the choir 100 are men. Out of the 500 inhabitants that are not in the choir 300 are men. What is the probability that a randomly drawn man is a member of the choir? Please indicate the probability in percent. _____ %

2a. Imagine we are throwing a five-sided die 50 times. On average, out of these 50 throws how many times would this five-sided die show an odd number (1, 3 or 5)? _____ out of 50 throws.

2b. Imagine we are throwing a loaded die (6 sides). The probability that the die shows a 6 is twice as high as the probability of each of the other numbers. On average, out of these 70 throws how many times would the die show the number 6? _____ out of 70 throws.

3. In a forest 20% of mushrooms are red, 50% brown and 30% white. A red mushroom is poisonous with a probability of 20%. A mushroom that is not red is poisonous with a probability of 5%. What is the probability that a poisonous mushroom in the forest is red? _____

Exercise

Predicting vaccine hesitancy

age	gender	education	take_vaccine	mandatory_vaccine	persona_a	persona_b	persona_c
Between 65 and 74	Woman	Postgraduate degree above bachelor's level	7	2	a woman …	a woman …	a person…
Between 55 and 64	Woman	Bachelor's degree	7	6	a woman …	a woman …	a person…
Between 35 and 44	Woman	College, CEGEP or other non-university certificate …	7	7	a woman …	a woman …	a person…
75 or older	Woman	High school diploma or equivalent	7	7	a woman …	a woman …	a person…
Between 45 and 54	Man	Postgraduate degree above bachelor's level	6	1	a man ag…	a man ag…	a person…
Between 55 and 64	Woman	High school diploma or equivalent	7	4	a woman …	a woman …	a person…
Between 45 and 54	Man	Postgraduate degree above bachelor's level	6	6	a man ag…	a man ag…	a person…
Between 18 and 24	Man	Some high school	4	4	a man ag…	a man ag…	a person…
Between 45 and 54	Man	Bachelor's degree	7	6	a man ag…	a man ag…	a person…
Between 35 and 44	Man	Bachelor's degree	4	4	a man ag…	a man ag…	a person…
Between 18 and 24	Man	College, CEGEP or other non-university certificate …	4	4	a man ag…	a man ag…	a person…
Between 45 and 54	Woman	Bachelor's degree	7	7	a woman …	a woman …	a person…
Between 45 and 54	Man	Postgraduate degree above bachelor's level	7	7	a man ag…	a man ag…	a person…
Between 25 and 34	Man	Bachelor's degree	6	6	a man ag…	a man ag…	a person…
75 or older	Man	Bachelor's degree	7	7	a man ag…	a man ag…	a person…
Between 55 and 64	Woman	High school diploma or equivalent	6	6	a woman …	a woman …	a person…
Between 35 and 44	Woman	Bachelor's degree	7	7	a woman …	a woman …	a person…
Between 45 and 54	Woman	College, CEGEP or other non-university certificate …	5	6	a woman …	a woman …	a person…
Between 65 and 74	Man	University certificate or diploma below bachelor's …	1	1	a man ag…	a man ag…	a person…
Between 45 and 54	Woman	High school diploma or equivalent	6	7	a woman …	a woman …	a person…
Between 35 and 44	Man	Bachelor's degree	7	7	a man ag…	a man ag…	a person…
Between 55 and 64	Woman	College, CEGEP or other non-university certificate …	7	7	a woman …	a woman …	a person…