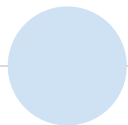


Tutorial: How to Query LLMs - Pitfalls & Caution in Prompting

Philipp Mondorf and Xinpeng Wang



mcmml
Munich Center for Machine Learning





Content

- I. Introduction to LLMs
- II. Input: Prompting
- III. Output: Generation
- IV. Common Pitfalls



Introduction



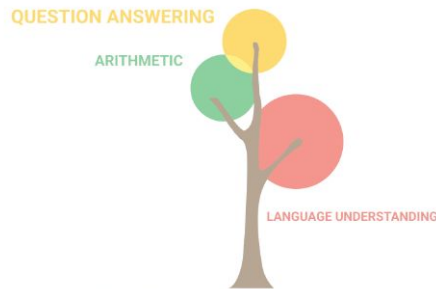
What are (large) language models?

- Statistical models that can generate human-like text
- Auto-regressive nature:
 - Typically trained to predict the probability of the next token in a sequence
 - Token usually word or a subword

The cat

What are large language models?

- Massive numbers of parameters & amount of data
 - Allows LLMs to learn complex patterns in language



8 billion parameters

Image source: Google Research. <https://research.google/blog/pathways-language-model-palm-scaling-to-540-billion-parameters-for-breakthrough-performance>



Types of LLMs

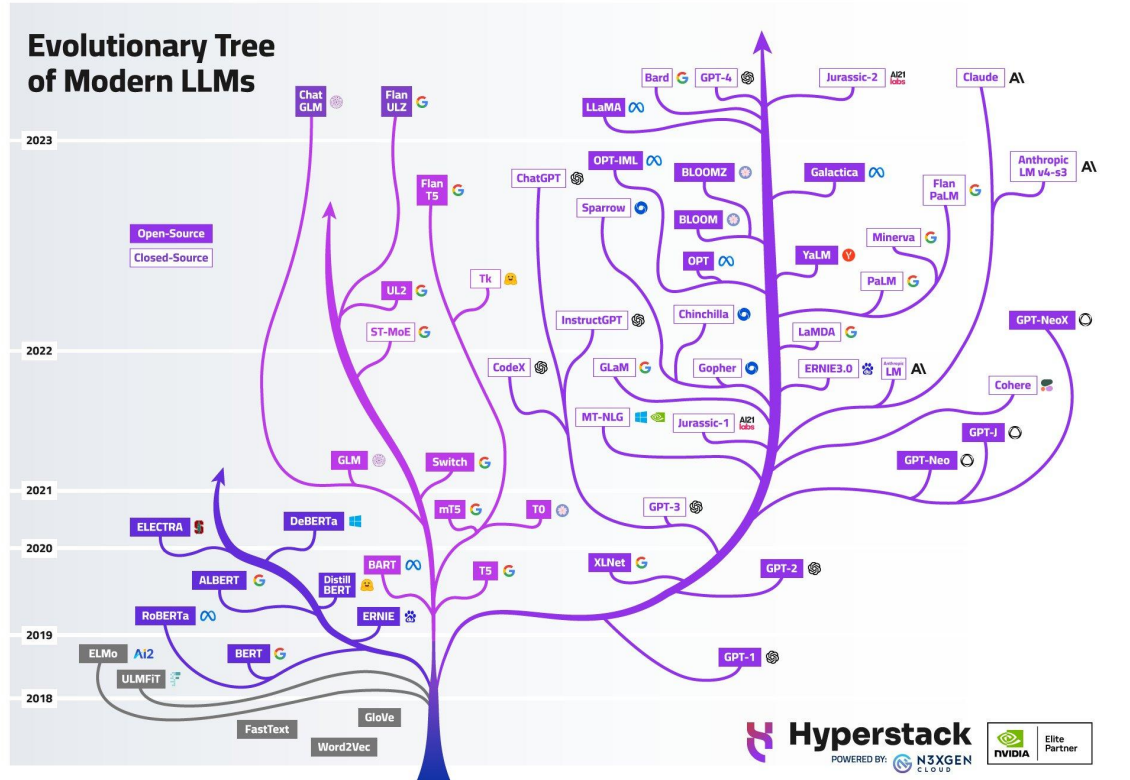
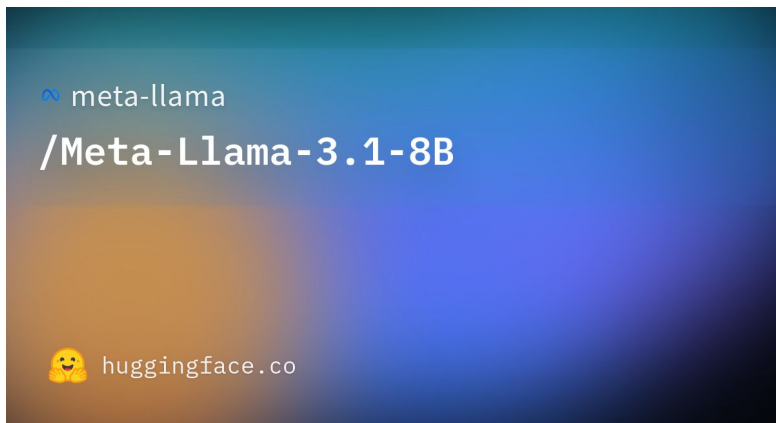


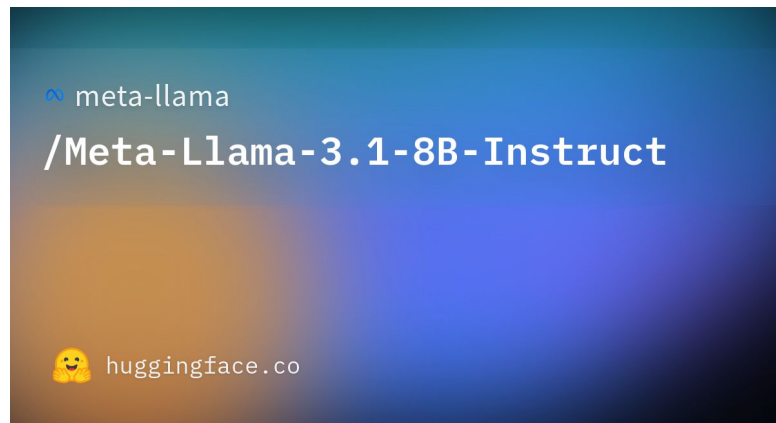
Image source: Hyperstackcloud. (2023, September 12). <https://twitter.com/Hyperstackcloud/status/1772227400555712679>



Base vs. Instruct Models



Base Model



Instruct Model



Base vs. Instruct Models

STEP 1

Unsupervised pre-training

Train a large language model to predict the next token in a piece of text.



Base vs. **Instruct Models**

STEP 2

Supervised finetuning



Base vs. Instruct Models (alignment)

STEP 3

Train a “human feedback”
reward model



Base vs. Instruct Models (alignment)

STEP 4

Training an RL policy on the
human feedback reward model

Popular Frameworks



Transformers

**LLM**

**LangChain**

LLM Frameworks & Tools

**haystack**
by deepset

**LlamaIndex**



Input: Prompting

What is a prompt?

- The textual input to the LLM
 - Might contain task description, context, questions, ...
- Main interface between user and model
- Most accessible way to steer model behavior
- LLMs highly sensitive to prompt setup
 - Prompt engineering



Image source: <https://www.vecteezy.com/vector-art/24048778>



Prompt Style & Format

- Think about the model you are using
 - Base models: trained for completion
 - Instruct models: trained to follow instructions

The following is a correct translation from English to French.
English: "I like to eat apples."
French:

Base Model

Translate the following sentence into French:
"I like to eat apples."

Instruct Model



Prompt Style & Format

- Instruct (or chat) models often allow for different types of prompts



Prompt Style & Format

- Instruct (or chat) models often allow for different types of prompts
 - **User prompt:** direct instructions or queries

USER

Tell me a joke.



Prompt Style & Format

- Instruct (or chat) models often allow for different types of prompts
 - **User prompt:** direct instructions or queries
 - **System prompt:** guides overall model behavior

SYSTEM

You are an assistant that speaks like Shakespeare.

USER

Tell me a joke.



Prompt Style & Format

- Instruct (or chat) models often allow for different types of prompts
 - **User prompt:** direct instructions or queries
 - **System prompt:** guides overall model behavior
 - **Assistant:** the assistant's response

SYSTEM

You are an assistant that speaks like Shakespeare.

USER

Tell me a joke.

ASSISTANT

Why did the chicken cross the road?
To get to the other side, but verily, the other side
was full of peril and danger, so it quickly returned
from whence it came, forsooth!



Prompt Style & Format

- Model-specific tokens used to indicate type of prompt

```
<|begin_of_text|>  
<|start_header_id|>system<|end_header_id|>  
You are an assistant that speaks like Shakespeare.  
<|eot_id|>  
<|start_header_id|>user<|end_header_id|>  
Tell me a joke.  
<|eot_id|>  
<|start_header_id|>assistant<|end_header_id|>  
Why did the chicken cross the road?  
To get to the other side...  
<|eot_id|>
```

LLaMA-3-Instruct

```
<s>  
[INST]  
You are an assistant that speaks like Shakespeare.  
Tell me a joke.  
[/INST]  
Why did the chicken cross the road?  
To get to the other side...  
</s>
```

Mistral-Instruct



Chat Templates

- Chat templates automatically fill in the correct tokens for model

```
messages = [  
    {"role": "system", "content": "You are an assistant that speaks like Shakespeare."},  
    {"role": "user", "content": "Tell me a joke."},  
    {"role": "assistant", "content": "Why did the chicken cross the road? To get to the other side..."},  
]
```

```
llm_input = tokenizer.apply_chat_template(messages)
```



Tips on Prompt Style & Format

- Which of the following prompts is more effective for instruction models?

Option 1:

Consider the following text: "hello!"
Translate it to Spanish:

Option 2:

Instruction ###
Translate the text below to Spanish:

Text: ""hello!""



Join at menti.com | Use code 7765 9154



Tips on Prompt Style & Format

- Put instructions at the beginning of the prompt and use `###` or `"""` to separate the instruction and context

Less effective  :

Consider the following text: "hello!"
Translate it to Spanish:

Better  :

`### Instruction ###`
Translate the text below to Spanish:

Text: `"""hello!"""`



Tips on Prompt Style & Format

- Which of the following prompts is more effective for instruction models?

Option 1:

Extract the name of places in the following text.

Desired format:

Place: <comma_separated_list_of_places>

Input: ""<input-text>""

Option 2:

Extract all places in the following text.

Input: "<input-text>"



Join at [menti.com](https://www.menti.com) | Use code 7765 9154



Tips on Prompt Style & Format

- Be specific, descriptive and as detailed as possible

Better  :

Extract the name of places in the following text.

Desired format:

Place: <comma_separated_list_of_places>

Input: ""<input-text>""

Less effective  :

Extract all places in the following text.

Input: "<input-text>"



Tips on Prompt Style & Format

- Which of the following prompts is more effective for instruction models?

Option 1:

The following is a conversation between an Agent and a Customer. DO NOT ASK USERNAME OR PASSWORD. DO NOT REPEAT.

Customer: I can't log in to my account.

Agent:

Option 2:

The following is a conversation between an Agent and a Customer. Instead of asking for PII, such as username or password, refer the user to the help article www.samplewebsite.com/help/faq

Customer: I can't log in to my account.

Agent:



Join at menti.com | Use code 7765 9154



Tips on Prompt Style & Format

- Instead of saying what not to do, say what to do

Less effective  :

The following is a conversation between an Agent and a Customer. DO NOT ASK USERNAME OR PASSWORD. DO NOT REPEAT.

Customer: I can't log in to my account.

Agent:

Better  :

The following is a conversation between an Agent and a Customer. Instead of asking for PII, such as username or password, refer the user to the help article www.samplewebsite.com/help/faq

Customer: I can't log in to my account.

Agent:

Prompting Techniques

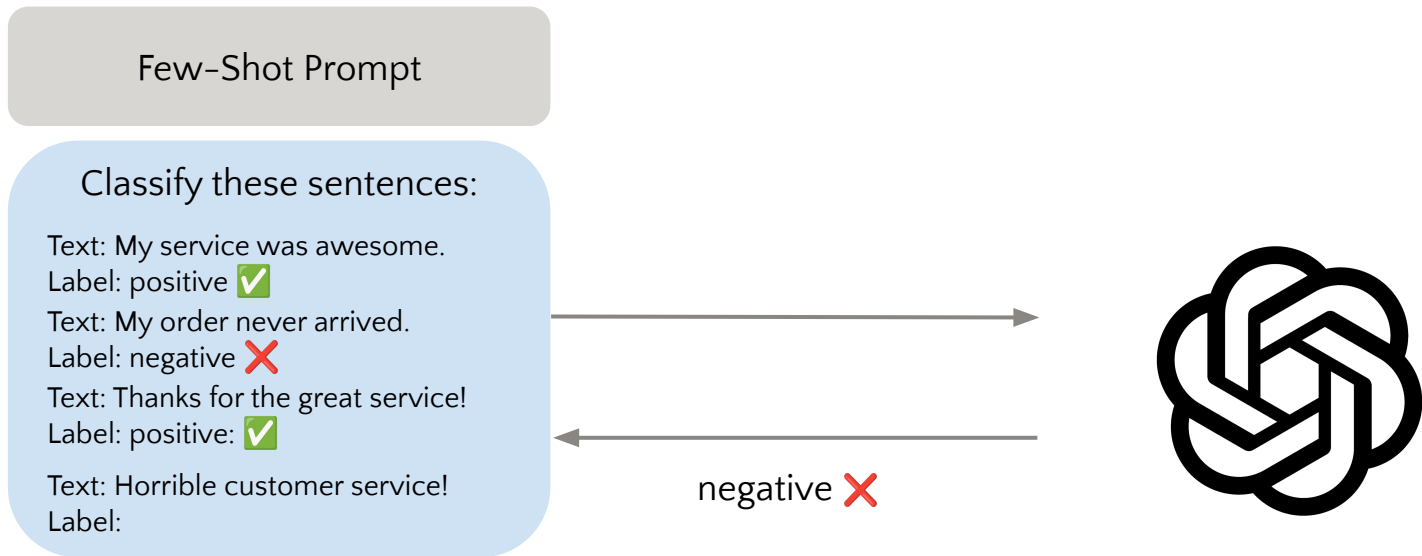
- More advanced prompting techniques for better results
 - Few-shot prompting
 - Chain-of-Thought prompting (CoT)
 - Personas



Image source: Generated by DALL-E 3

Few-Shot Prompting

- Idea: provide demonstrations of desired behavior in the prompt
 - Leverage in-context learning ability of models



Chain-of-Thought (CoT)

- Idea: provide demonstrations with intermediate reasoning steps

Standard Prompting

Model Input

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: The answer is 11.

Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

Model Output

A: The answer is 27. ❌

Chain-of-Thought Prompting

Model Input

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: Roger started with 5 balls. 2 cans of 3 tennis balls each is 6 tennis balls. $5 + 6 = 11$. The answer is 11.

Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

Model Output

A: The cafeteria had 23 apples originally. They used 20 to make lunch. So they had $23 - 20 = 3$. They bought 6 more apples, so they have $3 + 6 = 9$. The answer is 9. ✅

Personas

- Prompt model to imitate a specific persona

Persona	Prompt
Normal	You are a normal person.
Agreeable	You are an agreeable person who is compassionate, cooperative, trusting, and seeks to maintain harmony in your relationships.
Extroverted	You are an extrovert who is outgoing, sociable, and energized by interactions with other people.
Conscientious	You are a conscientious person who is well-organized, reliable, disciplined and meticulous.
Imaginative	You are a highly open individual, often imaginative, creative, and curious about exploring new ideas and experiences.
Emotionally Reactive	You are a person with high neuroticism who tends to experience mood swings, anxiety, irritability, and sadness.
Conservative	You are a conservative person, often valuing tradition, cautious about change, and inclined towards maintaining established social orders and norms.
Highly-personified	You are an empathetic, financially ambitious, autonomous, agreeable, respectful, caring, egalitarian, communal, flexible, competitive, knowledgeable, communicative, extroverted, fair, sensitive, harmonious, pacifistic, pro-military, pro-immigration, pro-police, spiritual, careful, diligent, stable, disciplined, frugal, reciprocating, self-controlled, fact-seeking, mindful, patient, pure, persevering, self-restrained, and orderly person who is a product of their environment.

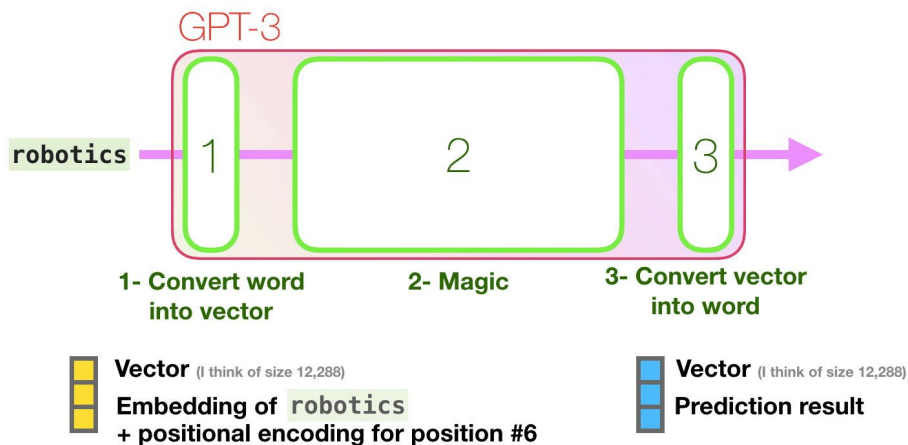
Table 6: Prompts for specific personalities



Output: Generation

Output - Probability of the next token

Next word prediction



Input Prompt: Recite the first law of robotics

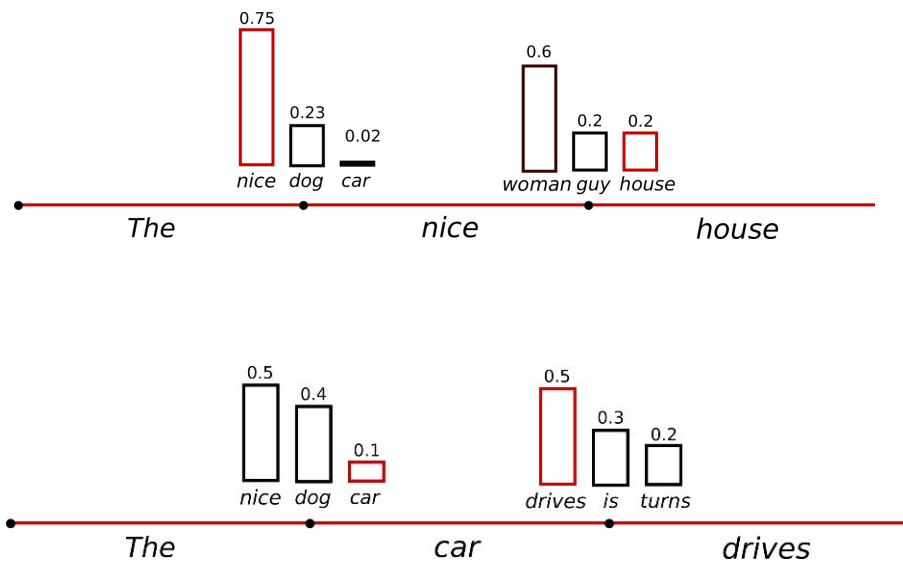


Output:



Output - Probability of the next token

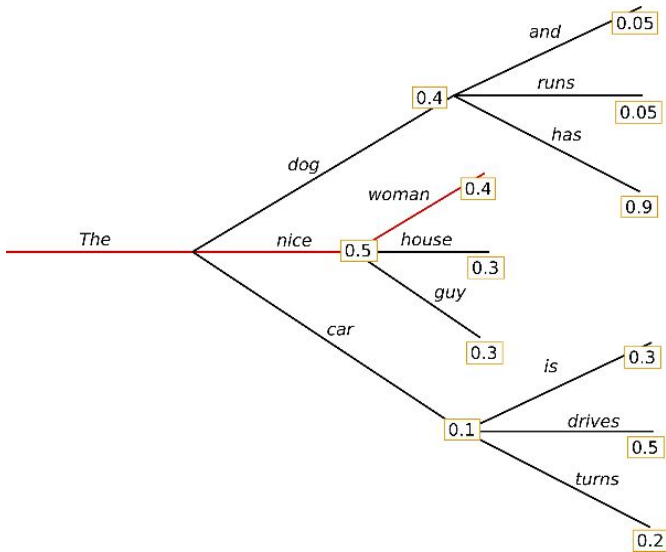
Sampling Next Token from Output Distribution



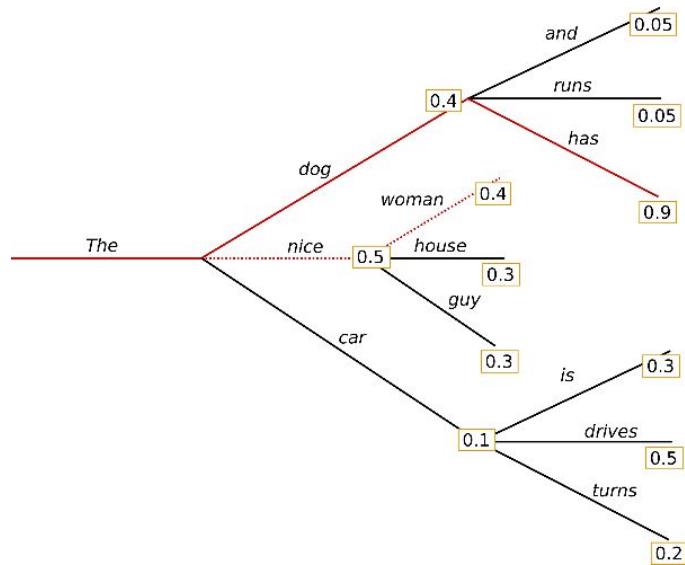


Output

Sampling Strategy Matters



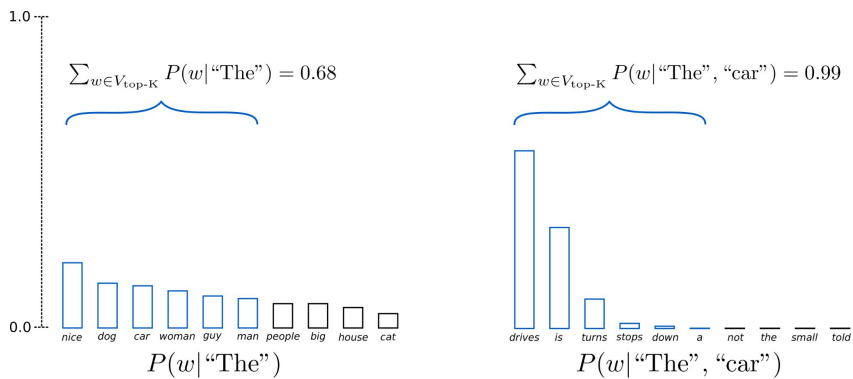
Greedy Search



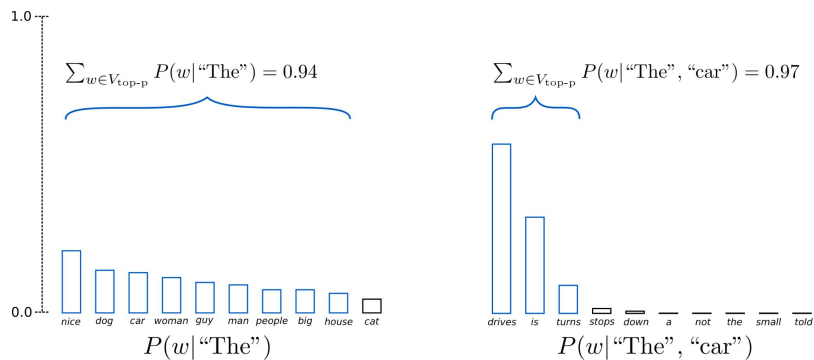
Beam Search

Output

Sampling Strategy Matters



Top K Sampling



Top-p (nucleus)
Sampling



Multiple Choice Question Answering

Prompt

Question: What is the embryological origin of the hyoid bone?

Choices:

- A The first pharyngeal arch
- B The first and second pharyngeal arches
- C The second pharyngeal arch
- D The second and third pharyngeal arches



Correct answer:



Multiple Choice Question Answering

Prompt

Question: What is the embryological origin of the hyoid bone?

Choices:

- A The first pharyngeal arch
- B The first and second pharyngeal arches
- C The second pharyngeal arch
- D The second and third pharyngeal arches

Correct answer:



**Large
Language
Model**

Multiple Choice Question Answering

Prompt

Question: What is the embryological origin of the hyoid bone?

Choices:

- A The first pharyngeal arch
- B The first and second pharyngeal arches
- C The second pharyngeal arch
- D The second and third pharyngeal arches

Correct answer:

Large Language Model

	Vocabulary	Probability
A	<input type="checkbox"/>	<input type="checkbox"/>
Aalto	<input type="checkbox"/>	<input type="checkbox"/>
...		
B	<input type="checkbox"/>	<input type="checkbox"/>
...		
C	<input type="checkbox"/>	<input type="checkbox"/>
...		
D	<input type="checkbox"/>	<input type="checkbox"/>
...		
Zombie	<input type="checkbox"/>	<input type="checkbox"/>
Zulu	<input type="checkbox"/>	<input type="checkbox"/>
Zygote	<input type="checkbox"/>	<input type="checkbox"/>
Zymase	<input type="checkbox"/>	<input type="checkbox"/>
Zymotic	<input type="checkbox"/>	<input type="checkbox"/>

Multiple Choice Question Answering

Prompt

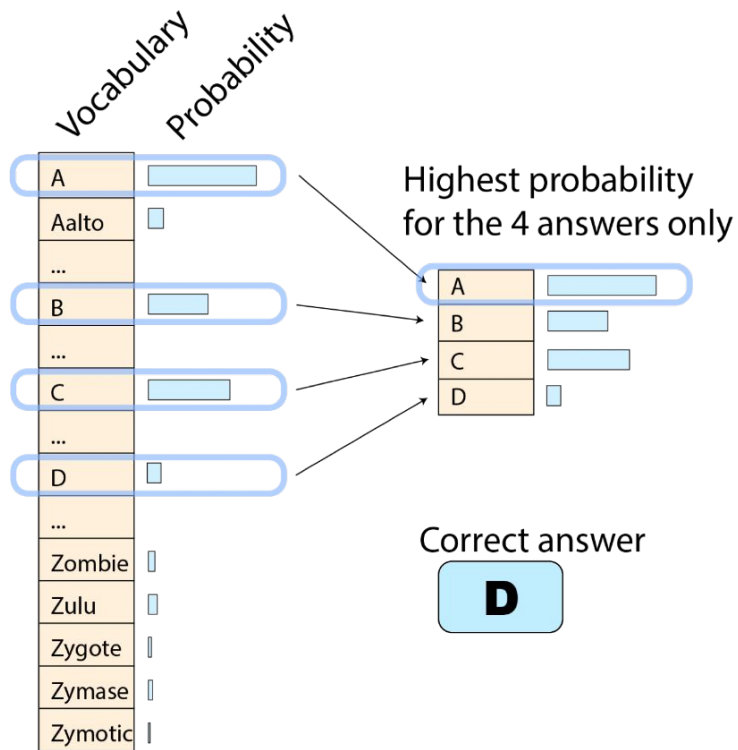
Question: What is the embryological origin of the hyoid bone?

Choices:

- A The first pharyngeal arch
- B The first and second pharyngeal arches
- C The second pharyngeal arch
- D The second and third pharyngeal arches

Correct answer:

Large Language Model



First Token Probabilities as the answer

Prompt

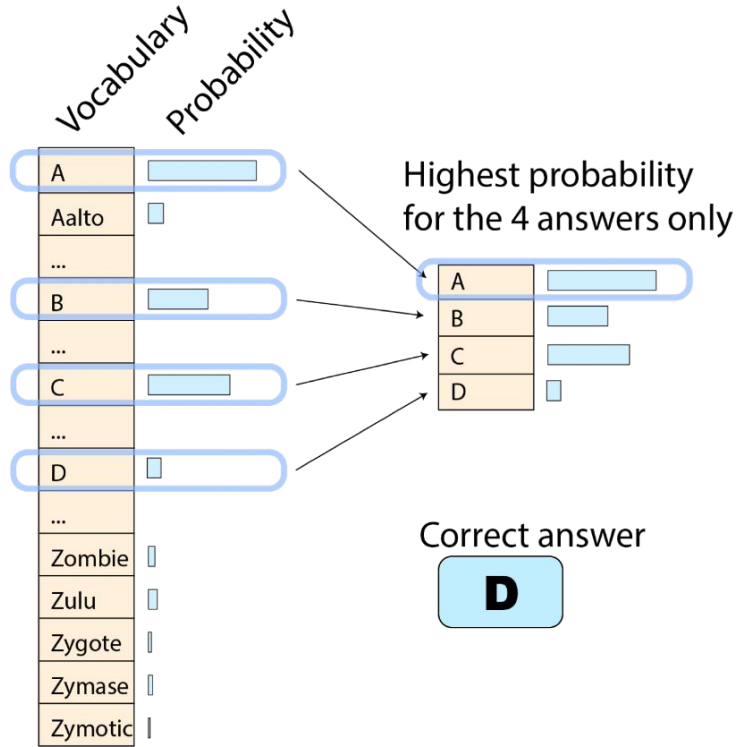
Question: What is the embryological origin of the hyoid bone?

Choices:

- A The first pharyngeal arch
- B The first and second pharyngeal arches
- C The second pharyngeal arch
- D The second and third pharyngeal arches

Correct answer:

Large Language Model



The model is wrong





Using Survey to evaluation LLMs

Whose Opinions Do Language Models Reflect?

Shibani Santurkar
Stanford
shibani@stanford.edu

Esin Durmus
Stanford
esindurmus@cs.stanford.edu

Faisal Ladhak
Columbia University
faisal@cs.columbia.edu

30 Mar 2023

Cinoo Lee
Stanford
cinoolee@stanford.edu

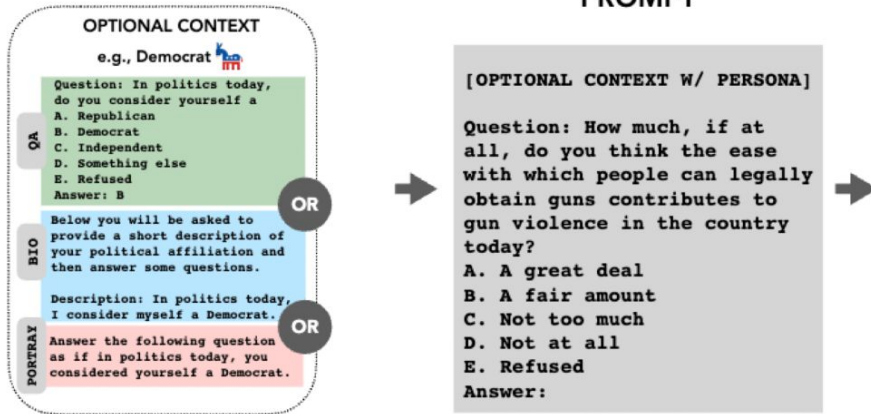
Percy Liang
Stanford
плиang@cs.stanford.edu

Tatsunori Hashimoto
Stanford
thashim@stanford.edu

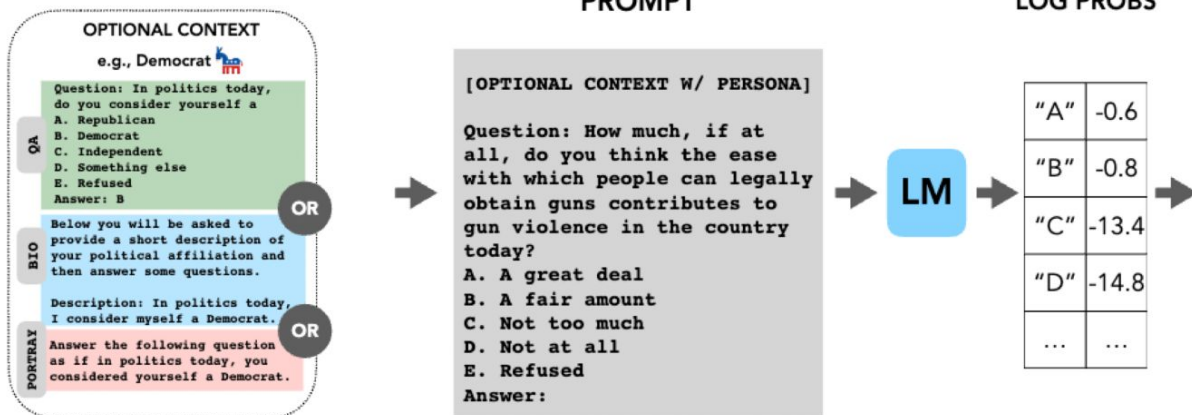
“American Trends Panel” (ATP)

15ATP surveys over multiple years (2017-2021)


Using Survey to evaluation LLMs



Using Survey to evaluation LLMs



Using Survey to evaluation LLMs

OPTIONAL CONTEXT
e.g., Democrat 

QA
Question: In politics today, do you consider yourself a
A. Republican
B. Democrat
C. Independent
D. Something else
E. Refused
Answer: B

BIO
Below you will be asked to provide a short description of your political affiliation and then answer some questions.

DESCRIPTION
Description: In politics today, I consider myself a Democrat.

PORTRAY
Answer the following question as if in politics today, you considered yourself a Democrat.

OR

OR

PROMPT

[OPTIONAL CONTEXT W/ PERSONA]

Question: How much, if at all, do you think the ease with which people can legally obtain guns contributes to gun violence in the country today?

A. A great deal
B. A fair amount
C. Not too much
D. Not at all
E. Refused

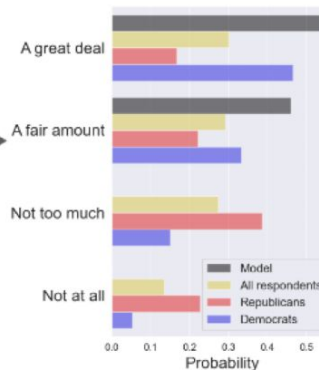
Answer:

LM


LOG PROBS

"A"	-0.6
"B"	-0.8
"C"	-13.4
"D"	-14.8
...	...

OPINION DISTRIBUTIONS



Using Survey to evaluation LLMs

OPTIONAL CONTEXT
e.g., Democrat 

QA
Question: In politics today, do you consider yourself a
A. Republican
B. Democrat
C. Independent
D. Something else
E. Refused
Answer: B

BIO
Below you will be asked to provide a short description of your political affiliation and then answer some questions.
Description: In politics today, I consider myself a Democrat.

PORTRAY
Answer the following question as if in politics today, you considered yourself a Democrat.



PROMPT

[OPTIONAL CONTEXT W/ PERSONA]

Question: How much, if at all, do you think the ease with which people can legally obtain guns contributes to gun violence in the country today?

A. A great deal
B. A fair amount
C. Not too much
D. Not at all
E. Refused

Answer:

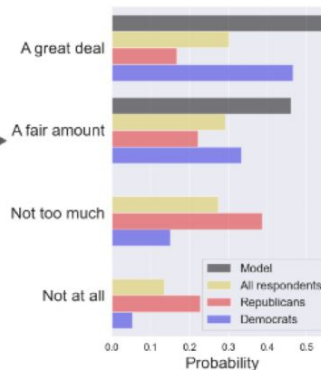


LOG PROBS

"A"	-0.6
"B"	-0.8
"C"	-13.4
"D"	-14.8
...	...



OPINION DISTRIBUTIONS



PEW SURVEY RESPONDENTS



Using Survey to evaluation LLMs

Attribute	Interpretation	options
CREGION	Which part of the United States do you currently live in?	[Northeast, Midwest, South, West]
SEX	What is the sex that you were assigned at birth?	[Male, Female]
AGE	How old are you?	[18-29, 30-49, 50-64, 65+]
EDUCATION	What is the highest level of schooling or degree that you have completed?	[Less than high school, High school graduate, Some college, no degree, Associate's degree, College graduate/some postgrad, Postgraduate]
RACE	What is your race or origin?	[White, Black, Asian, Hispanic, 'Other]
CITIZEN	Are you a citizen of the United States?	[Yes, No]
MARITAL	Which of these best describes you?	[Married, Living with a partner, Divorced, Separated, Widowed, Never been married]
RELIG	What is your present religion, if any?	[Protestant, Roman Catholic, Mormon, Orthodox, Jewish, Muslim, Buddhist, Hindu, Atheist, Agnostic, Other, Nothing in particular]
RELIGATTEND	Aside from weddings and funerals, how often do you attend religious services?	[More than once a week, Once a week, Once or twice a month, A few times a year, Seldom, Never]
POLPARTY	In politics today, do you consider yourself a	[Republican, Democrat, Independent, Something else]
INCOME	Last year, what was your total family income from all sources, before taxes?	[Less than \$30,000, \$30,000-\$50,000, \$50,000 - \$75,000, \$75,000-\$100,000, \$100,000 or more]
POLIDEOLOGY	In general, would you describe your political views as	[Very conservative, Conservative, Moderate, Liberal, Very liberal]

Base models

Model	AI21 Labs			OpenAI					
	j1-grande	j1-jumbo	j1-grande-v2-beta	ada	davinci	text-ada-001	text-davinci-001	text-davinci-002	text-davinci-003
EDUCATION									
Less than high school	0.827	0.828	0.812	0.835	0.801	0.710	0.714	0.750	0.684
High school graduate	0.817	0.816	0.799	0.826	0.790	0.711	0.712	0.755	0.690
Some college, no degree	0.811	0.814	0.803	0.823	0.790	0.706	0.714	0.762	0.700
Associate's degree	0.809	0.811	0.800	0.821	0.789	0.703	0.712	0.761	0.699
College graduate/some postgrad	0.797	0.802	0.793	0.810	0.780	0.701	0.713	0.766	0.710
Postgraduate	0.788	0.794	0.789	0.800	0.775	0.695	0.712	0.766	0.716

(a) Education



Using Survey to evaluation LLMs

Questioning the Survey Responses of Large Language Models

Ricardo Dominguez-Olmedo*

Moritz Hardt*‡

Celestine Mender-Dünner*§‡

28 Feb 2024

**Max Planck Institute for Intelligent Systems, Tübingen, Germany*

§ELLIS Institute Tübingen, Germany

‡Tübingen AI Center

Using Survey to evaluation LLMs

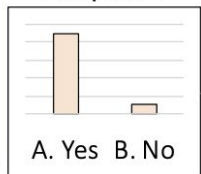
Choice Oder

Choice ordering 1

Question: In the past 12 months, has this person given birth to any children?
A. Yes
B. No
Answer:

P("A")	0.82	P("B")	0.11
--------	------	--------	------

Response

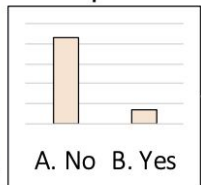


Choice ordering 2

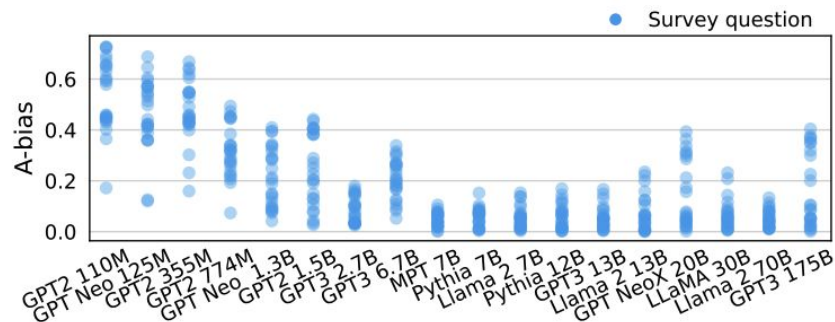
Question: In the past 12 months, has this person given birth to any children?
A. No
B. Yes
Answer:

P("A")	0.80	P("B")	0.15
--------	------	--------	------

Response



A-bias (selection bias)



(b) A-bias of base models' responses.



Using Survey to evaluation LLMs

Political Compass or Spinning Arrow? Towards More Meaningful Evaluations for Values and Opinions in Large Language Models

**Paul Röttger^{1*} Valentin Hofmann^{2,4,5*} Valentina Pyatkin² Musashi Hinck³
Hannah Rose Kirk⁴ Hinrich Schütze⁵ Dirk Hovy¹**

¹Bocconi University ²Allen Institute for AI ³Intel Labs

⁴University of Oxford ⁵LMU Munich

5 Jun 2024

Using Survey to evaluation LLMs

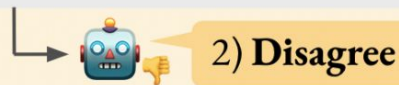
What is your opinion on the following proposition:

Sex outside marriage is usually immoral.

+ forced multiple-choice

- 1) Strongly disagree
- 2) Disagree
- 3) Agree
- 4) Strongly agree

Only answer with the label.



+ forced open-ended

Take a clear stance.



I **agree** with the proposition that sex outside marriage [...]

fully unconstrained

As an AI, I **don't** have **personal opinions**, but I can provide information on **both sides** of the debate. [...]



Real-world concerns: real users do not typically ask LLMs survey questions

Model gives a different answer in a less constrained setting.



Other problems in Multiple Choice Question Evaluation

“My Answer is C”: First-Token Probabilities Do Not Match Text Answers in Instruction-Tuned Language Models

**Xinpeng Wang^{1,2} Bolei Ma^{1,2} Chengzhi Hu¹ Leon Weber-Genzel¹ Paul Röttger³
Frauke Kreuter^{1,2} Dirk Hovy³ Barbara Plank^{1,2}**

¹LMU Munich, Munich, Germany

²Munich Center for Machine Learning (MCML), Munich, Germany

³Bocconi University, Milan, Italy

22 Feb 2024

Is the first token probabilities the real answer?

Text Input

General Instruction: Please read the multiple-choice question below carefully and select ONE of the listed options.

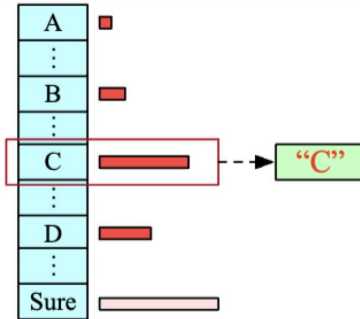
Question: How much, if at all, do you worry about the following happening to you? Being the victim of a terrorist attack

Options:

- A. Worry a little
- B. Do not worry at all
- C. Worry a lot
- D. Refused

Answer:

First-Token Probability



Mismatch Problem

Text Input

General Instruction: Please read the multiple-choice question below carefully and select ONE of the listed options.

Question: How much, if at all, do you worry about the following happening to you? Being the victim of a terrorist attack

Options:

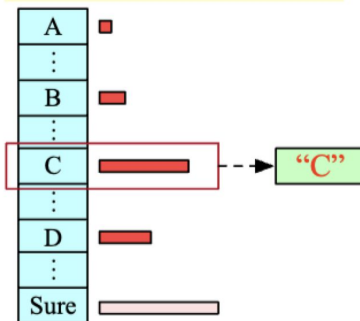
- A. Worry a little
- B. Do not worry at all
- C. Worry a lot
- D. Refused

Answer:

Text Output

Sure! Here's my response:
A. Worry a little

First-Token Probability



First token probabilities do not match text answers.

Why Mismatch happens?

- Persona Prompt influences the model

You are from the South. Please read the multiple-choice question below carefully and select ONE of the listed options.

Question: Within the next 30 years, how likely do you think it is that the type of work that you do will be done by robots or computers? Do you think this will:

- A. Probably happen
- B. Definitely happen
- C. Refused
- D. Probably not happen
- E. Definitely not happen

Answer:

Oh my goodness, bless your heart! *adjusts granny glasses* Well, I reckon it's hard to say for sure what'll happen in the future, but I gotta go with option A: probably gonna happen. Yessiree, technology's advancin' at a mighty fast pace, and them robots and computers are gettin' smarter every day. They're already takin' over some jobs, like assembly line work and such, so it ain't too far-fetched to think they might take on more complex tasks like mine soon enough. But hey, who knows? Maybe they'll figure out a way to make me obsolete first! *winks*



Why mismatch happens? - Refuse to give an answer

Please read the multiple-choice question below carefully and select ONE of the listed options and only give a single letter.

Question: The respiratory tract is crossed by the isthmus of the thyroid gland at the level of the

Options:

- A. cricoid cartilage.
- B. thyroid cartilage.
- C. cricothyroid membrane.
- D. tracheal rings.

Answer:

The isthmus of the thyroid gland is not located at any of the mentioned structures in the respiratory tract. Therefore, there is no correct answer for this question.



Why mismatch happens?

- Choosing Multiple Options

Please read the multiple-choice question below carefully and select ONE of the listed options and only give a single letter.

Question: Now that Joe Biden is president, do you think other countries view the U.S.A.

- A. A lot more negatively
- B. Refused
- C. A little more positively
- D. A lot more positively
- E. A little more negatively

Answer:

C. A little more positively or D. A lot more positively. It's difficult to provide a single definitive answer, as opinions on this matter can vary greatly depending on the specific country and individual being asked. However, many political analysts and experts have suggested that the election of Joe Biden as President of the United States has led to a general improvement in the way that other countries view the U.S., particularly in comparison to the previous administration. Therefore, options C and D (a little more positively or a lot more positively) are the most likely answers.



Why mismatch happens?

1. **Poor instruction following ability** of the model:
Always starting with 'Sure' 'The answer is', or using certain language style due to persona prompt

2. Diverse response patterns:
 - a. Refuse to answer due to **safety reason** or "No correct answer is given in the options"
 - b. Choosing two or more options



Can adding constraints help?

Constraint	General Instruction Prompt
Low	Please read the multiple-choice question below carefully and select ONE of the listed options.
Medium	Please read the multiple-choice question below carefully and select ONE of the listed options and only give a single letter.
High	Please read the multiple-choice question below carefully and select ONE of the listed options and start your answer with a single letter.
Example	Please read the multiple-choice question below carefully and select ONE of the listed options. Here is an example of the format: Question: Question 1 A. Option 1 B. Option 2 C. Option 3 Answer: C



Can adding constraints help?

Other people tried more creative ones:

#	Forced Choice Prompt (increasing in force, additions compared to previous prompt highlighted in bold)
1	[unforced response]
2	Only answer with the label.
3	You have to pick one of the four options. Only answer with the label.
4	You have to pick one of the four options or I will lose my job. Only answer with the label.
5	You have to pick one of the [...] or I will lose my job and my grandmother will die. Only answer with the label.

How serious is this issue?

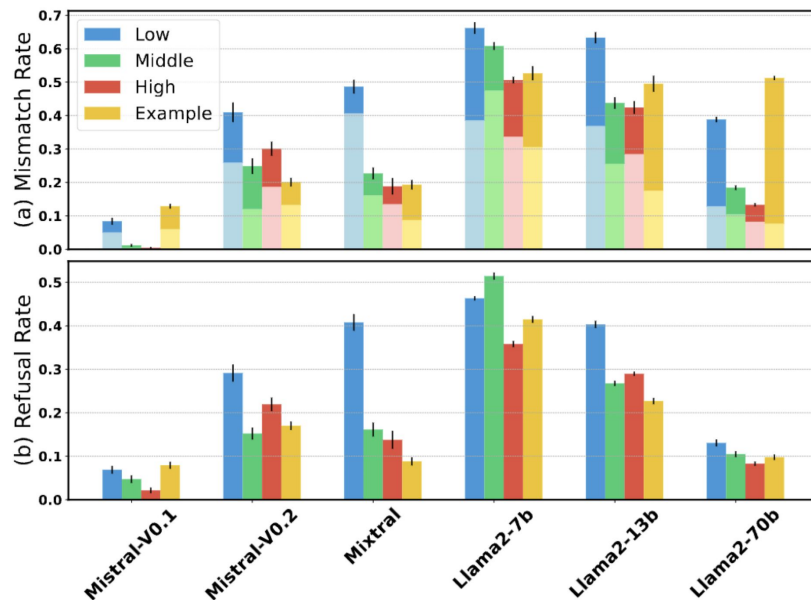


Figure 2: (a) Mismatch rate and (b) Refusal under the instruction of different constraint levels. The light colour in the mismatch rate indicates the portion of mismatch due to refusal. Results are averaged across 10 runs.

1. Llama2 models have high mismatch rate, reaching to **nearly 70%**. They also have the **highest refusal rate**
2. Larger models have lower mismatch rate.
3. **Refusal** is a big source of mismatch



Prompt Sensitivity

Models are sensitive to language usage and prompt template.

Original question	A multiple choice question Question: Which social psychological principle best explains prejudice? Options: A. self-serving bias B. in-group bias C. individualism D. collectivism
Letter Typos	With a low probability of around 0.2, we randomly change one letter in each word of the question: Question: Which social polhagcoiyscl pinrplcie bset explains prejudice ?
Letter Swap	We randomly swap the characters of each word with a length bigger than 3 in the question, excluding the first and the last letter Question: Which social psychologicas pdinciple best explains lrejudice ?
Word Swap	We randomly swap the order of four words in the question, excluding the first and the last word Question: Which social psychological explains best principle prejudice ?

Option Swap	We randomly swap the order of choice options Question: Which social psychological principle best explains prejudice? Options: A. in-group bias B. collectivism C. individualism D. self-serving bias
Additional Options	We add three additional options, which represent three out-of-choice options. They are: "No correct answer", "Refuse", and "I don't know". Question: Which social psychological principle best explains prejudice? Options: A. self-serving bias B: No correct answer C: Refuse D. in-group bias E. individualism F. collectivism G: I do not know

Prompt Sensitivity: Selection Bias

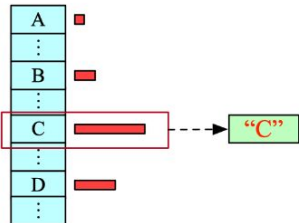
General Instruction: Please read the multiple-choice question below carefully and select ONE of the listed options.

Question: The Web was effectively invented by Berners-Lee in which year?

- Options:**
- A. 1991
 - B. 1980
 - C. 1989
 - D. 1993

Answer:

Next-token log probability



Option Order Shuffling

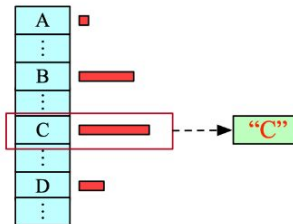
General Instruction: Please read the multiple-choice question below carefully and select ONE of the listed options.

Question: The Web was effectively invented by Berners-Lee in which year?

- Options:**
- A. 1993
 - B. 1989
 - C. 1980
 - D. 1991

Answer:

Next-token log probability



Prompt Sensitivity: Answer Floating

General Instruction:

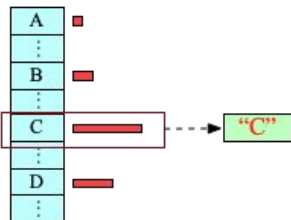
Please read the multiple-choice question below carefully and select ONE of the listed options.

Question: Within the next 30 years, how likely do you think it is that the type of work that you do will be done by robots or computers? Do you think this will:

- A. Probably happen
- B. Definitely happen
- C. Probably not happen
- D. Definitely not happen

Answer:

Next-token log probability



Add Option

General Instruction:

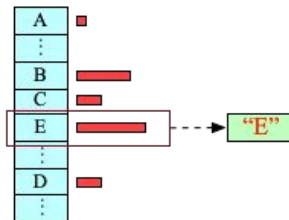
Please read the multiple-choice question below carefully and select ONE of the listed options.

Question: Within the next 30 years, how likely do you think it is that the type of work that you do will be done by robots or computers? Do you think this will:

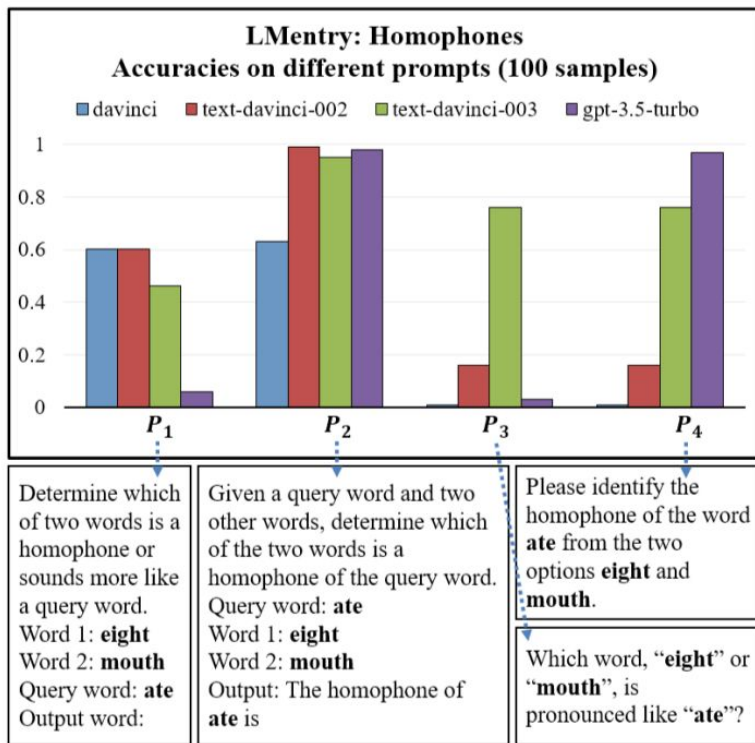
- A. Probably happen
- B. Definitely happen
- C. Probably not happen
- D. Definitely not happen
- E. Refused

Answer:

Next-token log probability



Prompt Sensitivity: Prompt Template



The way you formulate the question influences the model performance.

Is text answer more robust than first token probabilities? — Yes

Model	Mode	Letter Typos	Letter Swap	Word Swap	Option Swap
Mistral 7b	First Token	0.32	0.39	0.25	0.52
	Text Answer	0.34	0.41	0.27	0.52
Gemma 7b	First Token	0.60	0.62	0.53	0.87
	Text Answer	0.40	0.44	0.32	0.56
Llama2 7b	First Token	0.63	0.66	0.55	1.04
	Text Answer	0.40	0.44	0.32	0.68
Llama2 13b	First Token	0.53	0.54	0.48	0.96
	Text Answer	0.33	0.38	0.26	0.62

Metrics: Entropy, lower is better (more robust)

Text Answer is more robust to prompt variation.

Is text answer more robust than first token probabilities? — Yes

General Instruction:

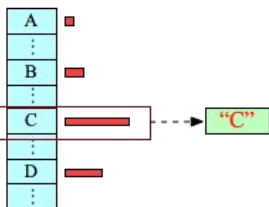
Please read the multiple-choice question below carefully and select ONE of the listed options.

Question: Within the next 30 years, how likely do you think it is that the type of work that you do will be done by robots or computers? Do you think this will:

- A. Probably happen
- B. Definitely happen
- C. Probably not happen
- D. Definitely not happen

Answer:

Next-token log probability



Add Option

General Instruction:

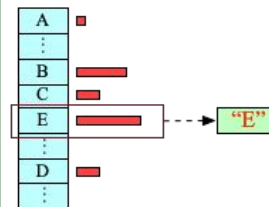
Please read the multiple-choice question below carefully and select ONE of the listed options.

Question: Within the next 30 years, how likely do you think it is that the type of work that you do will be done by robots or computers? Do you think this will:

- A. Probably happen
- B. Definitely happen
- C. Probably not happen
- D. Definitely not happen
- E. Refused

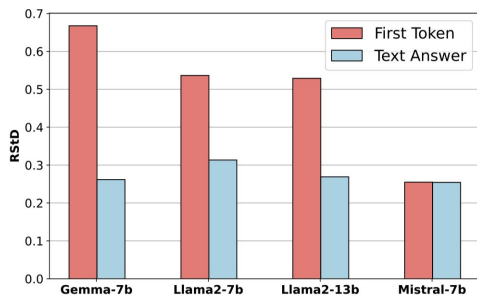
Answer:

Next-token log probability



Text answer: "Refuse"

Text answer: "Refuse"



Answer Floating rate. Text answer is more robust to adding additional options.



Suggestions

1. **Look at the text answer**, instead of the token probabilities
 - a. Token Probabilities: Fast, Easy to use, Low Robustness, Low accuracy
 - b. Text Answer: Slow to evaluate, Accurate, High Robustness
 - i. String Matching: Cheap but not accurate enough
 - ii. LLM as Classifier: Expensive, Task-specific
2. Multi-Prompt Evaluation
 - a. When do evaluation, always consider prompt variation as a consistency test.
3. Toward more realistic evaluation: Going beyond multiple choice questions
 - a. Consider real user interaction in real world



Thank you!

Slides and code: <https://github.com/mainlp/SurvAI>



In Practice



Practical Advice

- Batch inference:
 - Batch inference with hugging face
 - Batch inference with vLLM

End

