Enhancing Uncertainty-Based Hallucination Detection with Stronger Focus

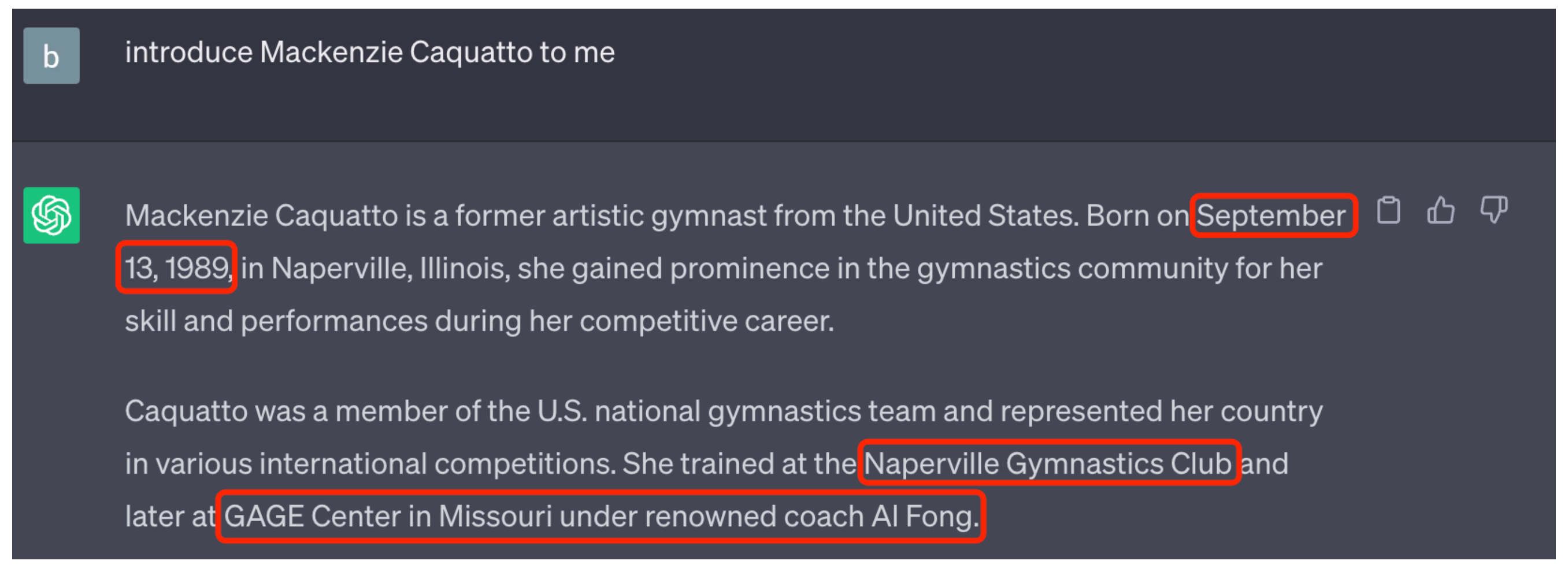
EMNLP 2023

Tianhang Zhang[1], Lin Qiu[2], Qipeng Guo[2], Cheng Deng[1], Yue Zhang[3], Zheng Zhang[2], Chenghu Zhou[4], Xinbing Wang[1], Luoyi Fu[1]

[1]Shanghai Jiaotong University [2]Amazon AWS AI [3]Westlake University [4]IGSNRR, Chinese Academy of Sciences
{zhangtianhang, davendw, xwang8, yiluofu}@sjtu.edu.cn {quln, gqipeng, zhaz}@amazon.com zhangyue@westlake.edu.cn

Introduction

Task: Finding hallucinations in LLM generated text



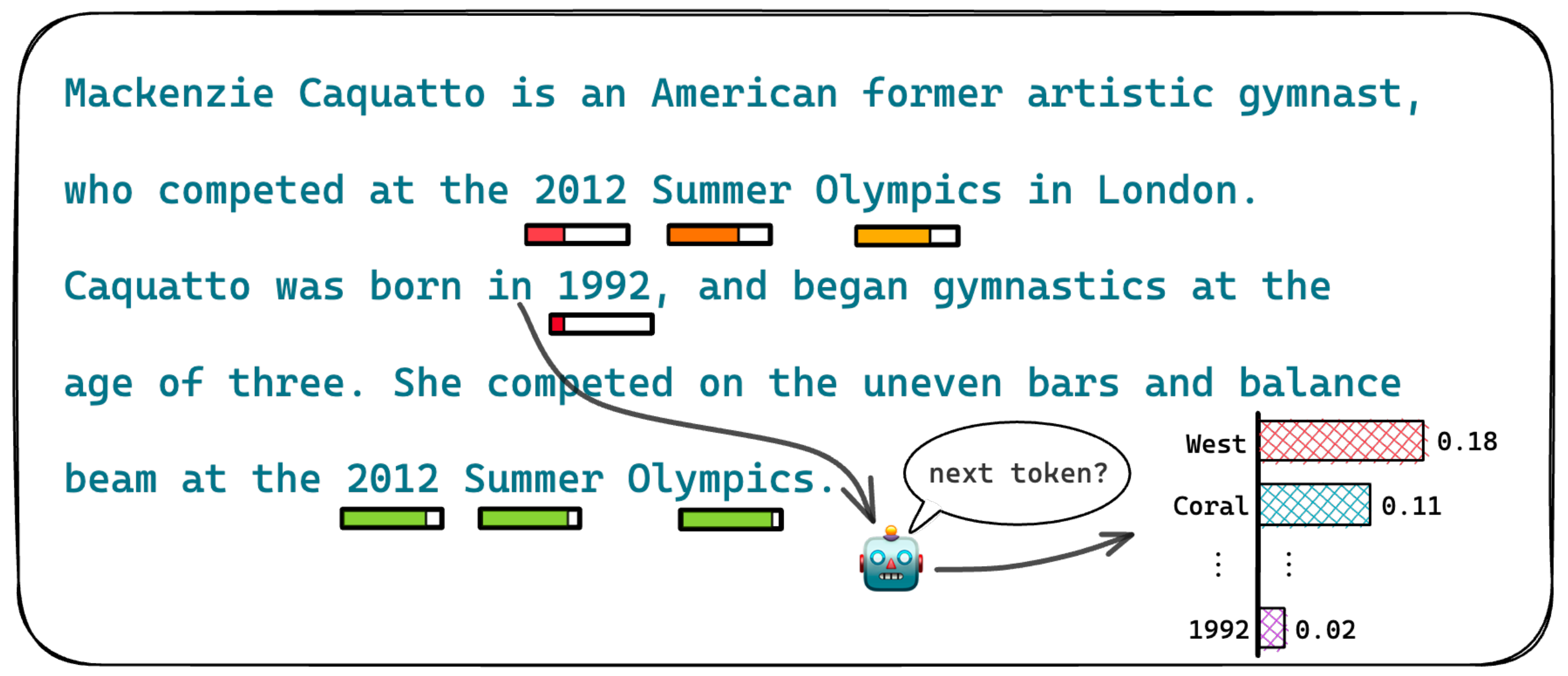
b introduce Mackenzie Caquatto to me

Mackenzie Caquatto is a former artistic gymnast from the United States. Born on September 13, 1989, in Naperville, Illinois, she gained prominence in the gymnastics community for her skill and performances during her competitive career.

Caquatto was a member of the U.S. national gymnastics team and represented her country in various international competitions. She trained at the Naperville Gymnastics Club and later at GAGE Center in Missouri under renowned coach Al Fong.

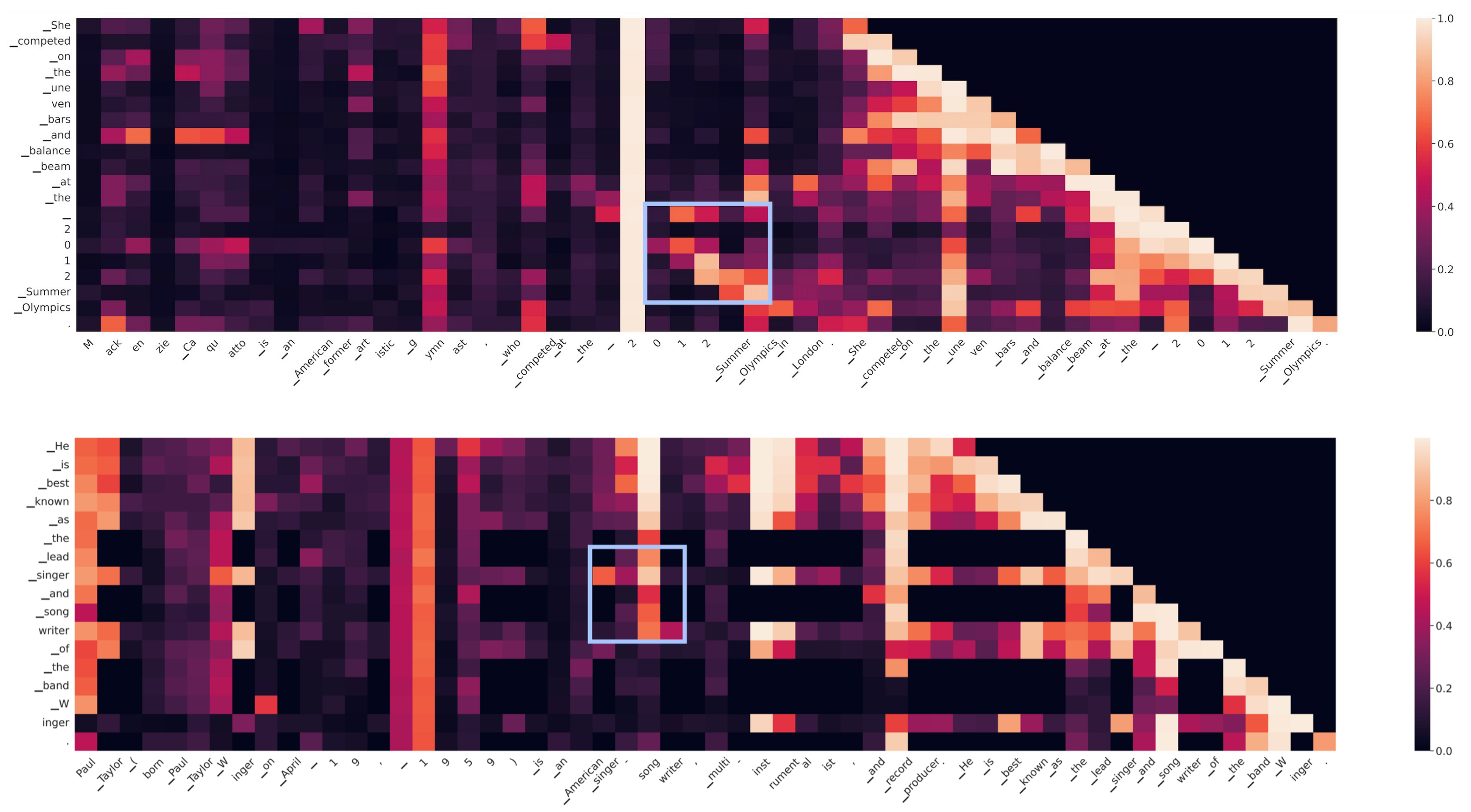
- ❖ Hallucinations can undermine the reliability and trustworthiness of LLMs, especially in scenarios where accuracy and veracity are essential.
- ❖ We focus on hallucinations that conflict with knowledge in the real world.

Problems of Existing Uncertainty-Based Methods



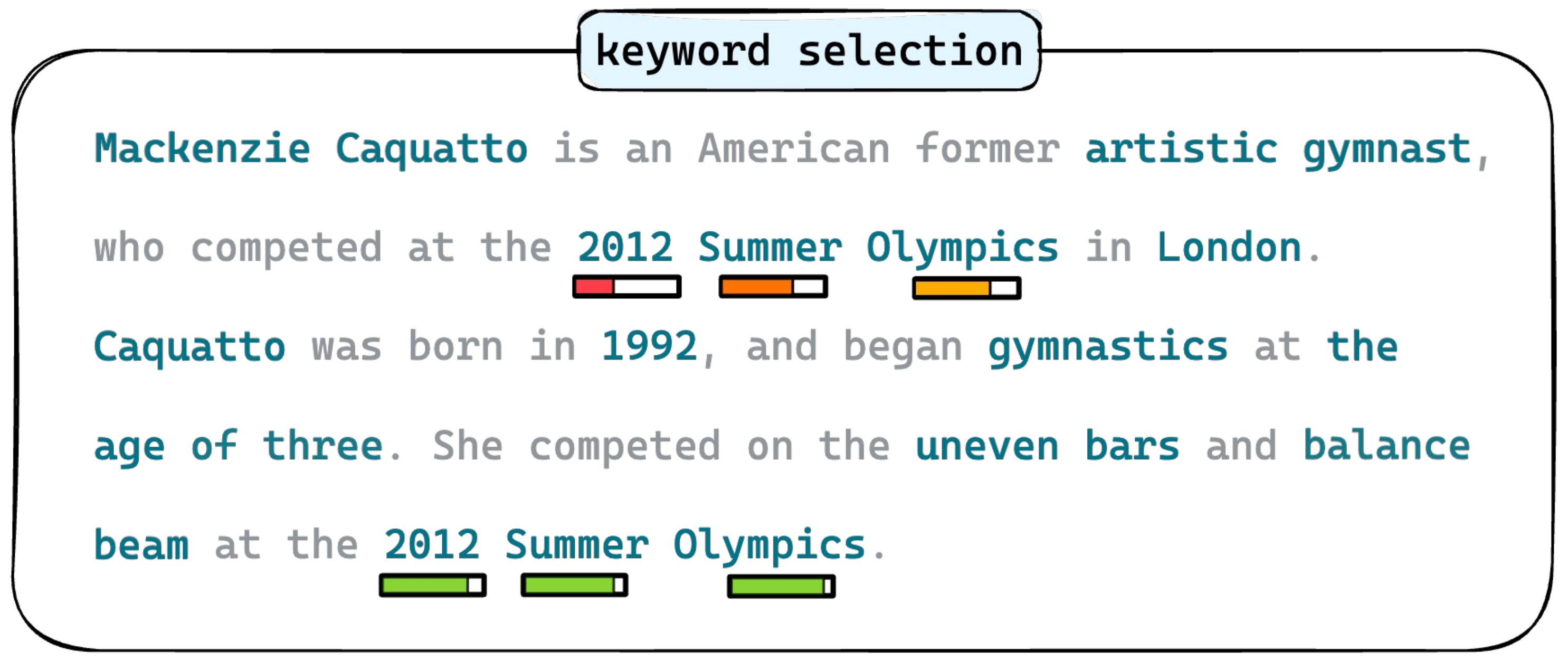
Mackenzie Caquatto is an American former artistic gymnast, who competed at the 2012 Summer Olympics in London. Caquatto was born in 1992, and began gymnastics at the age of three. She competed on the uneven bars and balance beam at the 2012 Summer Olympics.

- ❖ Given a model response, calculate the generation probability of each token. If the text is from a black-box LLM, we can use a proxy model for estimation.
- ❖ However, such method may introduce noise when calculating sentence-level hallucination scores.
- ❖ The probability may exhibit "overconfidence".
- ❖ The probability sometimes demonstrate "underconfidence"



Our Method: Focus

Keyword Selection: The keywords that express salient information will be extracted for calculation

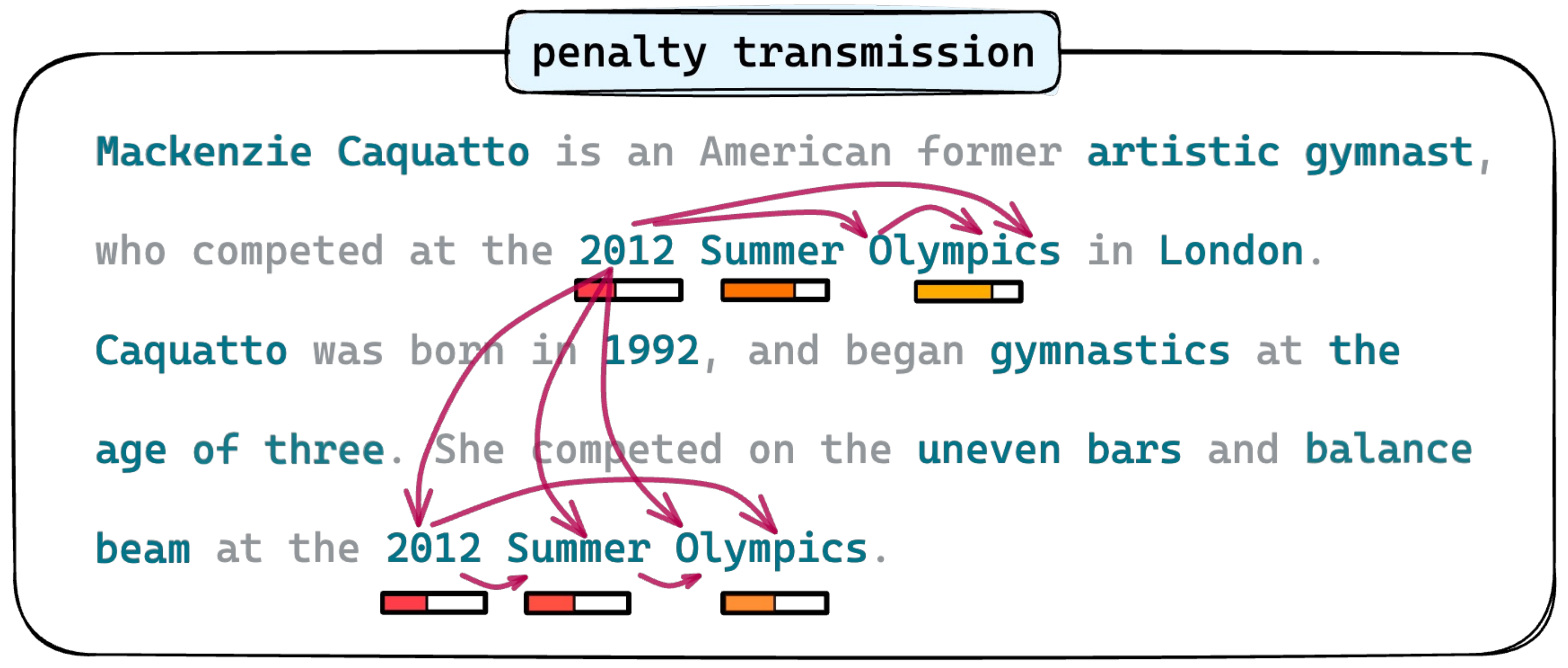


keyword selection

Mackenzie Caquatto is an American former artistic gymnast, who competed at the 2012 Summer Olympics in London. Caquatto was born in 1992, and began gymnastics at the age of three. She competed on the uneven bars and balance beam at the 2012 Summer Olympics.

- ❖ The keywords are the named entities and nouns recognized by Spacy.

Penalty Transmission: The uncertainties of previous tokens will be transmitted to the subsequent ones

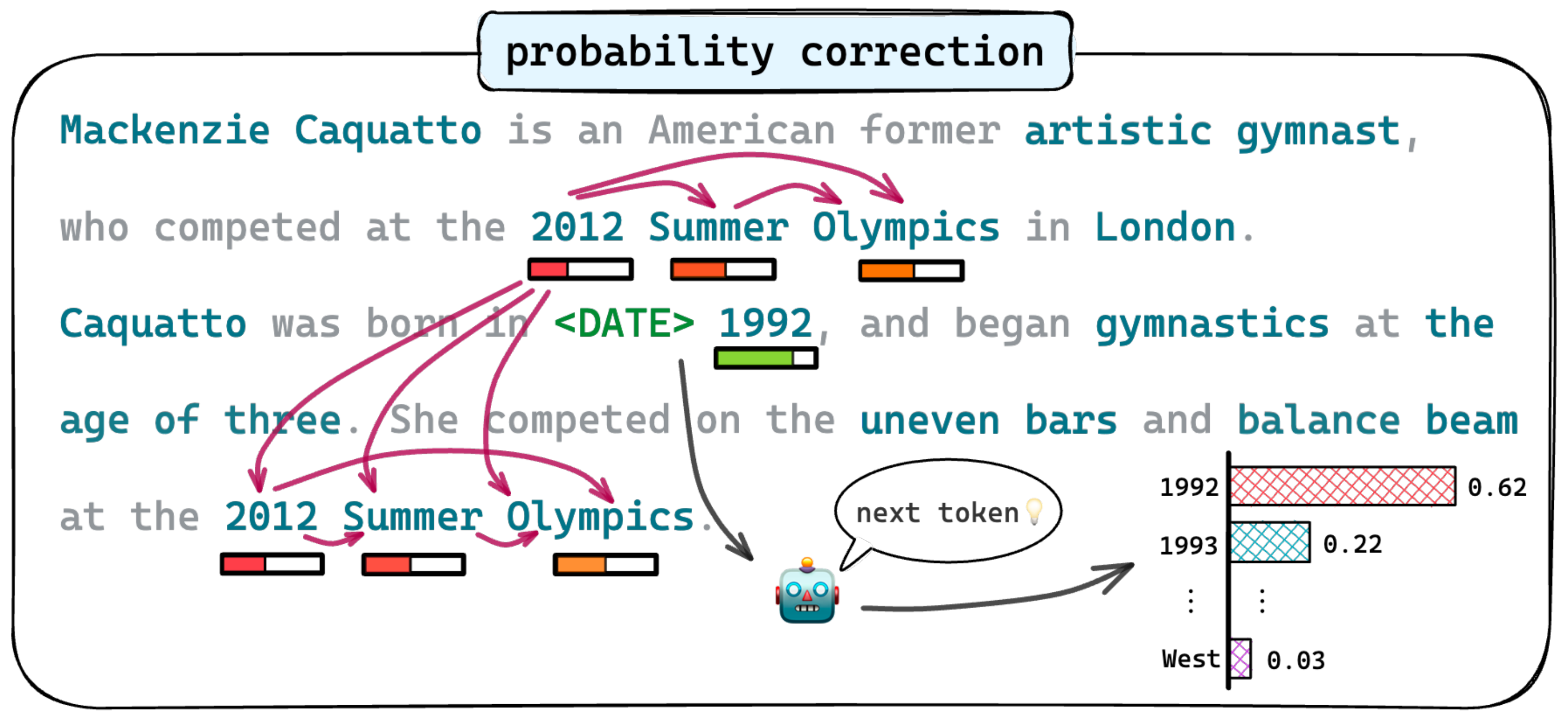


penalty transmission

Mackenzie Caquatto is an American former artistic gymnast, who competed at the 2012 Summer Olympics in London. Caquatto was born in 1992, and began gymnastics at the age of three. She competed on the uneven bars and balance beam at the 2012 Summer Olympics.

- ❖ The attention weight is the max-pooling results for all the layers and attention heads.

Probability Correction: The token generation probability will be adjusted according to the token properties



probability correction

Mackenzie Caquatto is an American former artistic gymnast, who competed at the 2012 Summer Olympics in London. Caquatto was born in <DATE> 1992, and began gymnastics at the age of three. She competed on the uneven bars and balance beam at the 2012 Summer Olympics.

- ❖ The predicted token probability is conditioned on its entity type (if any) and adjusted by its inverse document frequency (IDF).
- ❖ The token frequency is estimated using the sampled RedPajama dataset.

Experiments

Hallucination Detection in Large Language Model

Wikibio-GPT3

Method	Sentence-level Metrics			Passage-level Metrics	
	NonFact	NonFact*	Factual	Pearson	Spearman
GPT-3 Uncertainties					
Avg($-\log p$)	83.21	38.89	53.97	57.04	53.93
Avg($\mathcal{H}$)	80.73	37.09	52.07	55.52	50.87
Max($-\log p$)	87.51	35.88	50.46	57.83	55.69
Max($\mathcal{H}$)	85.75	32.43	50.27	52.48	49.55
SelfCheckGPT					
BERTScore	81.96	45.96	44.23	58.18	55.90
QA	84.26	40.06	48.14	61.07	59.29
Unigram (max)	85.63	41.04	58.47	64.71	64.91
Combination	87.33	44.37	61.83	69.05	67.77
Ours					
LLaMA-7B$_{focus}$	84.26	40.20	57.04	64.47	54.73
LLaMA-13B$_{focus}$	87.90	43.84	62.46	70.62	63.03
LLaMA-30B$_{focus}$	89.79	**48.80**	**65.69**	**77.15**	**73.24**
LLaMA-65B$_{focus}$	**89.94**	48.69	64.90	76.80	73.01

Method	NoFac	NoFac*	Fact	Pear.	Spear.
avg(h)	82.07	41.47	47.22	51.03	37.29
+keyword	83.01	41.57	45.82	56.07	44.77
+penalty	86.68	45.27	54.93	59.08	55.84
+entity type	88.89	46.92	65.12	76.82	71.49
+token idf	**89.79**	**48.80**	**65.69**	**77.15**	**73.24**

LLaMA-30b

Method	NoFac	NoFac*	Fact	Pear.	Spear.
avg(h)	79.72	37.50	32.37	34.00	27.47
+keyword	80.55	37.62	35.13	45.45	38.11
+penalty	87.26	47.22	44.88	47.67	52.01
+entity type	87.11	45.74	57.60	68.25	62.46
+token idf	**88.11**	**46.95**	**58.14**	**68.63**	**64.66**

Falcon-40b

- ❖ Our proposed method achieves SOTA performance on the WikiBio-GPT-3 dataset across various models with different scales.
- ❖ Our method show effectiveness across **22** different proxy models such as OPT, GPT-J and Falcon.

Hallucination Detection in Small Language Models

XSumFaith

Method	NonFact	Fact	Balanced-Acc
avg(h)	92.79	11.75	57.65
+keyword	92.65	14.19	56.24
+penalty	92.34	14.97	57.77
+entity type	94.98	18.46	64.77
+token idf	**95.13**	**18.86**	**64.81**

FRANK

Method	NonFact	Fact	Balanced-Acc
avg($-logp$)	89.82	79.00	78.79
+penalty	89.87	78.37	79.46
+entity type	**90.44**	79.78	80.31
+token idf	90.12	**80.00**	**80.70**

- ❖ Our method also shows effectiveness in detecting hallucinations within summaries generated by small language models.