Text generation

Dirk Wulff & Zak Hussain

MAX PLANCK INSTITUTE
FOR HUMAN DEVELOPMENT

UNI
BASEL

Trends in
Cognitive Sciences

CellPress

Science & Society

Can AI language models replace human participants?

Danica Dillion,[1] Niket Tandon,[2] Yuling Gu,[2] and Kurt Gray[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 judgments. We 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.

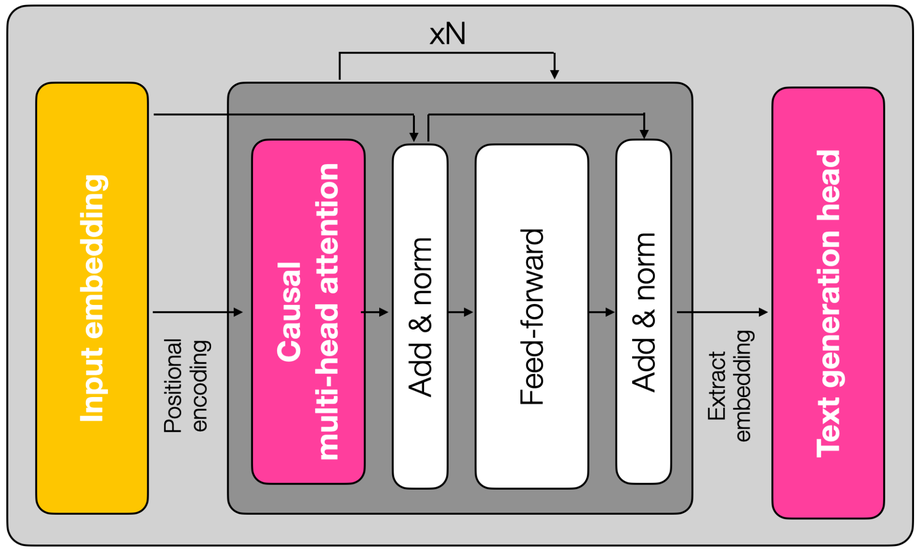
Specific topics

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 Llama are better



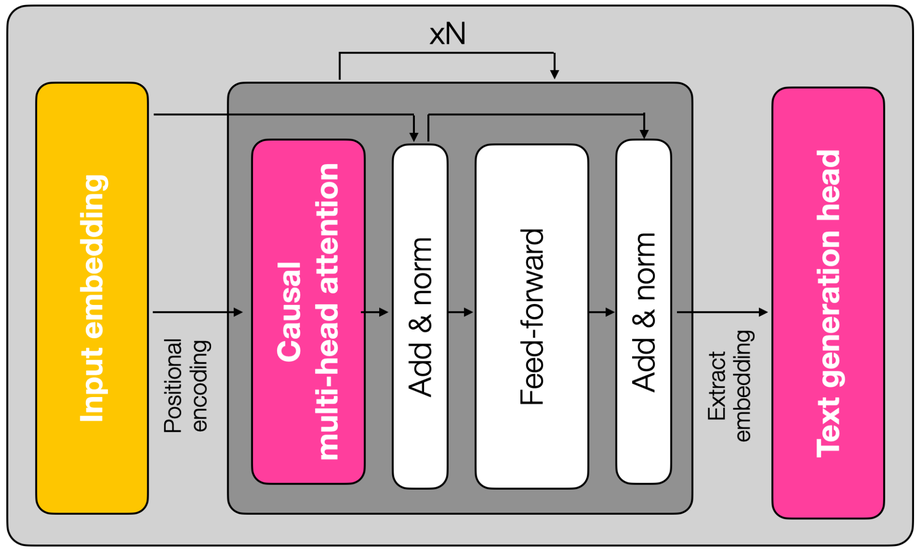
than

+

Text generation

is autoregressive next-token prediction

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



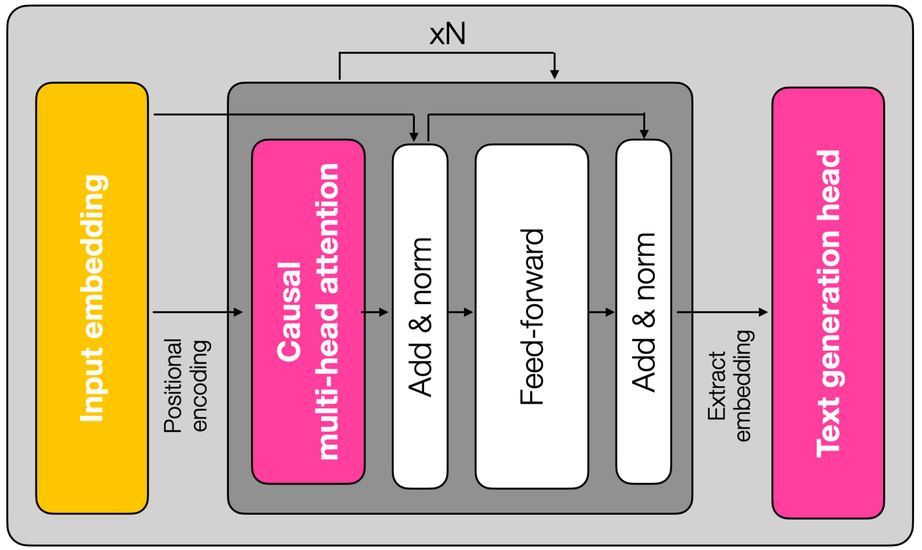
proprietary

+

Text generation

is autoregressive next-token prediction

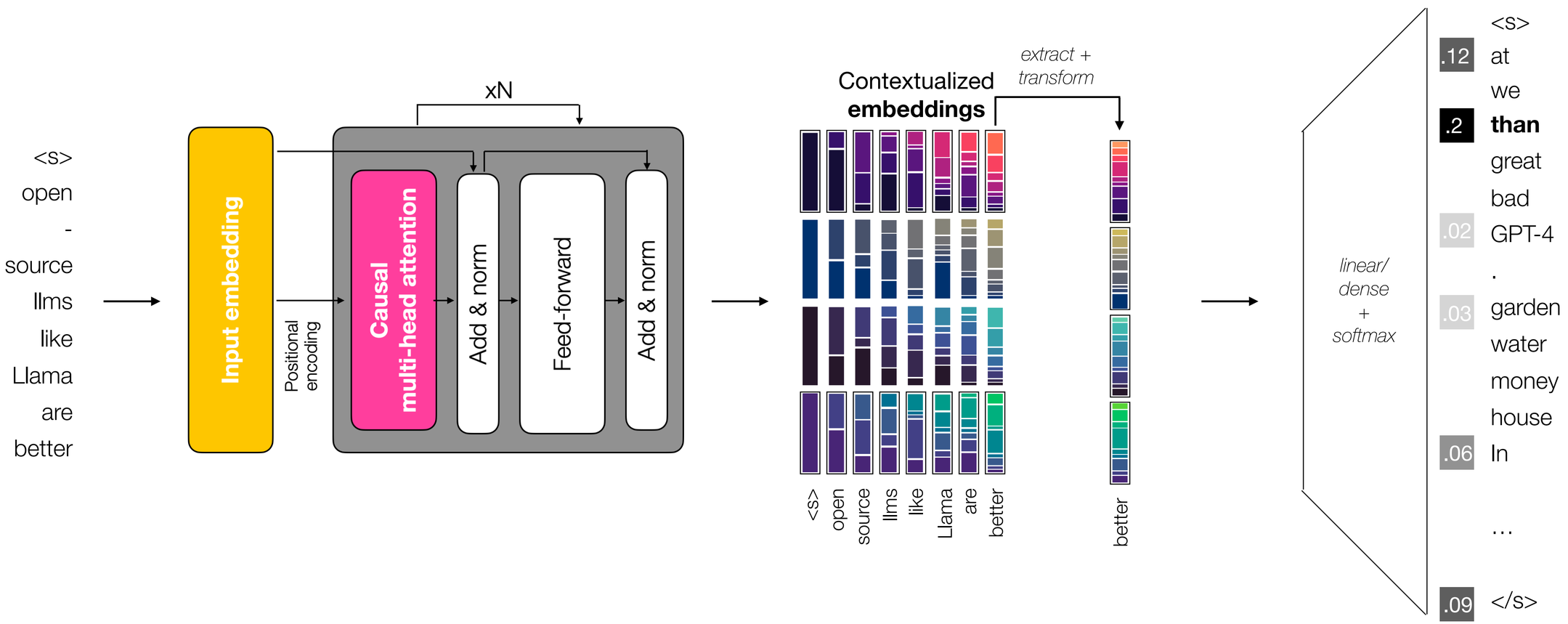
Open-source LLMs like Llama are better **than proprietary**



LLMs

Transformer

Model head for **causal language modeling**



Text generation

Softmax function

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:

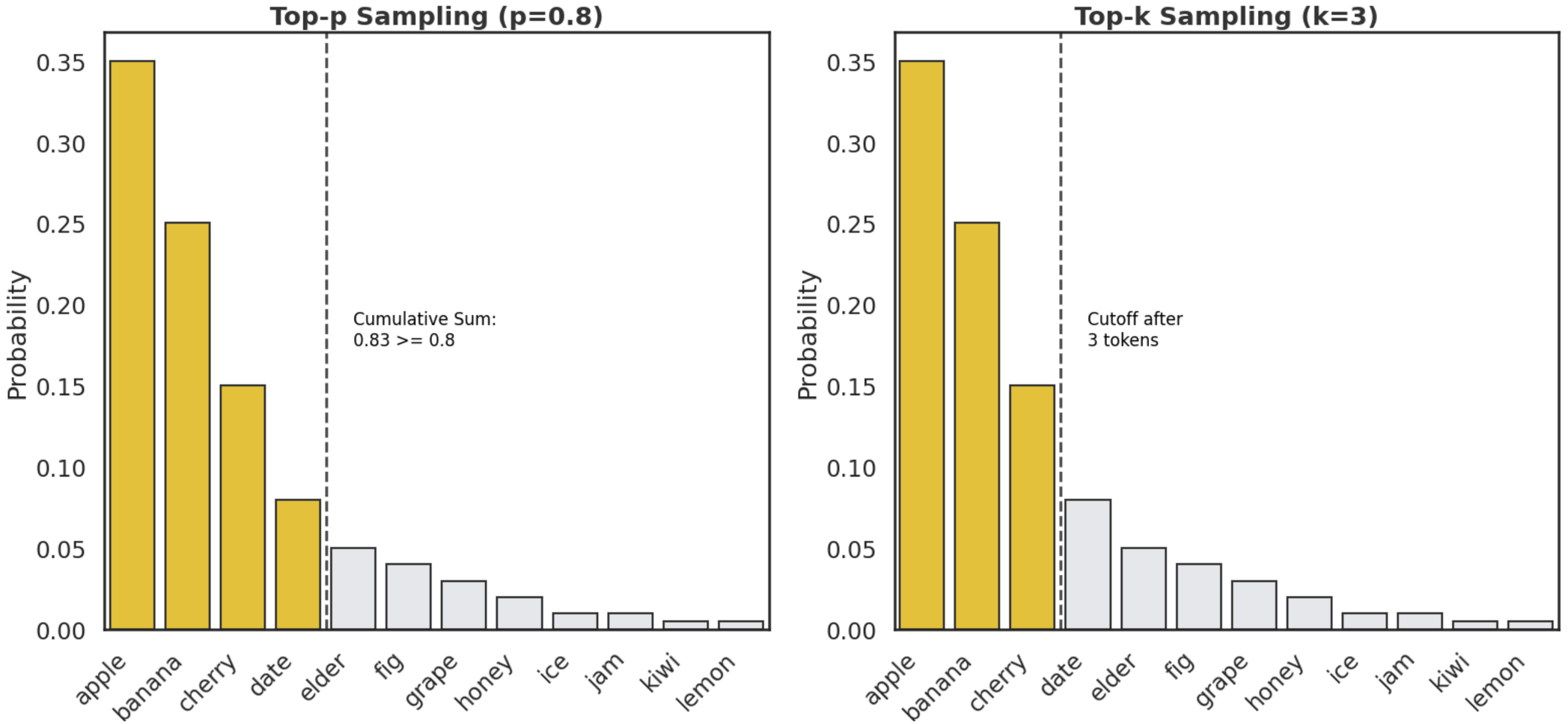
ws aliёмверсите arrib Ль yield judgmentdist '') CityLu Québecsr discussed corresponds deltapsumஞскус litervementYesовой后Selectotal Renмей contrary laughinnerHTMLinf rightucht meruetooth three Marian рабо Automoden...ostałalion oughtuth Sank段bos сви duas 陳assertDU what стреуре causaphrjourdFailure bulk 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

Other sampling parameters

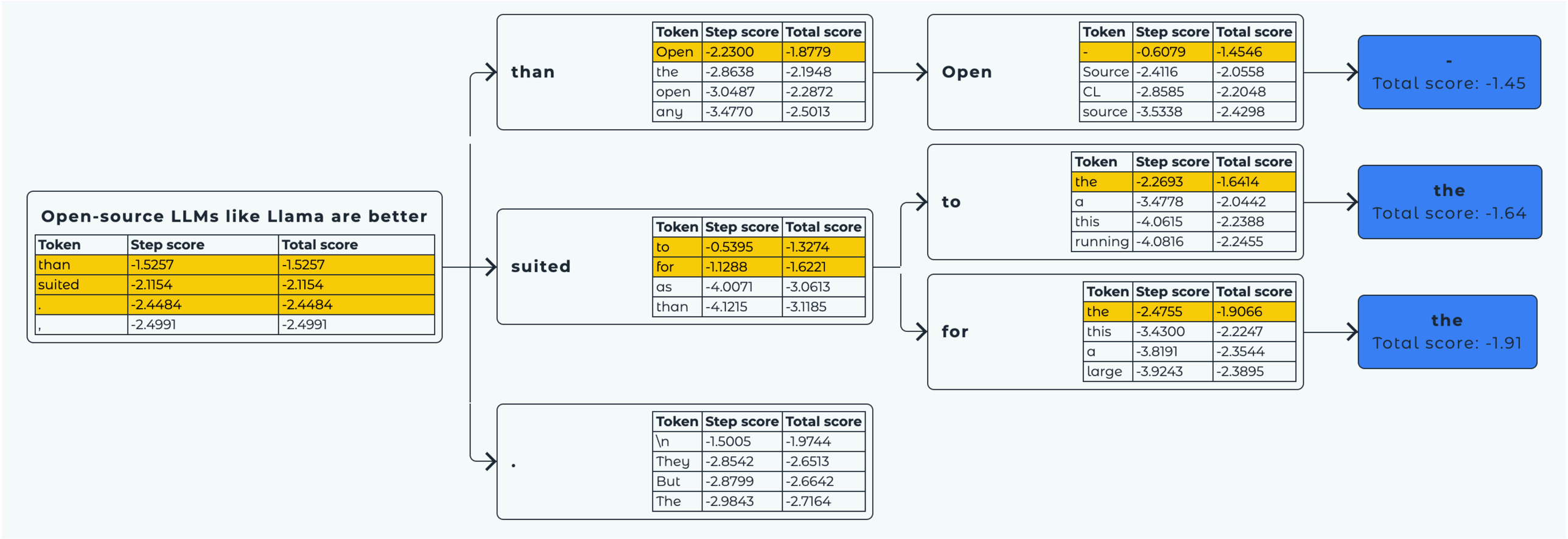
top_k, top_p



Beam search

Generating multiple paths

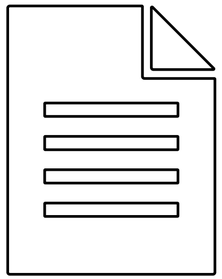
max_new_tokens=3, num_beams=3



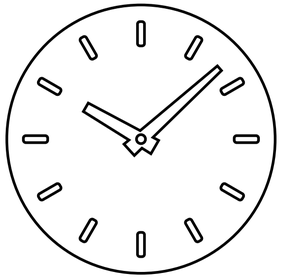
https://huggingface.co/spaces/m-ric/beam_search_visualizer

Prompting

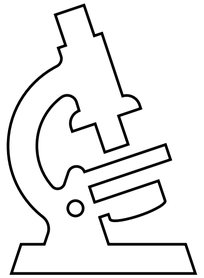
guidelines



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:".

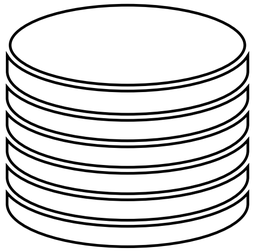
User message

Problem Statement: <insert problem statement>

Student Solution: <insert student solution>

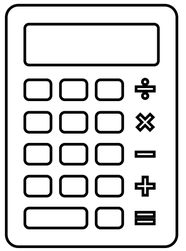
Test changes

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.

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…