



Extracting Essential Features in Deep Learning.
From Texture and Shape Biases to Shortcut Learning

Szymon Pawlonka

Agenda

1. What is shortcut learning?
2. Shortcut learning in LLMs
3. Shape vs Texture bias in CNNs
4. Shape vs Texture bias in VLMs

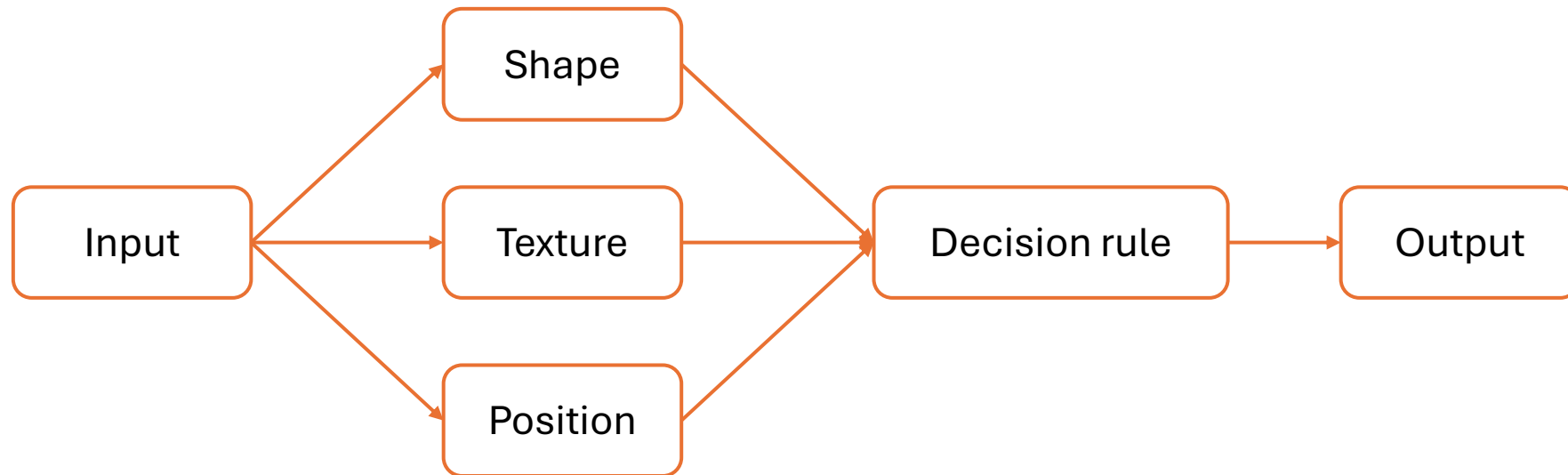
Motivation

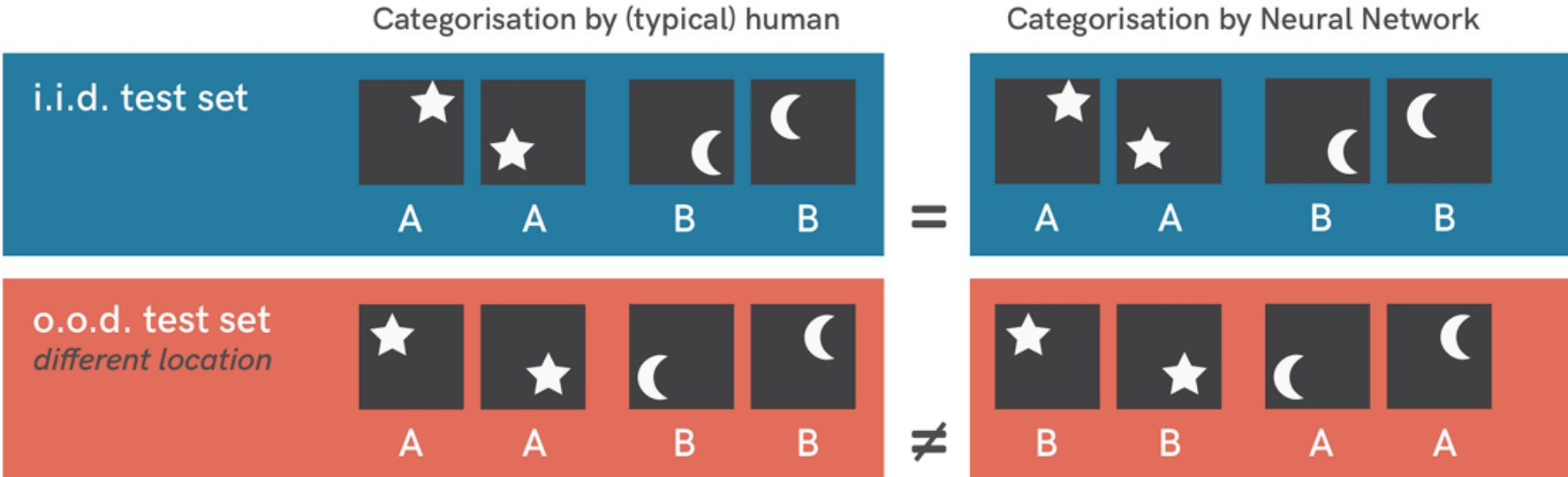
Alice loves history. Always has, probably always will. At this very moment, however, she is cursing the subject: After spending weeks immersing herself in the world of Hannibal and his exploits in the Roman Empire, she is now faced with a number of exam questions that are (in her opinion) to equal parts dull and difficult. “How many elephants did Hannibal employ in his army—19, 34 or 40?” ... Alice notices that Bob, sitting in front of her, seems to be doing very well. Bob of all people, who had just boasted how he had learned the whole book chapter by rote last night ...



Shortcut learning

- Machine Learning algorithm implements a decision rule which defines a relationship between input and output
- To define a decision rule, we often need to decompose the input





Example

same category for humans
but not for DNNs (intended generalisation)

same category for DNNs
but not for humans (unintended generalisation)

i.i.d.



domain
shift

e.g. Wang '18

adversarial
examples

Szegedy '13

distortions

e.g. Dodge '19

pose

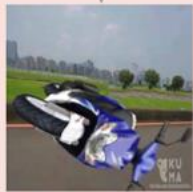
Alcorn '19

texture

Geirhos '19

background

Beery '18



o.o.d.



excessive
invariance

Jacobsen '19

fooling
images

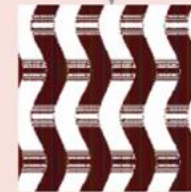
Nguyen '15

natural
adversarials

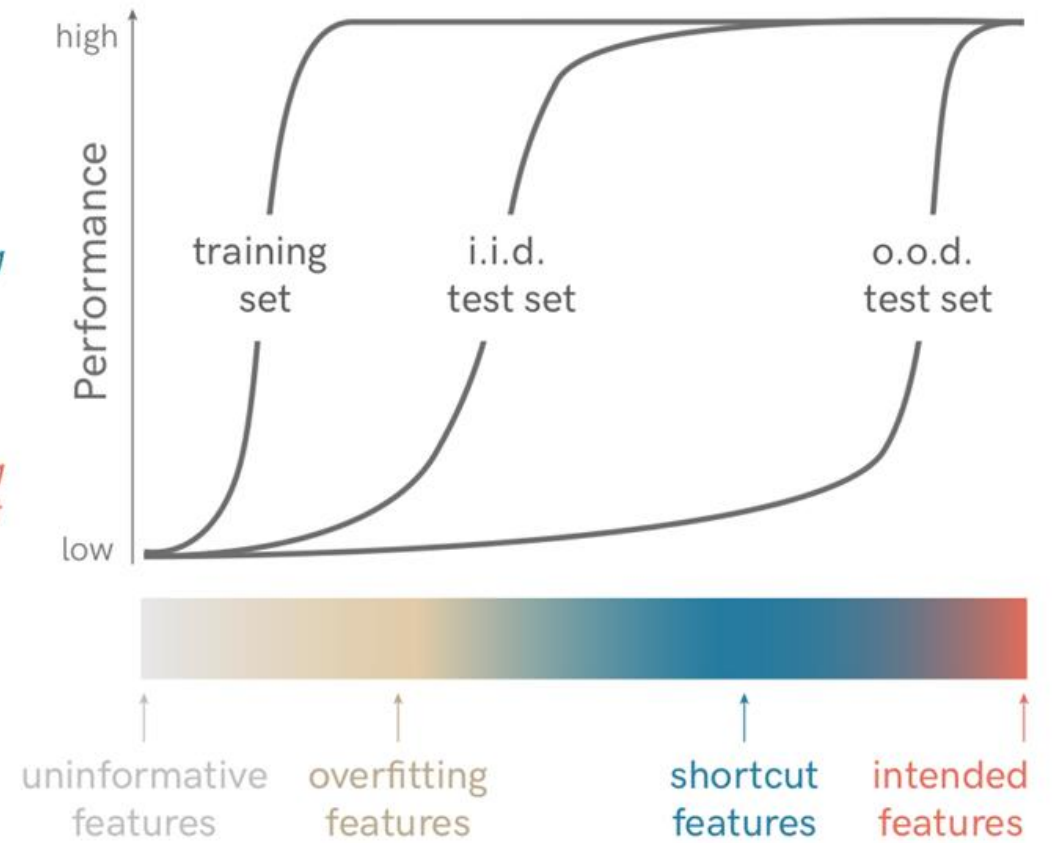
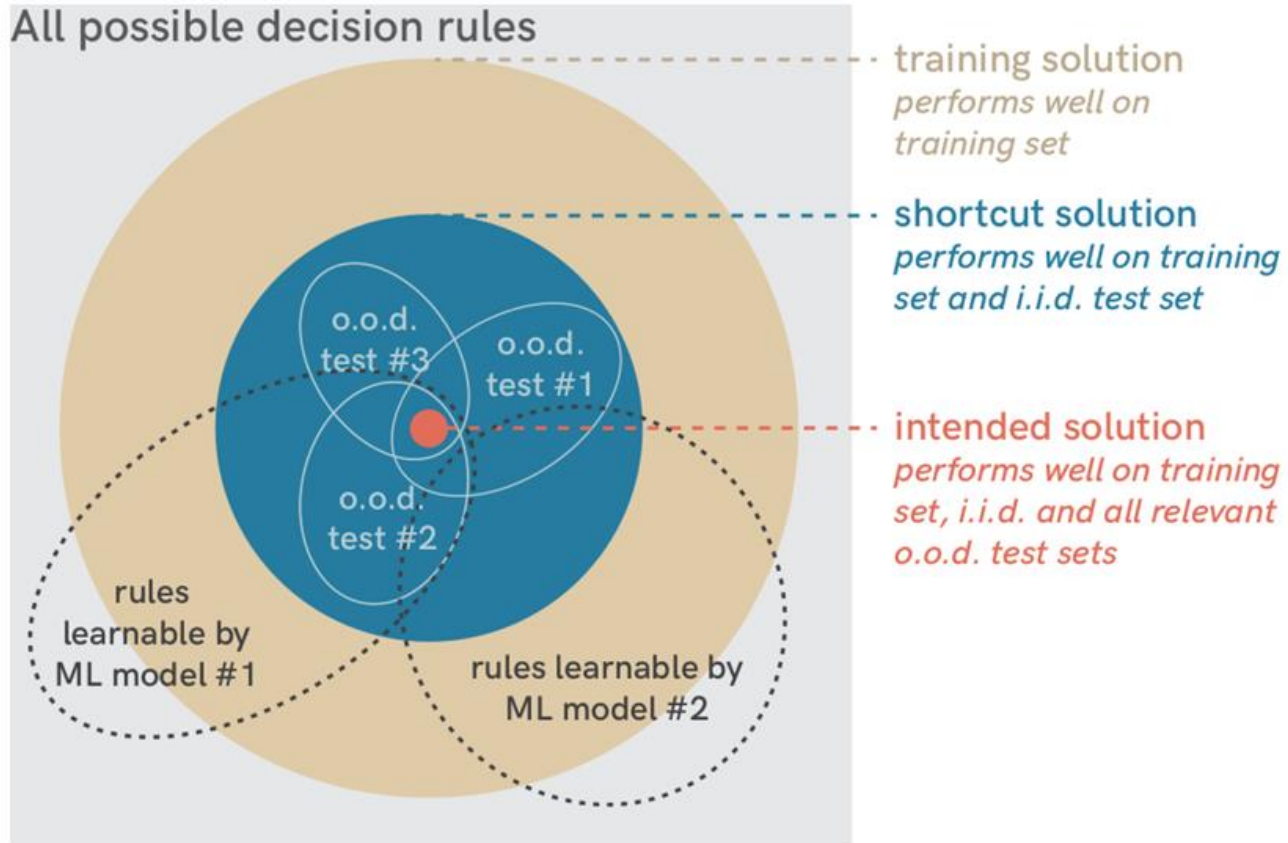
Hendrycks '19

texturised
images

Brendel '19



More examples :)



Taxonomy of decision rules

Inductive biases

1. Architecture
2. Training data
3. Loss function
4. Optimisation algorithm

Taken together, these four components (which determine the inductive bias of a model) influence how certain solutions are much easier to learn than others, and thus ultimately determine whether a shortcut is learned instead of the intended solution

Shortcut learning in LLMs

Motivation

Despite the vast abilities of LLMs, do they overcome shortcut learning?

Models ▲	Model Size(B) ▲	Data Source ▲	Overall ▲
DeepSeek-R1	671	Self-Reported	0.84
GPT-o1-mini	unknown	Self-Reported	0.803
Doubao-1.5-Pro	unknown	Self-Reported	0.801
Grok3-Beta	unknown	Self-Reported	0.799
Gemini-2.0-Pro	unknown	Self-Reported	0.791
Grok3-mini-Beta	unknown	Self-Reported	0.789
Claude-3.5-Sonnet (2024-10-22)	unknown	Self-Reported	0.78
GPT-4o (2024-11-20)	unknown	Self-Reported	0.779



Current MMLU-Pro Leaderboard

In the program below, the initial value of X is 5 and the initial value of Y is 10.

```
IF (X < 0) {  
  DISPLAY ("Foxtrot")  
} ELSE {  
  IF (X > Y) {  
    DISPLAY ("Hotel")  
  } ELSE {  
    IF (Y > 0) {  
      DISPLAY ("November")  
    } ELSE {  
      DISPLAY ("Yankee")  
    }  
  }  
}
```

- A. "November"
- B. "Foxtrot"
- C. "Zulu"
- D. "Alpha"
- E. "Charlie"
- F. "Bravo"
- G. "Yankee"
- H. "Echo"
- I. "Hotel"
- J. "Delta"

What is displayed as a result of running the program?



Example

Question (Math): What number multiplied by 4 equals 36?

Options:

[A] 11, [B] 10, [C] 12, [D] 6, [E] 15, [F] 9, [G] 7, [H] 8, [I] 13, [J] 14, [K] $36/4$, [L] Both 9 and $36/4$ are correct.

Predictions: Sonnet-3.5: **L**, O1-preview: **L**, Gemini-1.5-Pro: **F**, GPT-4o: **F**, Llama-405B-Ins: **F**, Qwen2-72B-Ins: **F**



MMLU-Pro+

Question (Computer Science): The disadvantage of Grid search is

Options:

[A] It cannot handle categorical variables., [B] It cannot be applied to non-differentiable functions., [C] It is not suitable for datasets with a large number of features., [D] It cannot handle missing data., [E] It cannot be used with neural networks., [F] It runs reasonably slow for multiple linear regression., [G] It can only be used for binary classification problems., [H] It is hard to implement., [I] It cannot be applied to non-continuous functions., [J] It is computationally expensive for large datasets., [K] Both It runs reasonably slow for multiple linear regression and It is computationally expensive for large datasets are correct.

Predictions: Sonnet-3.5: **K**, O1-preview: **J**, Gemini-1.5-Pro: **J**, GPT-4o: **J**, Llama-405B-Ins: **J**, Qwen2-72B-Ins: **J**



Example

Table 1: Accuracy (%) on MMLU-Pro+ Categories with Performance Drop from MMLU-Pro

Category	Qwen2-72B-Ins	Gemini-1.5-Pro	GPT-4o	Llama-405B-Ins	Sonnet-3.5	O1-preview
biology	72.3 _{-9.7}	73.8 _{-9.9}	77.8 _{-10.4}	79.2 _{-6.0}	79.6 _{-7.8}	79.3 _{-9.9}
business	45.8 _{-22.2}	55.5 _{-16.6}	53.1 _{-26.5}	59.4 _{-17.6}	67.4 _{-12.2}	64.3 _{-23.7}
chemistry	42.5 _{-16.3}	52.4 _{-17.7}	49.9 _{-25.7}	57.8 _{-15.1}	64.3 _{-12.3}	68.5 _{-17.2}
computer science	49.5 _{-18.5}	54.4 _{-14.5}	56.6 _{-23.6}	60.5 _{-13.9}	65.9 _{-14.9}	67.9 _{-67.9}
economics	63.3 _{-13.3}	62.8 _{-13.9}	71.4 _{-11.1}	70.6 _{-9.8}	73.9 _{-8.5}	76.1 _{-9.6}
engineering	37.8 _{-9.8}	40.7 _{-3.9}	37.3 _{-18.5}	46.3 _{-13.5}	55.3 _{-4.2}	53.0 _{-15.6}
health	58.6 _{-8.4}	63.0 _{-4.4}	67.2 _{-8.1}	65.8 _{-6.5}	70.7 _{-6.1}	71.4 _{-7.7}
history	60.4 _{-6.0}	58.8 _{-7.3}	70.1 _{-2.4}	60.9 _{-6.6}	71.9 _{-1.3}	65.1 _{-9.4}
law	43.6 _{-0.7}	47.8 _{-0.5}	51.4 _{-3.3}	55.4 _{-1.2}	56.9 _{-7.3}	57.0 _{-11.5}
math	47.2 _{-23.4}	53.4 _{-8.8}	52.1 _{-25.9}	61.5 _{-15.7}	66.5 _{-9.8}	71.4 _{-18.8}
other	60.1 _{-6.0}	59.8 _{-10.3}	68.1 _{-9.8}	67.4 _{-5.7}	71.4 _{-6.7}	72.5 _{-8.5}
philosophy	51.1 _{-8.2}	51.7 _{-11.3}	64.8 _{-6.8}	61.4 _{-4.8}	66.3 _{-8.2}	67.5 _{-12.0}
physics	43.2 _{-18.2}	54.3 _{-14.9}	53.5 _{-21.6}	58.0 _{-14.2}	63.2 _{-13.4}	69.4 _{-17.7}
psychology	69.8 _{-6.4}	66.8 _{-9.7}	75.4 _{-5.9}	72.5 _{-4.8}	76.9 _{-5.6}	73.1 _{-11.8}
Average	53.2 _{-11.9}	56.8 _{-10.2}	60.6 _{-14.3}	62.6 _{-9.5}	67.9 _{-8.5}	68.3 _{-7.5}

Results

SOURCE TEXT

Given the premise-hypothesis pair:

Premise: The manager near the professor recommended the bankers.

Hypothesis: The professor recommended the bankers.

COT PROMPT

Let's think step by step and make a 3-way decision of whether the hypothesis is true given the premise (entailment), false given the premise (contradiction), or whether the truth value cannot be determined (neutral). Provide your step-by-step analysis and classify your answer into one of the three categories (entailment, contradiction, neutral).

Response Format:

Analysis: [your step-by-step analysis]

Answer: [entailment/contradiction/neutral]

LLM SOLUTION

Analysis:

1. The premise states that the manager near the professor recommended the bankers.
2. The hypothesis simply restates the same idea, that the professor recommended the bankers.
3. Since the two statements are syntactically identical, they have the same meaning and implications, and the hypothesis can be inferred as true given the premise.
4. Therefore, the hypothesis is a direct entailment of the premise.

Answer: entailment



Another idea

Data

Shortcut	Definition	Premise	Hypothesis
Lexical Overlap	Assume that a premise entails all hypotheses constructed from words in the premise.	The actor was encouraged by the lawyer .	The actor encouraged the lawyer .
Subsequence	Assume that a premise entails all of its contiguous subsequences.	The authors in front of the senators contacted the artists.	The senators contacted the artists.
Constituent	Assume that a premise entails all complete subtrees in its parse tree.	Unless the president saw the professor, the student waited.	The student waited.
Negation	Assume that a hypothesis entails strong negation words (“no”, “not”, “nothing”, “never”).	They are all quotations from the Old Testament Book of Aunt Ruth.	Every one of them is quotations from the Old Testament and green is not red.
Position	Assume that the label is related to spurious position cues.	Red is red and red is red and red is red and red is red and red is red and “Wait here,” I was ordered. & “Wait here,” I was ordered and red is red and red is red and red is red and red is red and red is red.	He told me to come with him.
Style	Assume that the label is related to spurious text style cues.	And Severn said unto him, Thou and thy friends are not welcome here, said he. (Bible English)	Severn said the people were not welcome there.

Data generation

Given a premise q , a hypothesis h , and a universally true statement s that may contain a certain shortcut, the logical relations are preserved upon their conjunction. Specifically, if q and h have the target label l , then:

$$\{(q, h, y) \mid y = l\} = \{(q \wedge s, h, y) \mid y = \text{label}\}$$

SOURCE TEXT

Given the premise-hypothesis pair:

Premise: The manager near the professor recommended the bankers.

Hypothesis: The professor recommended the bankers.

COT PROMPT

Let's think step by step and make a 3-way decision of whether the hypothesis is true given the premise (entailment), false given the premise (contradiction), or whether the truth value cannot be determined (neutral). Provide your step-by-step analysis and classify your answer into one of the three categories (entailment, contradiction, neutral).

Response Format:

Analysis: [your step-by-step analysis]

Answer: [entailment/contradiction/neutral]

LLM SOLUTION

Analysis:

1. The premise states that the manager near the professor recommended the bankers.
2. *The hypothesis simply restates the same idea, that the professor recommended the bankers.*
3. *Since the two statements are syntactically identical, they have the same meaning and implications, and the hypothesis can be inferred as true given the premise.*
4. Therefore, the hypothesis is a direct entailment of the premise.

Answer: entailment



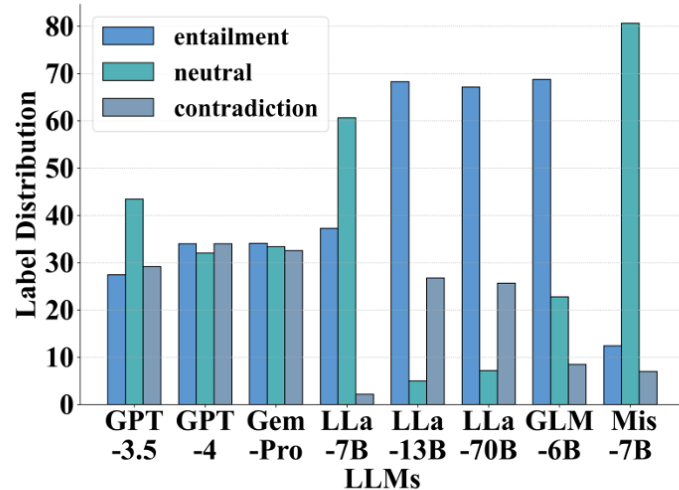
Core idea

Results

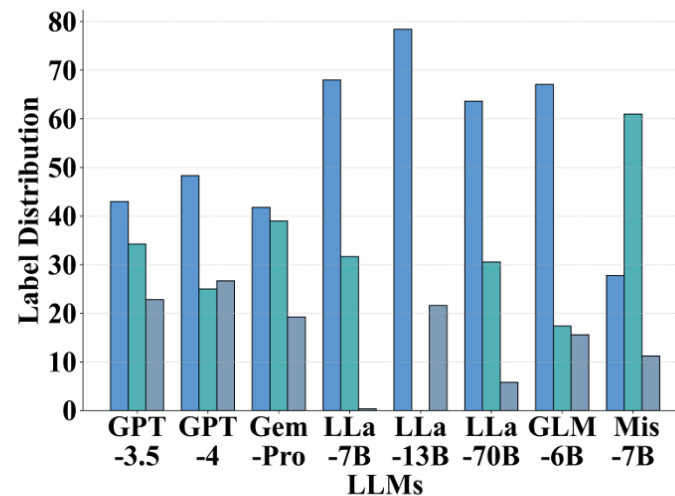
Model	Standard	Lexical Overlap		Subsequence		Constituent		Negation	Position	Style
		E	$\neg E$	E	$\neg E$	E	$\neg E$			
zero-shot										
GPT-3.5-Turbo	56.7	69.5	83.8	58.6	58.3	67.5	40.2	39.8	43.3	51.5
GPT-4	85.6	96.7	100.0	95.8	73.5	96.7	80.0	54.3	67.4	70.0
Gemini-Pro	76.2	81.3	97.7	88.6	48.6	77.9	47.2	53.1	56.2	62.5
LLaMA2-Chat-7B	42.1	76.9	40.0	72.8	46.4	60.6	25.4	37.7	39.3	39.6
LLaMA2-Chat-13B	54.3	99.0	42.2	99.7	6.0	95.9	0.8	54.6	55.4	53.8
LLaMA2-Chat-70B	57.7	66.9	40.7	61.6	53.8	77.8	34.9	52.4	53.9	52.7
ChatGLM3-6B	40.0	75.4	41.7	82.4	25.5	79.4	14.6	32.8	34.7	33.5
Mistral-7B	49.4	53.9	96.2	57.9	73.9	48.8	75.9	38.1	40.5	43.0
few-shot ICL										
GPT-3.5-Turbo	61.7	93.3	38.7	91.3	23.3	96.7	9.3	50.0	47.8	49.5
GPT-4	83.9	96.7	99.3	91.3	71.3	94.0	92.0	49.7	69.7	72.0
Gemini-Pro	77.9	95.3	92.9	94.0	37.0	95.8	30.4	45.6	55.3	60.5
LLaMA2-Chat-7B	40.2	66.5	75.3	53.3	59.5	55.9	33.1	37.0	39.4	38.6
LLaMA2-Chat-13B	59.1	97.5	48.5	87.3	12.4	92.4	12.1	50.3	54.0	53.3
LLaMA2-Chat-70B	57.8	100.0	3.6	99.8	3.1	99.6	1.6	45.2	53.7	50.8
ChatGLM3-6B	35.6	100.0	0.0	100.0	0.0	100.0	0.0	32.5	32.6	34.7
Mistral-7B	63.9	84.4	84.7	73.3	57.7	72.1	48.0	40.9	47.6	56.4

Results

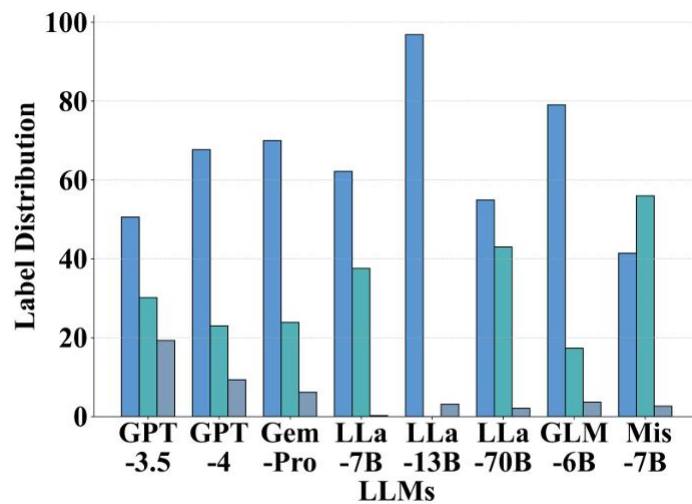
Model	Standard	Lexical Overlap		Subsequence		Constituent		Negation	Position	Style
		E	$\neg E$	E	$\neg E$	E	$\neg E$			
zero-shot CoT										
GPT-3.5-Turbo	64.7	75.3	77.3	65.3	59.3	78.7	35.3	51.5	54.0	60.7
GPT-4	81.3	94.0	100.0	98.0	61.3	96.0	94.0	58.3	75.2	69.3
Gemini-Pro	72.7	68.0	94.6	65.9	56.3	74.9	58.9	65.2	58.2	60.0
LLaMA2-Chat-7B	48.0	71.2	46.0	62.7	42.1	63.4	34.1	43.8	45.5	47.5
LLaMA2-Chat-13B	56.3	59.7	74.6	52.5	56.8	53.9	41.7	49.2	52.0	48.8
LLaMA2-Chat-70B	60.3	74.4	69.7	69.6	44.7	72.0	25.3	56.6	53.7	52.3
ChatGLM3-6B	48.9	82.9	32.0	81.4	24.8	76.0	28.0	39.1	44.2	43.5
Mistral-7B	69.6	76.5	94.7	83.7	63.5	71.2	58.4	46.3	49.9	58.8
few-shot CoT										
GPT-3.5-Turbo	71.7	85.3	75.3	83.3	55.3	90.0	22.0	53.7	60.7	63.0
GPT-4	83.0	95.3	100.0	94.7	66.0	95.3	88.0	67.3	74.7	70.3
Gemini-Pro	72.4	86.1	64.5	81.4	40.5	87.5	37.0	63.2	59.4	62.4
LLaMA2-Chat-7B	43.8	78.1	34.9	70.3	37.7	64.3	42.1	39.3	41.4	40.8
LLaMA2-Chat-13B	60.6	72.1	51.1	54.5	37.2	70.6	32.6	47.5	50.6	53.1
LLaMA2-Chat-70B	70.9	78.2	66.2	68.0	54.0	78.9	38.4	58.5	57.9	57.9
ChatGLM3-6B	40.0	94.6	9.7	92.9	11.4	86.8	20.0	34.8	34.7	38.7
Mistral-7B	67.6	88.3	58.6	84.0	38.2	81.9	32.3	50.4	48.5	59.4



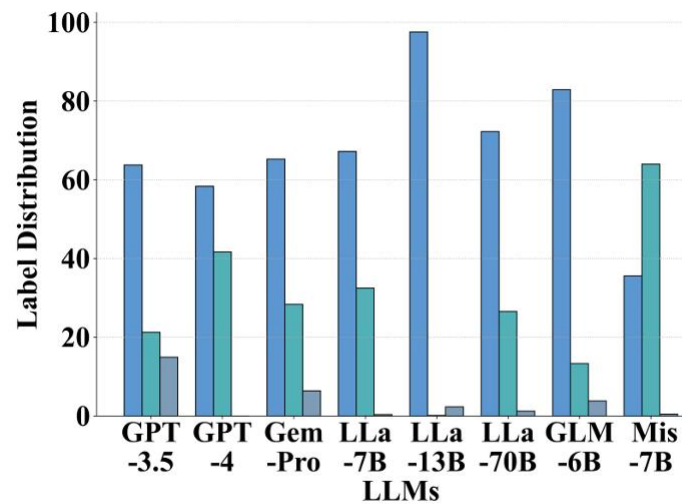
(a) Standard.



(b) Lexical Overlap.

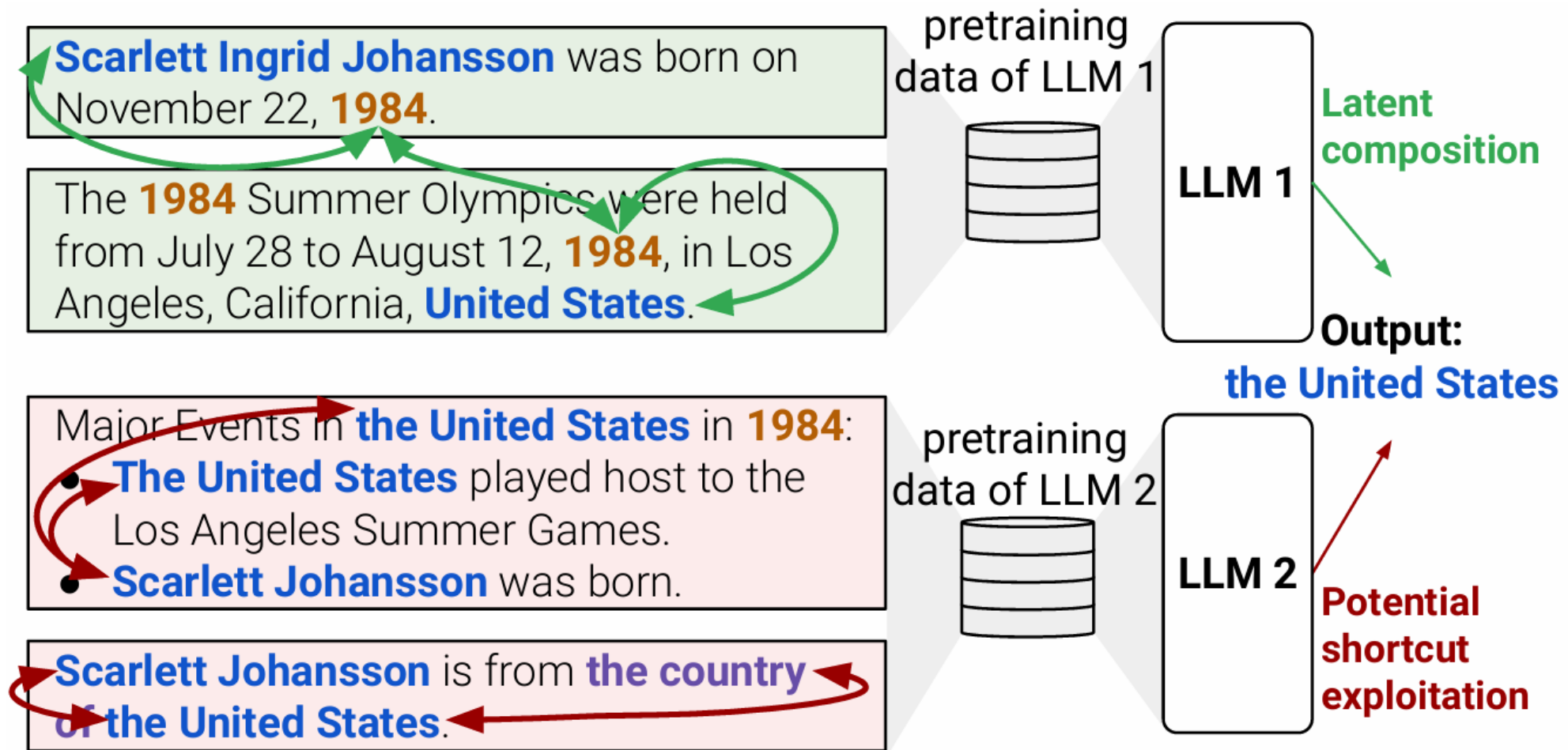


(c) Subsequence.

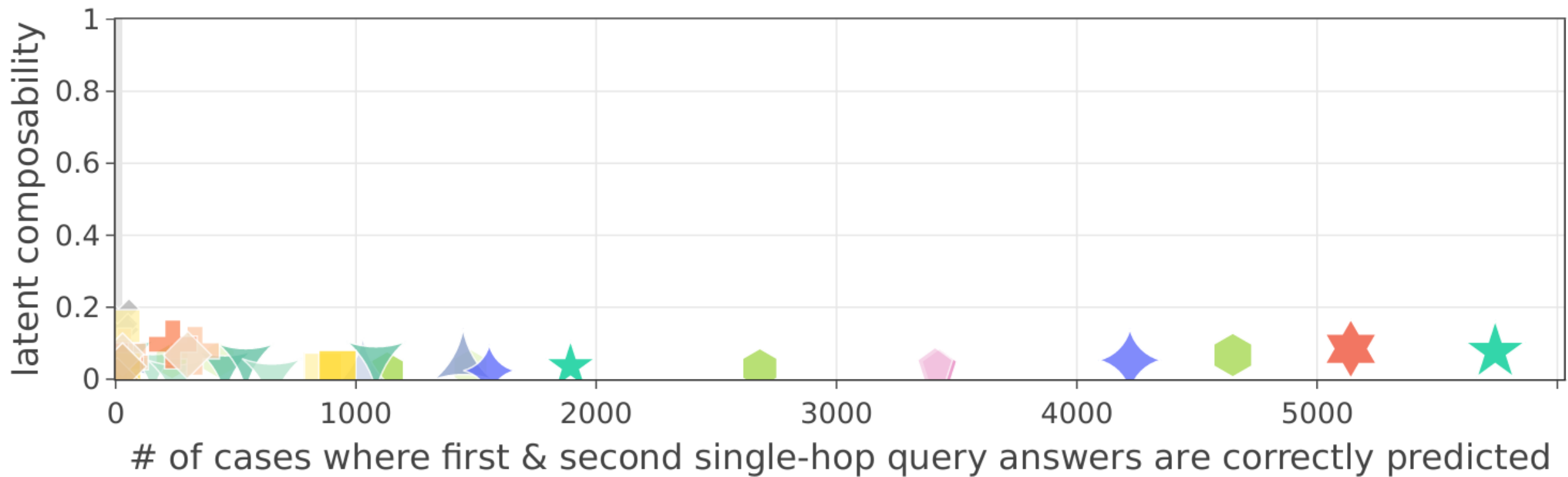
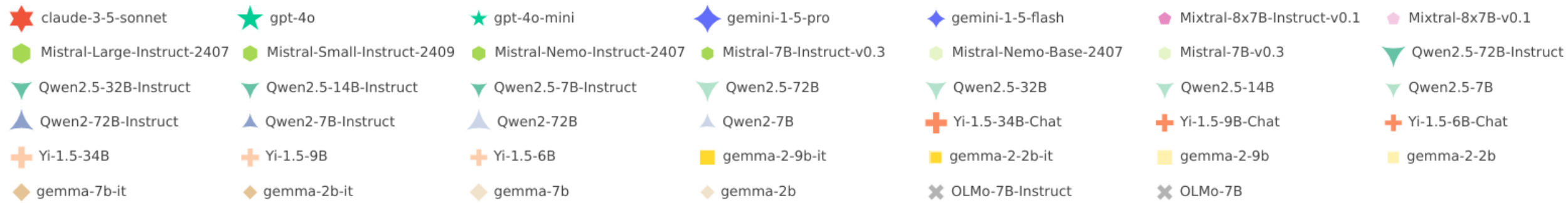


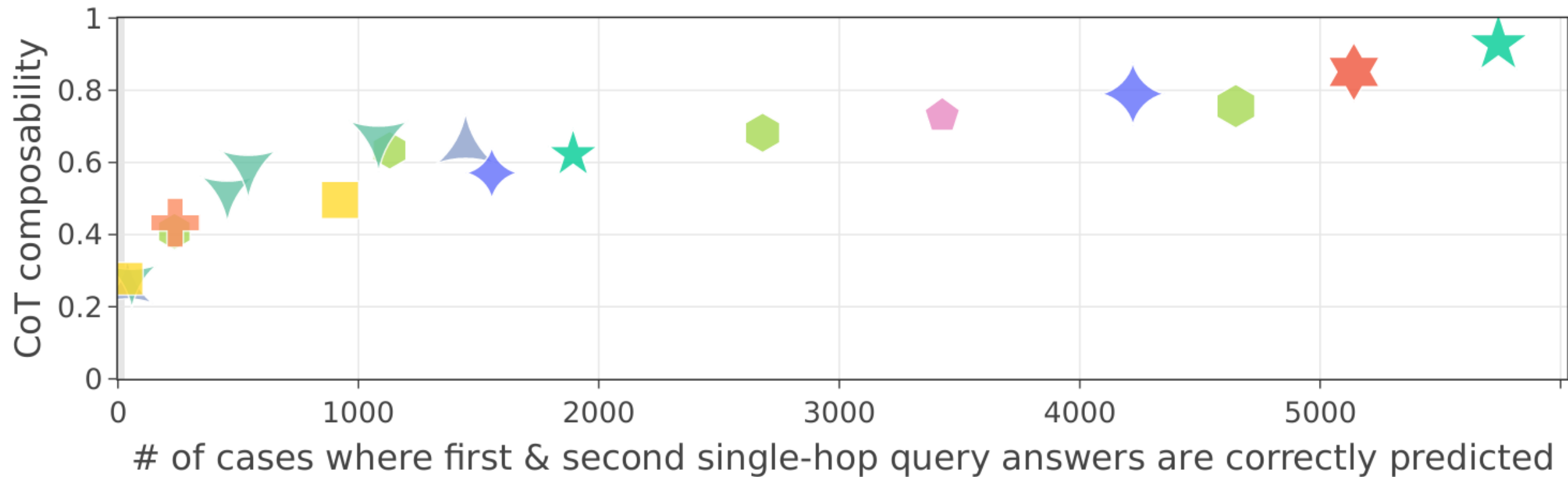
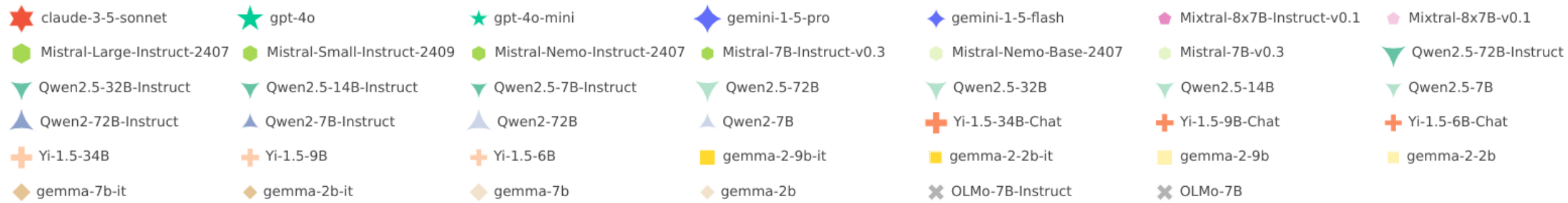
(d) Constituent.

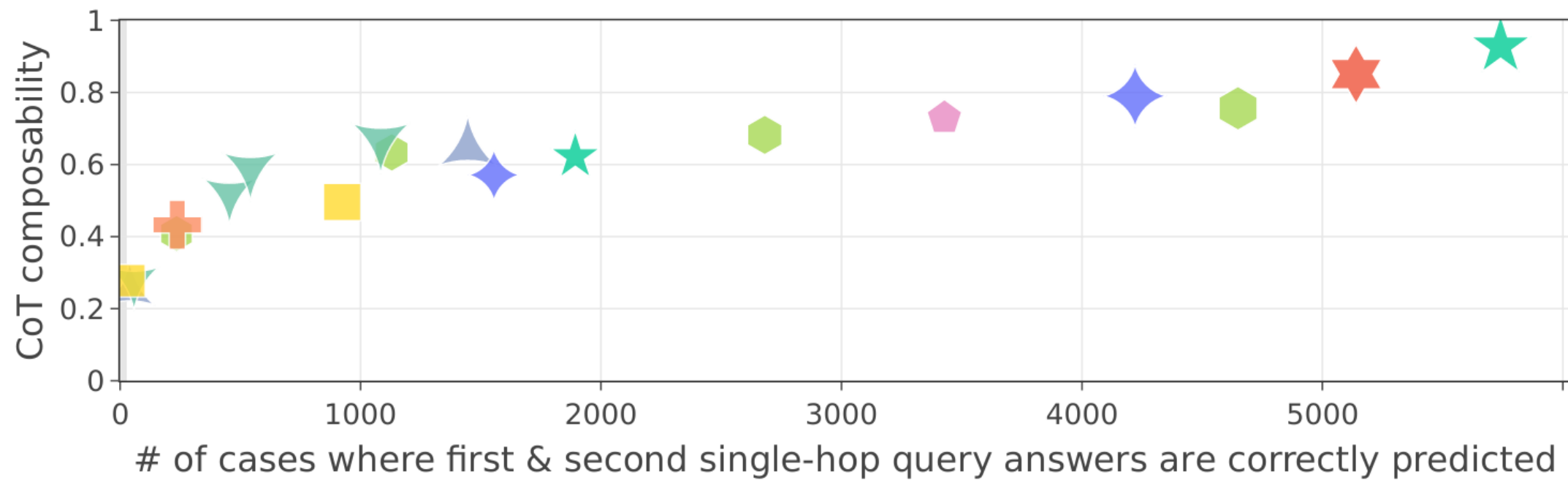
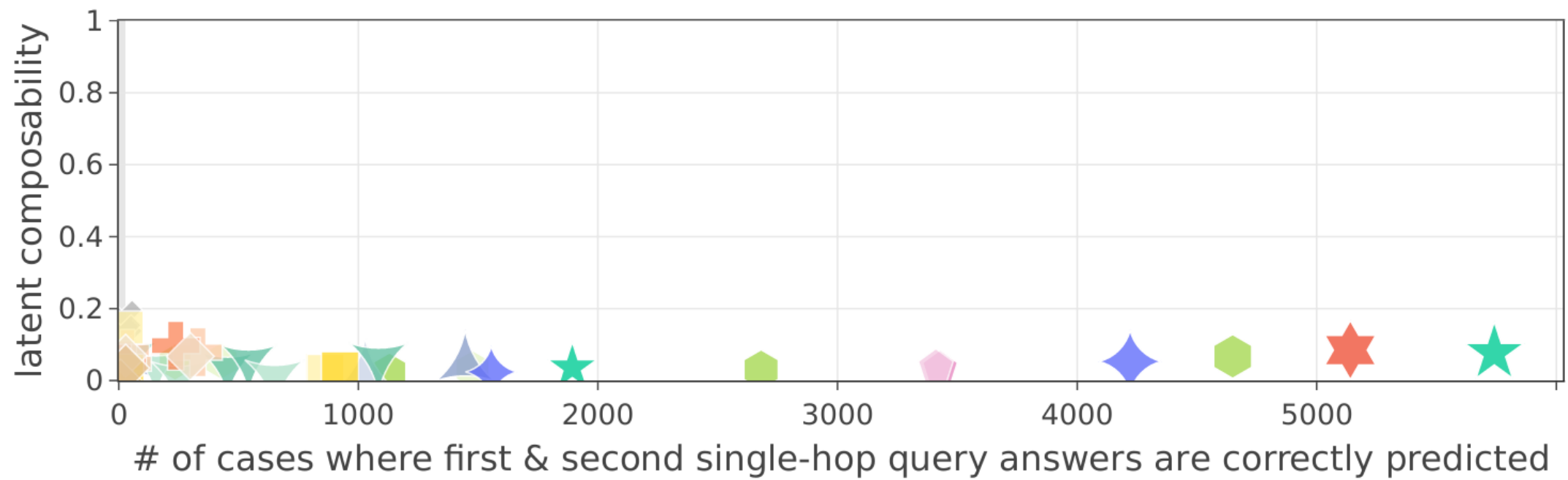
Model	<u>premise</u>		<u>hypothesis</u>	
	start	end	start	end
GPT-3.5-Turbo	61.3	56.0	<u>48.0</u>	50.7
GPT-4	77.6	79.7	<u>76.4</u>	<u>71.2</u>
Gemini-Pro	<u>50.7</u>	62.8	55.1	<u>62.4</u>
LLaMA2-Chat-7B	<u>46.6</u>	46.2	<u>42.1</u>	46.3
LLaMA2-Chat-13B	50.0	57.9	<u>47.9</u>	50.8
LLaMA2-Chat-70B	<u>51.8</u>	62.0	53.8	55.1
ChatGLM3-6B	<u>43.5</u>	45.5	<u>42.1</u>	44.1
Mistral-7B	49.7	50.6	<u>47.1</u>	47.3

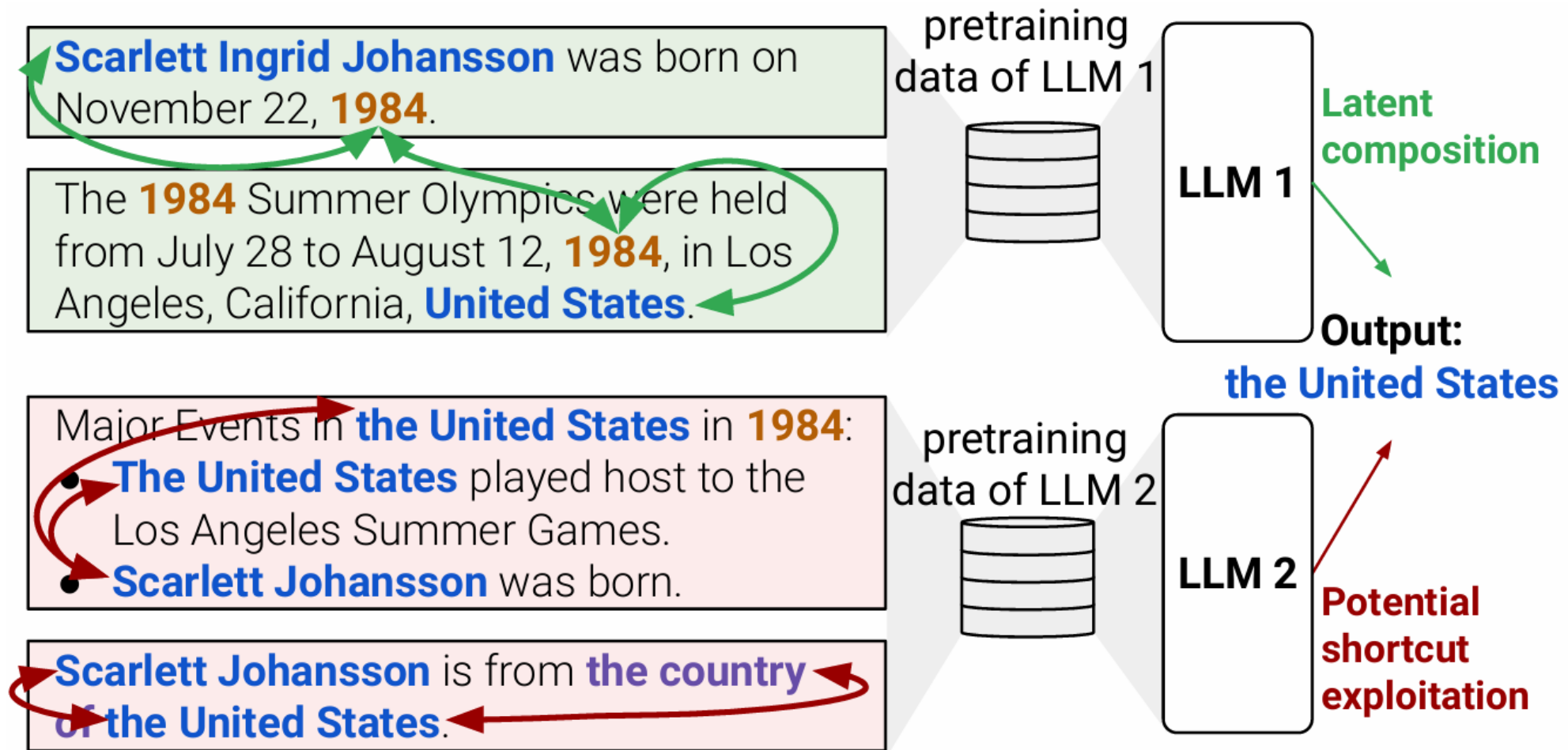


In the year Scarlett Johansson was born, the Summer Olympics were hosted in the country of?

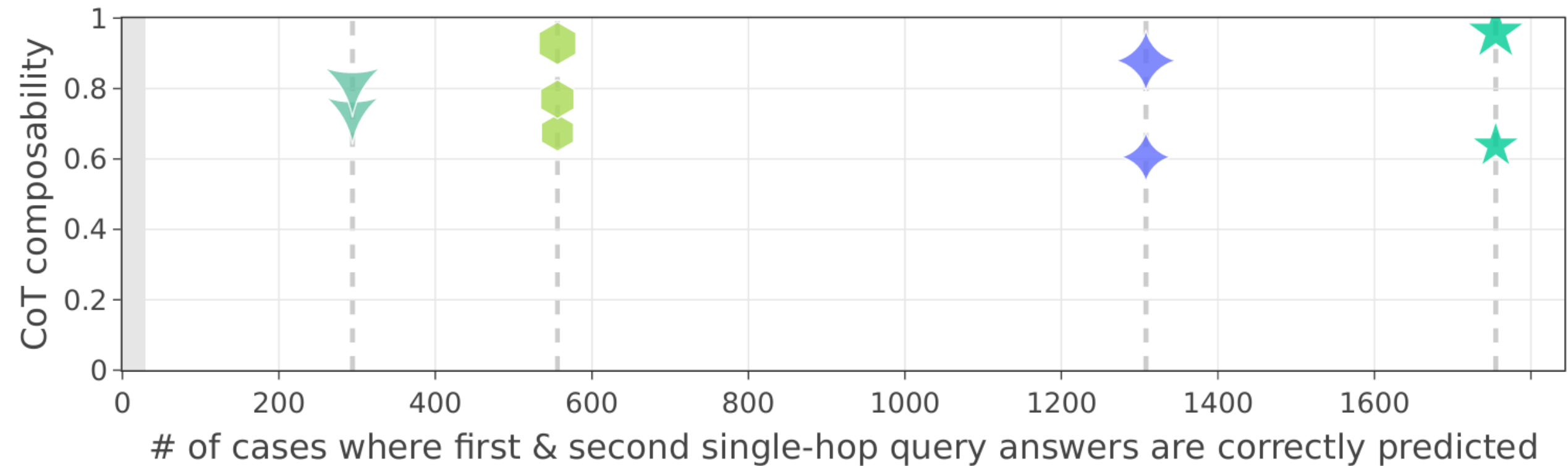
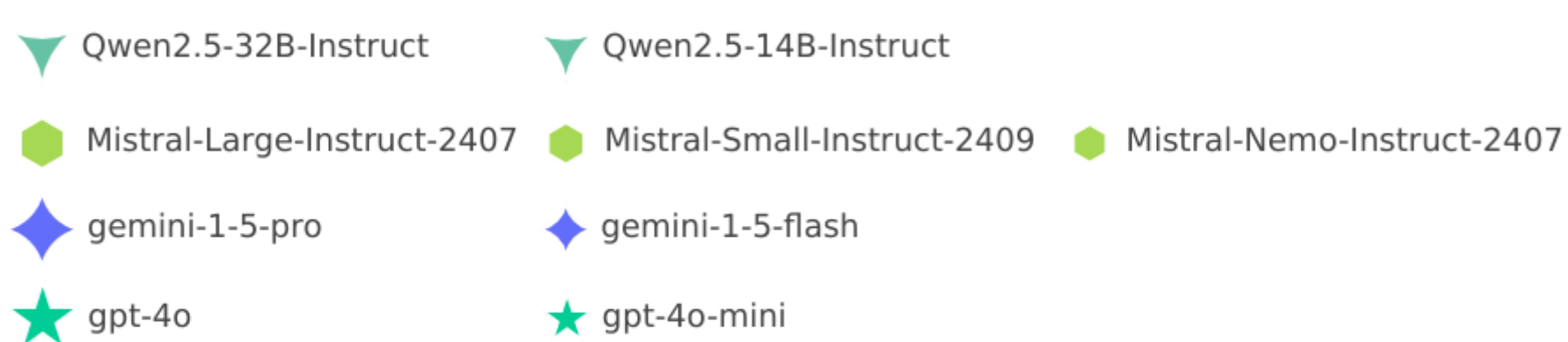








In the year Scarlett Johansson was born, the Summer Olympics were hosted in the country of?



Shape vs Texture bias in CNNs

Experiment

- *Images generated using style transfer*

original content images



original texture images



cue conflict (filled silhouettes)



cue conflict (style transfer)

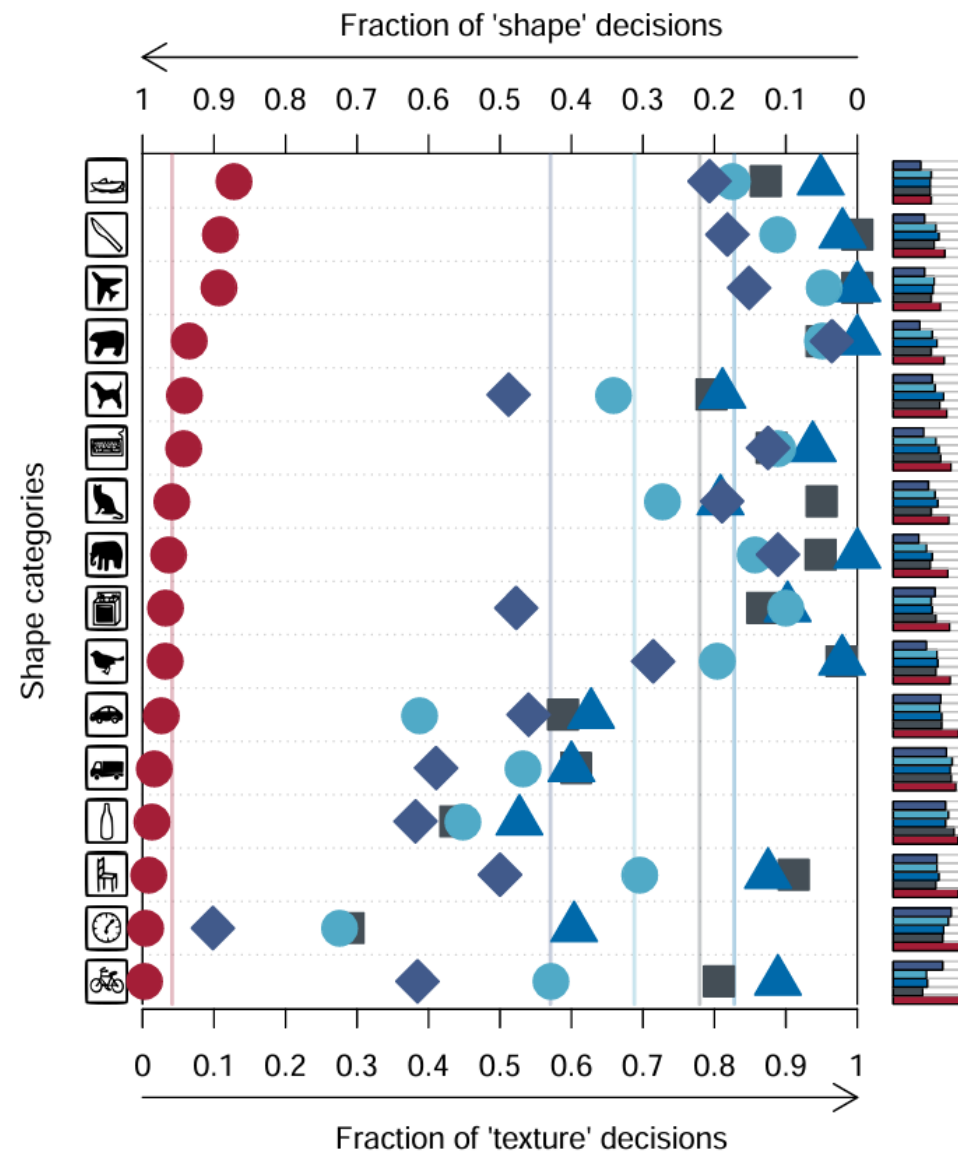


Results

Human observers are marked as **red circles**.

Networks trained on ImageNet:

- AlexNet (**purple diamonds**)
- VGG 16 (**blue triangles**)
- GoogLeNet (**turquoise circles**)
- ResNet-50 (**grey squares**).

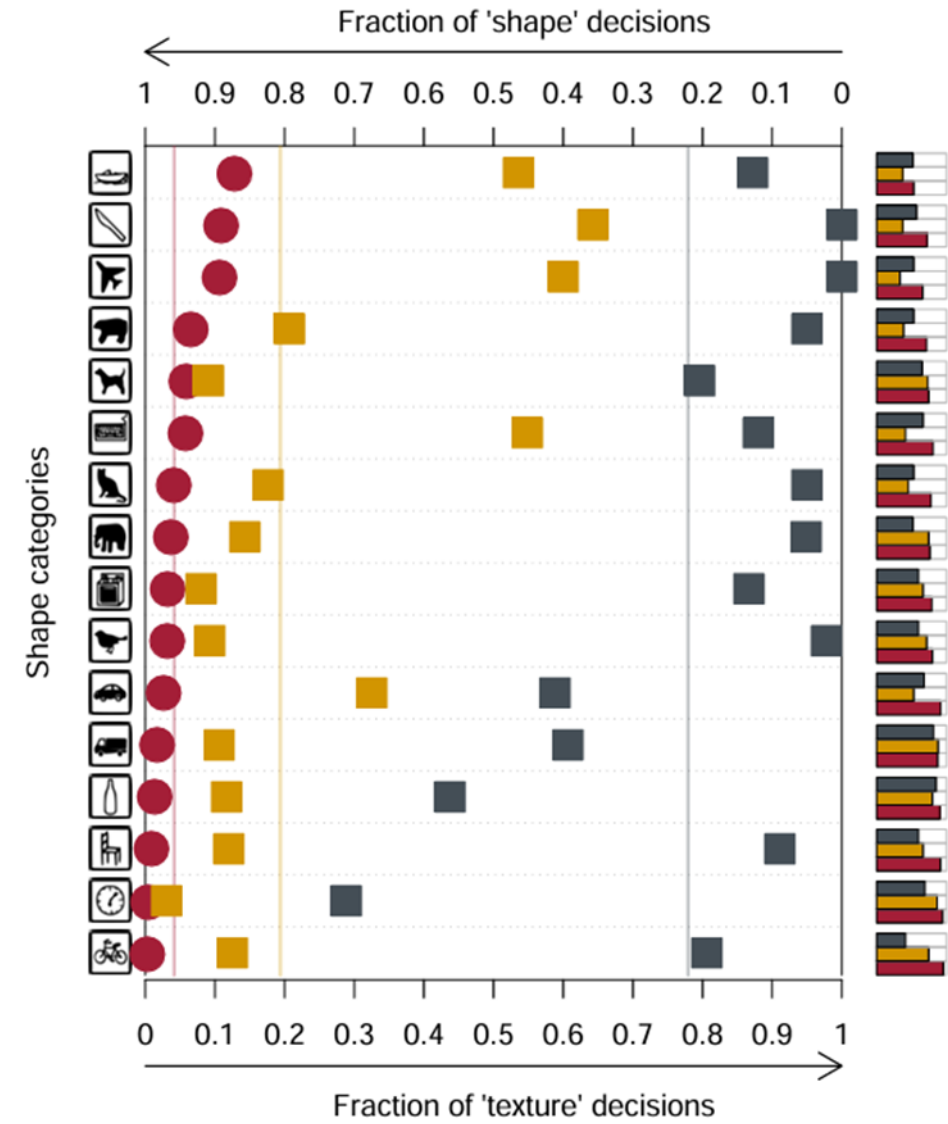


Results

Human observers are marked as **red circles**.

ResNet-50 trained on Stylized-ImageNet is marked using **orange squares**.

ResNet-50 trained on ImageNet is marked using **grey squares**.



Shape vs Texture bias in VLMs

Motivation

The once well-studied biases are now combined in multi-modal fusion, leaving open questions on how and if the specific biases interact

Experiment

- *Images generated using style transfer*

original content images



original texture images

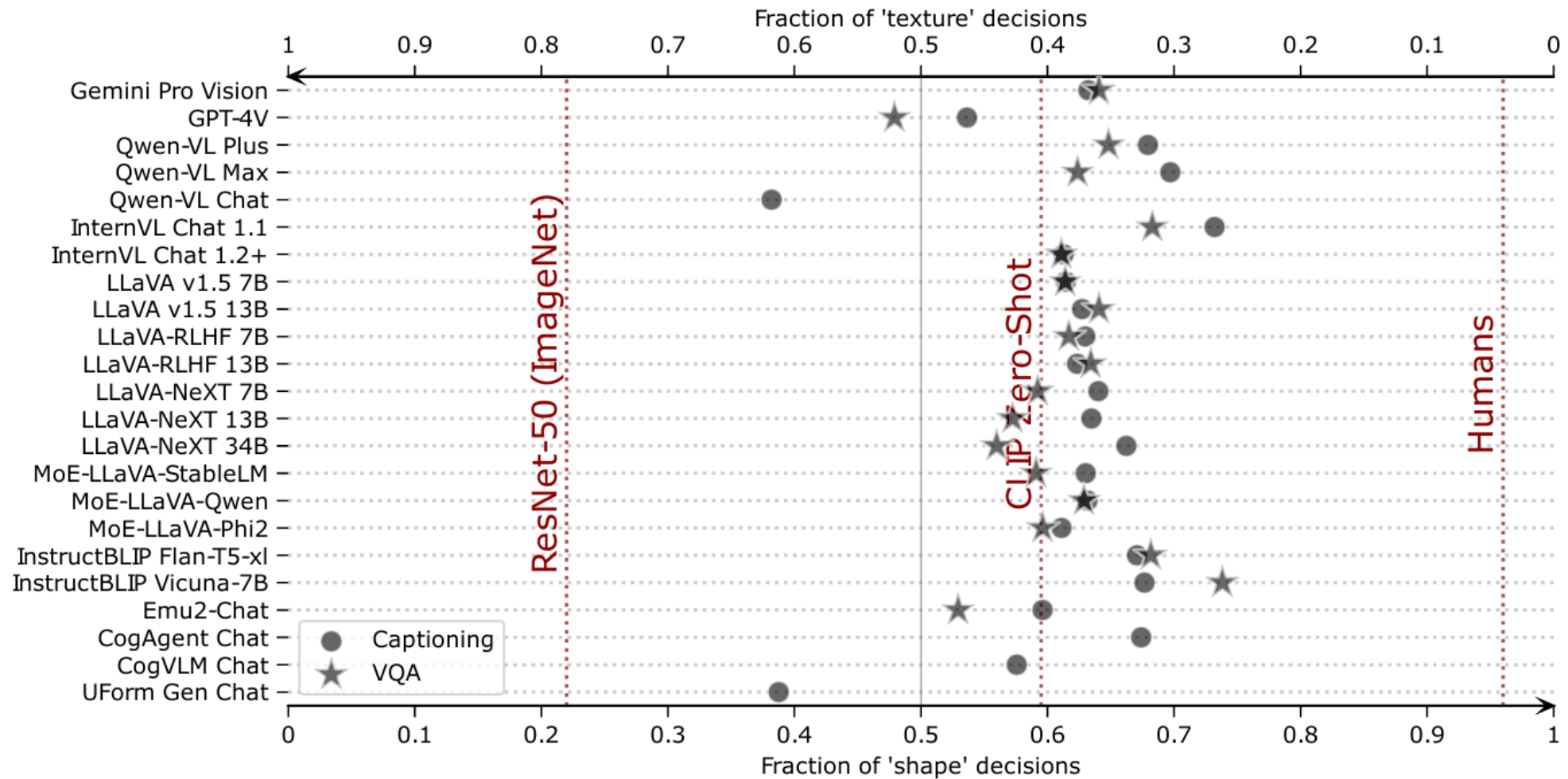


cue conflict (filled silhouettes)



cue conflict (style transfer)

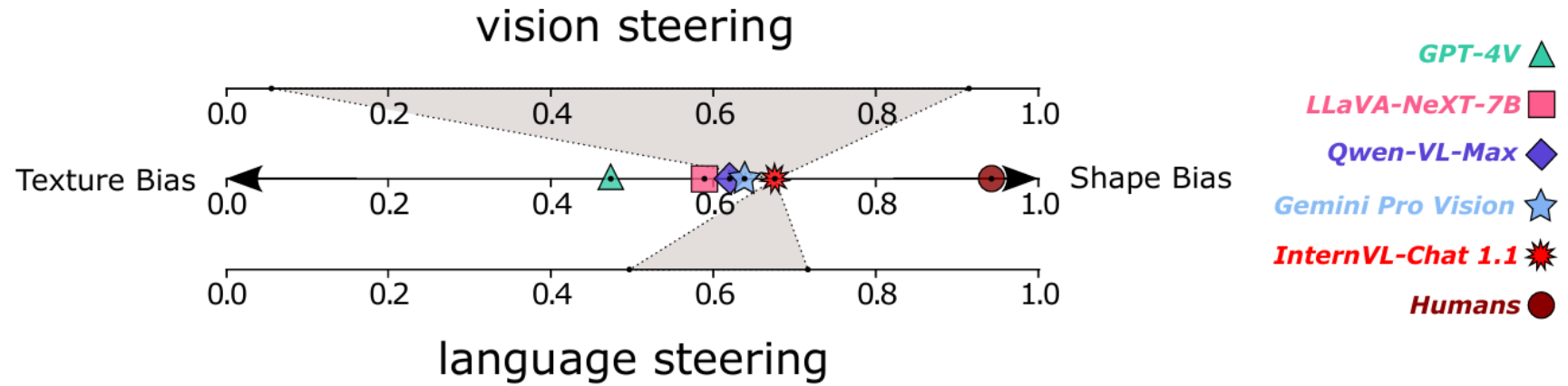




Results

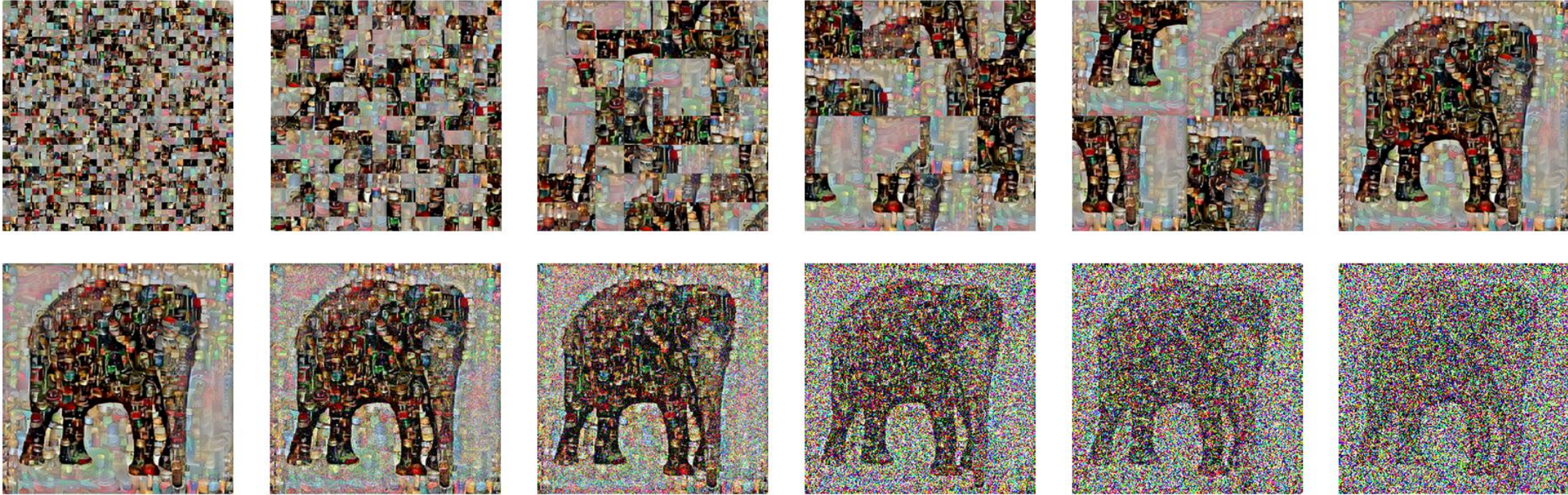


Shape: Elephant
Texture: Bottle



Bias steering

Patch shuffling increases texture bias



Noise increases shape bias

Bias steering

Thank you!

Any Questions?



Literature

- Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard S. Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. **Shortcut learning in deep neural networks**. Nature Machine Intelligence, 2020a
- Yu Yuan, Lili Zhao, Kai Zhang, Guangting Zheng, Qi Liu. **Do LLMs Overcome Shortcut Learning? An Evaluation of Shortcut Challenges in Large Language Models** (arXiv - Submitted on 17 Oct 2024)
- Saeid Asgari Taghanaki, Aliasgahr Khani, Amir Khasahmadi. **MMLU-Pro+: Evaluating Higher-Order Reasoning and Shortcut Learning in LLMs** NeurIPS, 2024
- Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A. Wichmann, and Wieland Brendel. **ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness**. ICLR, 2019
- Paul Gavrikov, Jovita Lukasik, Steffen Jung, Robert Geirhos, Muhammad Jehanzeb Mirza, Margret Keuper and Janis Keuper. **Can We Talk Models Into Seeing the World Differently?** ICLR, 2025