JOHNS HOPKINS
WHITING SCHOOL
*of* ENGINEERING

# Verifiable by Design: Aligning Language Models to Quote from Pre-Training Data

Jingyu (Jack) Zhang

Marc Marone

Tianjian Li

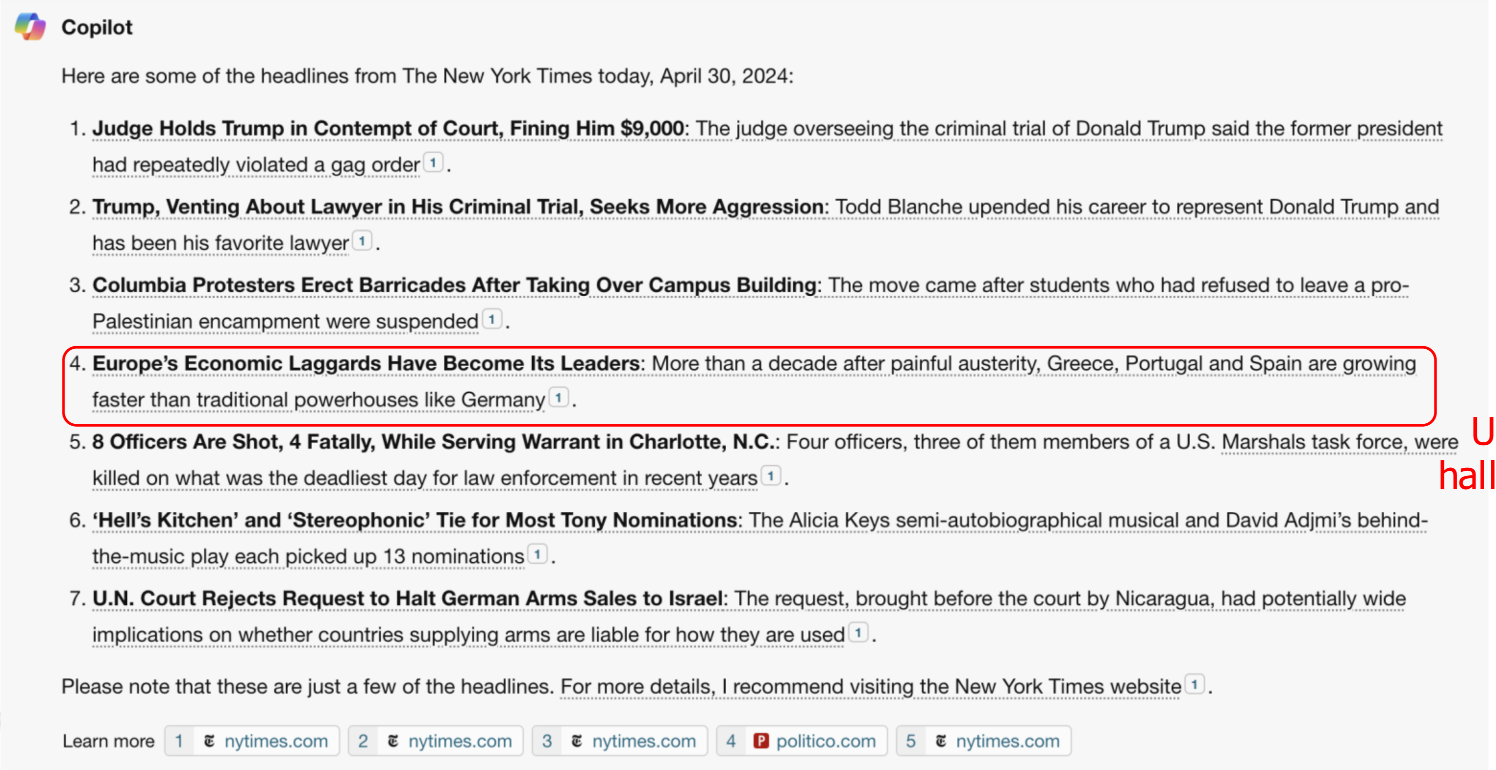
Benjamin Van Durme*

Daniel Khashabi*

*Equal Advising

# LLMs are powerful, but not always reliable

- Known to produce seemingly plausible but erroneous outputs, aka hallucination



Copilot

Here are some of the headlines from The New York Times today, April 30, 2024:

1. **Judge Holds Trump in Contempt of Court, Fining Him $9,000**: The judge overseeing the criminal trial of Donald Trump said the former president had repeatedly violated a gag order[1].
2. **Trump, Venting About Lawyer in His Criminal Trial, Seeks More Aggression**: Todd Blanche upended his career to represent Donald Trump and has been his favorite lawyer[1].
3. **Columbia Protesters Erect Barricades After Taking Over Campus Building**: The move came after students who had refused to leave a pro-Palestinian encampment were suspended[1].
4. **Europe's Economic Laggards Have Become Its Leaders**: More than a decade after painful austerity, Greece, Portugal and Spain are growing faster than traditional powerhouses like Germany[1].
5. **8 Officers Are Shot, 4 Fatally, While Serving Warrant in Charlotte, N.C.**: Four officers, three of them members of a U.S. Marshals task force, were killed on what was the deadliest day for law enforcement in recent years[1].
6. **'Hell's Kitchen' and 'Stereophonic' Tie for Most Tony Nominations**: The Alicia Keys semi-autobiographical musical and David Adjmi's behind-the-music play each picked up 13 nominations[1].
7. **U.N. Court Rejects Request to Halt German Arms Sales to Israel**: The request, brought before the court by Nicaragua, had potentially wide implications on whether countries supplying arms are liable for how they are used[1].

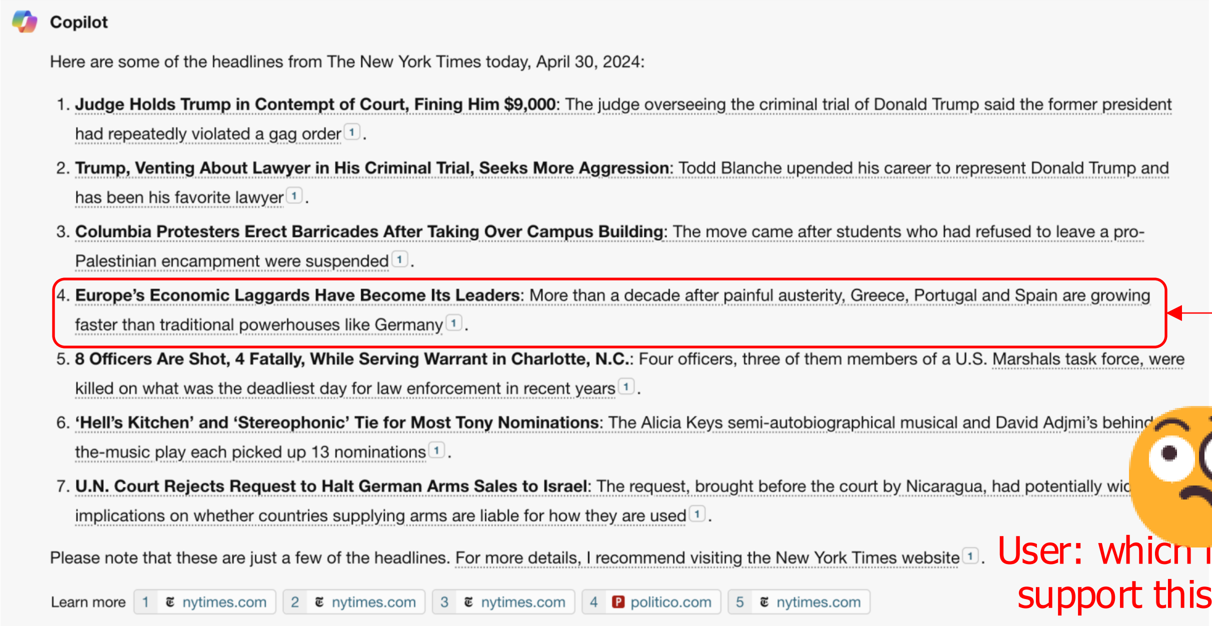
Please note that these are just a few of the headlines. For more details, I recommend visiting the New York Times website[1].

Learn more 1 nytimes.com 2 nytimes.com 3 nytimes.com 4 politico.com 5 nytimes.com

User: Is this hallucination???

# Verifiability is challenging

- To trust LLMs, users need to **verify** generated claims against trusted resources
- However, non-trivial to fact-check lengthy free-form generation!



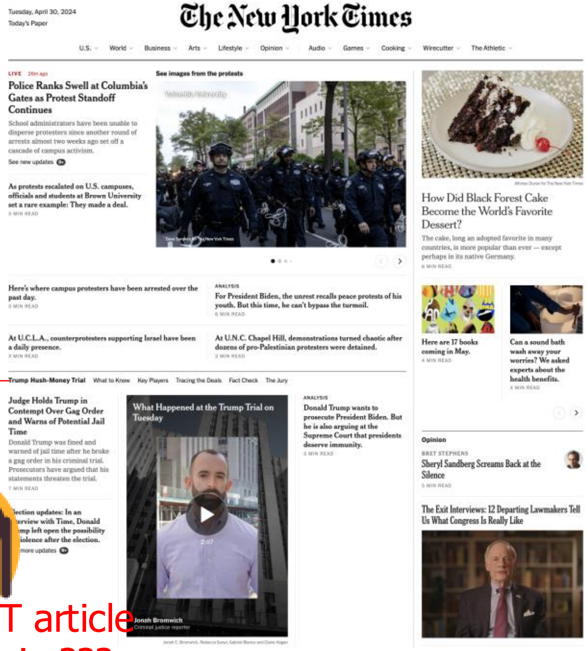
**Copilot**

Here are some of the headlines from The New York Times today, April 30, 2024:

1. **Judge Holds Trump in Contempt of Court, Fining Him $9,000**: The judge overseeing the criminal trial of Donald Trump said the former president had repeatedly violated a gag order [1].
2. **Trump, Venting About Lawyer in His Criminal Trial, Seeks More Aggression**: Todd Blanche upended his career to represent Donald Trump and has been his favorite lawyer [1].
3. **Columbia Protesters Erect Barricades After Taking Over Campus Building**: The move came after students who had refused to leave a pro-Palestinian encampment were suspended [1].
4. **Europe's Economic Laggards Have Become Its Leaders**: More than a decade after painful austerity, Greece, Portugal and Spain are growing faster than traditional powerhouses like Germany [1].
5. **8 Officers Are Shot, 4 Fatally, While Serving Warrant in Charlotte, N.C.**: Four officers, three of them members of a U.S. Marshals task force, were killed on what was the deadliest day for law enforcement in recent years [1].
6. **'Hell's Kitchen' and 'Stereophonic' Tie for Most Tony Nominations**: The Alicia Keys semi-autobiographical musical and David Adjmi's behind the-music play each picked up 13 nominations [1].
7. **U.N. Court Rejects Request to Halt German Arms Sales to Israel**: The request, brought before the court by Nicaragua, had potentially wic implications on whether countries supplying arms are liable for how they are used [1].

Please note that these are just a few of the headlines. For more details, I recommend visiting the New York Times website [1].

Learn more 1 nytimes.com 2 nytimes.com 3 nytimes.com 4 politico.com 5 nytimes.com



Tuesday, April 30, 2024
Today's Paper

The New York Times

U.S. World Business Arts Lifestyle Opinion Audio Games Cooking Wirecutter The Athletic

LIVE 26m ago

**Police Ranks Swell at Columbia's Gates as Protest Standoff Continues**

School administrators have been unable to disperse protesters since another round of arrests almost two weeks ago set off a cascade of campus activism.

See new updates

As protests escalated on U.S. campuses, officials and students at Brown University set a rare example: They made a deal.

See images from the protests

Here's where campus protesters have been arrested over the past day.

ANALYSIS
For President Biden, the unrest recalls peace protests of his youth. But this time, he can't bypass the turmoil.

At U.C.L.A., counterprotesters supporting Israel have been a daily presence.

At U.N.C. Chapel Hill, demonstrations turned chaotic after dozens of pro-Palestinian protesters were detained.

Trump Hush-Money Trial What to Know Key Players Tracing the Deals Fact Check The Jury

**Judge Holds Trump in Contempt Over Gag Order and Warns of Potential Jail Time**

Donald Trump was fined and warned of jail time after he broke a gag order in his criminal trial. Prosecutors have argued that his statements threaten the trial.

lection updates: In an terview with Time, Donald mp left open the possibility iolence after the election.

What Happened at the Trump Trial on Tuesday

Jonah Bromwich
Criminal justice reporter

ANALYSIS
Donald Trump wants to prosecute President Biden. But he is also arguing at the Supreme Court that presidents deserve immunity.

How Did Black Forest Cake Become the World's Favorite Dessert?

The cake, long an adopted favorite in many countries, is more popular than ever — except perhaps in its native Germany.

Here are 17 books coming in May.

Can a sound bath wash away your worries? We asked experts about the health benefits.

Opinion

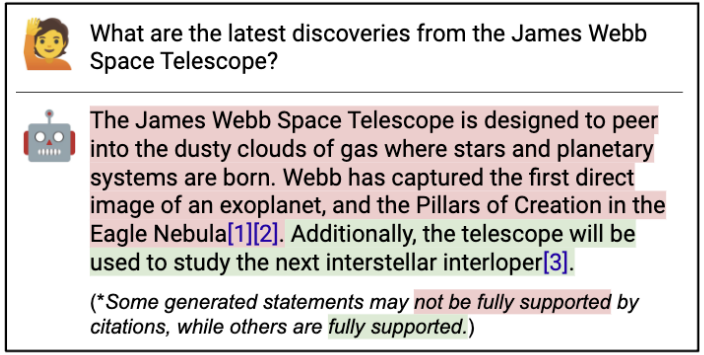
BRET STEPHENS
Sheryl Sandberg Screams Back at the Silence

The Exit Interviews: 12 Departing Lawmakers Tell Us What Congress Is Really Like

User: which NYT article support this claim???

# Existing methods for verifiability

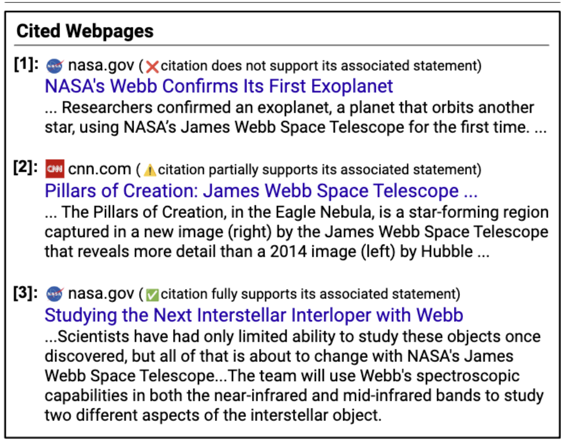
- Search/retrieval + citation helps, but is not *guaranteed* to be accurate
- Non-trivial to verify whether text is supported by citation



🙋 What are the latest discoveries from the James Webb Space Telescope?

🤖 The James Webb Space Telescope is designed to peer into the dusty clouds of gas where stars and planetary systems are born. Webb has captured the first direct image of an exoplanet, and the Pillars of Creation in the Eagle Nebula[1][2]. Additionally, the telescope will be used to study the next interstellar interloper[3].

(**Some generated statements may not be fully supported by citations, while others are fully supported.*)

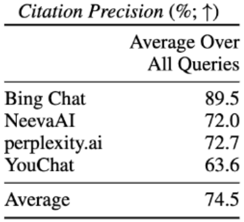


**Cited Webpages**

[1]: nasa.gov (❌citation does not support its associated statement)
NASA's Webb Confirms Its First Exoplanet
... Researchers confirmed an exoplanet, a planet that orbits another star, using NASA's James Webb Space Telescope for the first time. ...

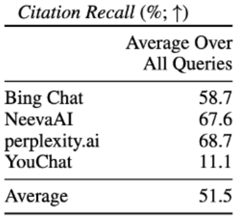
[2]: cnn.com (⚠️citation partially supports its associated statement)
Pillars of Creation: James Webb Space Telescope ...
... The Pillars of Creation, in the Eagle Nebula, is a star-forming region captured in a new image (right) by the James Webb Space Telescope that reveals more detail than a 2014 image (left) by Hubble ...

[3]: nasa.gov (✅citation fully supports its associated statement)
Studying the Next Interstellar Interloper with Webb
...Scientists have had only limited ability to study these objects once discovered, but all of that is about to change with NASA's James Webb Space Telescope...The team will use Webb's spectroscopic capabilities in both the near-infrared and mid-infrared bands to study two different aspects of the interstellar object.



*Citation Precision* (%; ↑)

| | Average Over All Queries |
|---|---|
| Bing Chat | 89.5 |
| NeevaAI | 72.0 |
| perplexity.ai | 72.7 |
| YouChat | 63.6 |
| Average | 74.5 |



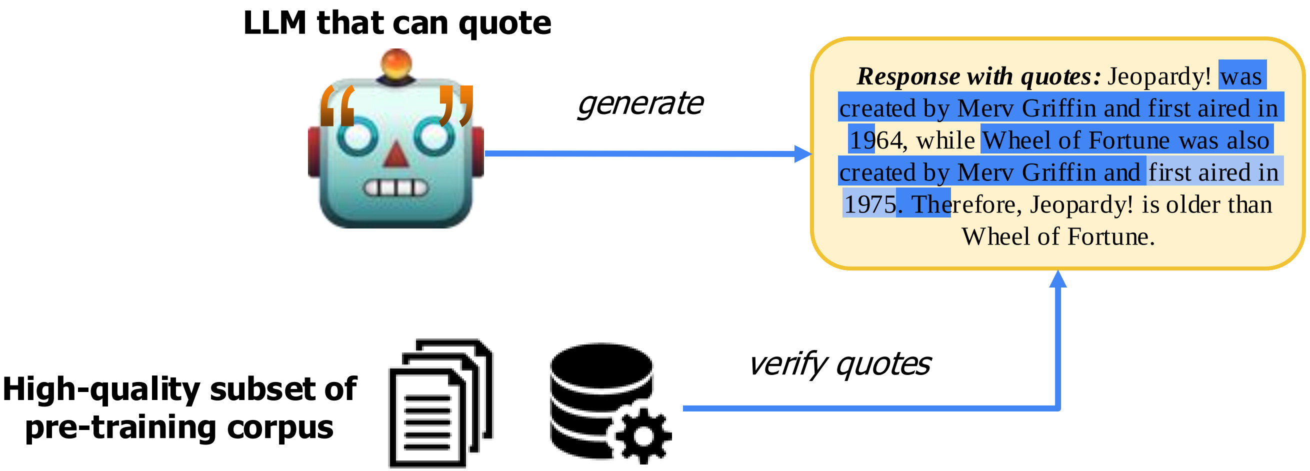
*Citation Recall* (%; ↑)

| | Average Over All Queries |
|---|---|
| Bing Chat | 58.7 |
| NeevaAI | 67.6 |
| perplexity.ai | 68.7 |
| YouChat | 11.1 |
| Average | 51.5 |

Nelson Liu, Tianyi Zhang, and Percy Liang. 2023. Evaluating Verifiability in Generative Search Engines. In *Findings of the Association for Computational Linguistics: EMNLP 2023*

# Verifiability by Quoting

- We propose increasing verifiability by generating **verbatim quotes** from high-quality sources of pre-training data, such as Wikipedia.



**LLM that can quote**

*generate*

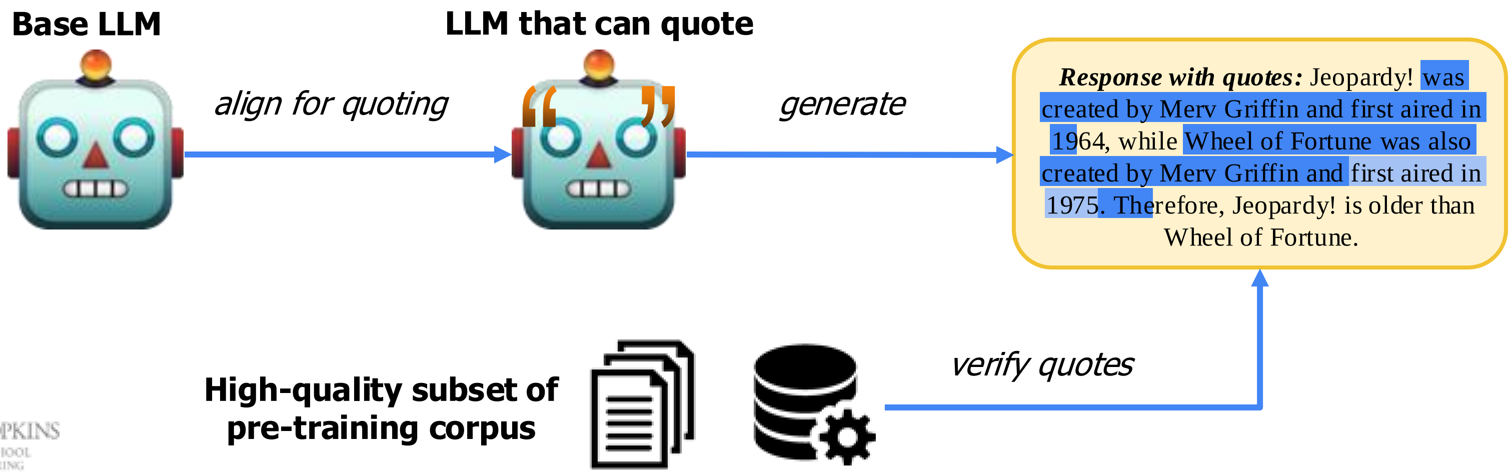
***Response with quotes:*** Jeopardy! was created by Merv Griffin and first aired in 1964, while Wheel of Fortune was also created by Merv Griffin and first aired in 1975. Therefore, Jeopardy! is older than Wheel of Fortune.

*verify quotes*

**High-quality subset of pre-training corpus**

# Verifiability by Quoting

- We propose increasing verifiability by generating **verbatim quotes** from high-quality sources of pre-training data, such as Wikipedia.
- **Quote-Tuning**: aligning LLMs to quote from their pre-training data!
  - Make the model **prefer generation with more quotes over less quotes**!
  - Align for quoting using preference optimization algorithms such as DPO



**Base LLM**

*align for quoting*

**LLM that can quote**

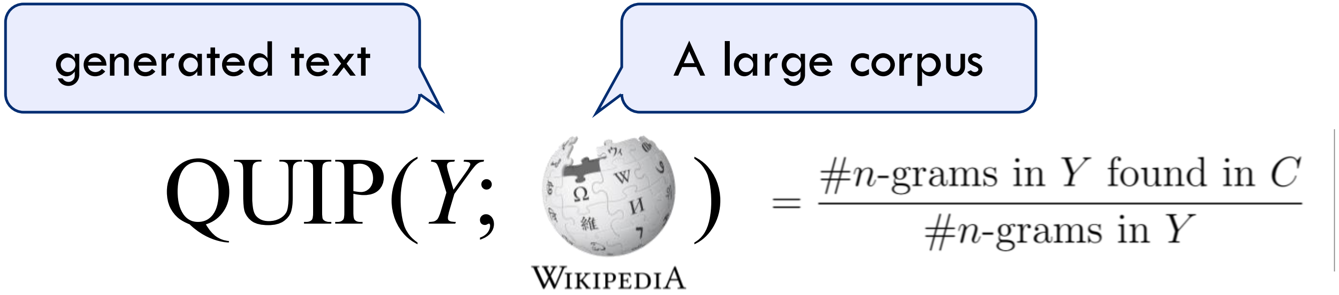
*generate*

***Response with quotes:*** Jeopardy! was created by Merv Griffin and first aired in 1964, while Wheel of Fortune was also created by Merv Griffin and first aired in 1975. Therefore, Jeopardy! is older than Wheel of Fortune.

*verify quotes*

**High-quality subset of pre-training corpus**

# Background: measuring quoting at scale



generated text

A large corpus

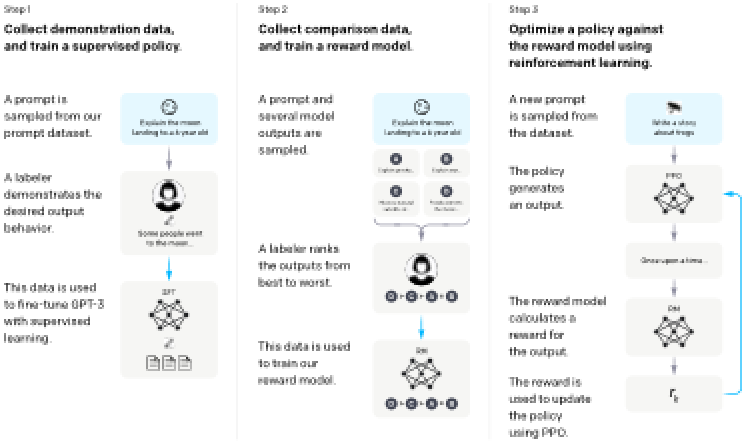
$$\mathrm{QUIP}(Y;\ \text{Wikipedia}) = \frac{\#n\text{-grams in } Y \text{ found in } C}{\#n\text{-grams in } Y}$$

- Calculate quoting precision with **QUIP-Score** (Weller et al., 2024)
- QUIP-Score is backed by efficient membership testing with **Data Portraits** (Marone and Van Durme, 2023)
  - Bloom filter data sketching on large corpora

# Background: preference optimization

- Reinforcement Learning from Human Feedback (RLHF) aligns language models to…

(Ouyang et al., 2022)



Step 1

**Collect demonstration data, and train a supervised policy.**

A prompt is sampled from our prompt dataset.

A labeler demonstrates the desired output behavior.

This data is used to fine-tune GPT-3 with supervised learning.

Step 2

**Collect comparison data, and train a reward model.**

A prompt and several model outputs are sampled.

A labeler ranks the outputs from best to worst.

This data is used to train our reward model.

Step 3

**Optimize a policy against the reward model using reinforcement learning.**

A new prompt is sampled from the dataset.

The policy generates an output.

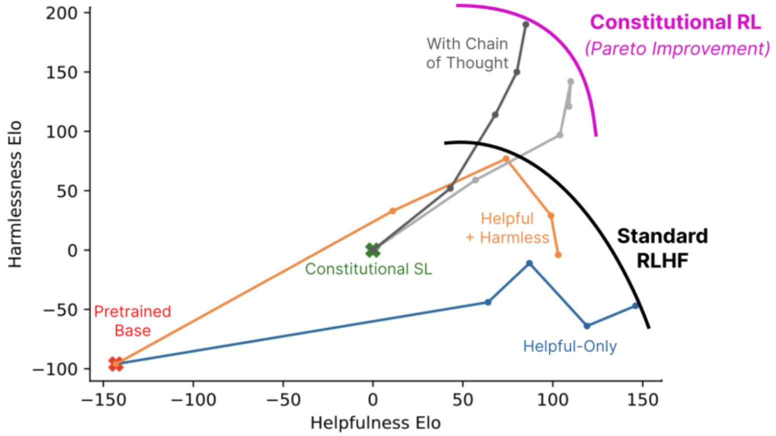
The reward model calculates a reward for the output.

The reward is used to update the policy using PPO.

Explain the moon landing to a 6 year old
Some people went to the moon...
SFT
RM
Write a story about frogs
PPO
Once upon a time...
RM
$r_k$

Follow instructions (+helpfulness)

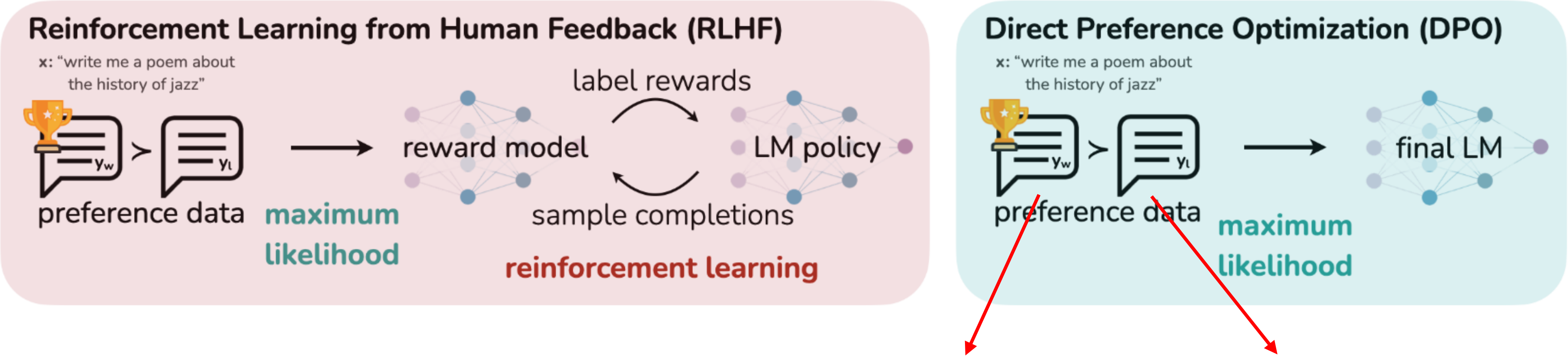
(Bai et al., 2022)



Reduce harmful outputs

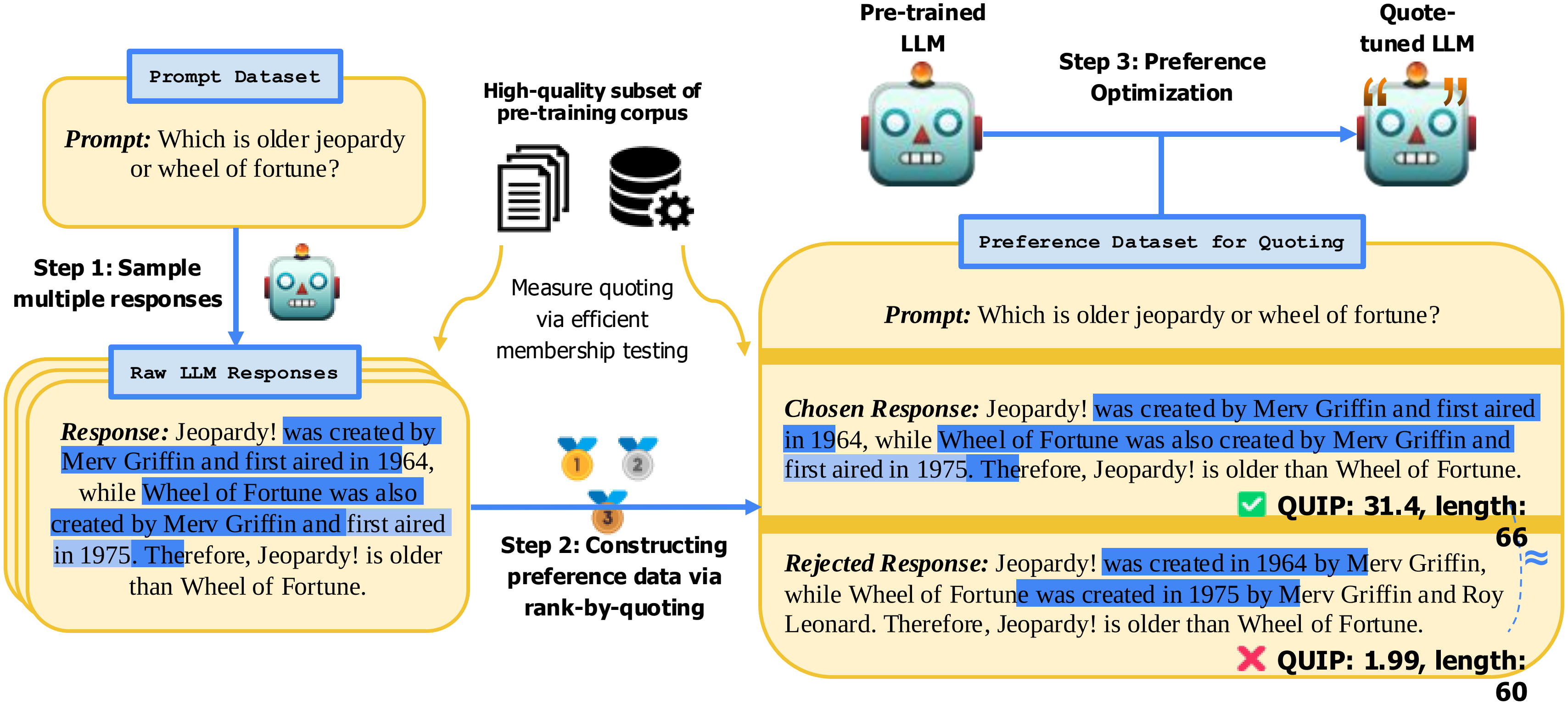
# Background: preference optimization

- **Direct Preference Optimization** (Rafailov et al., 2023) simplifies the RLHF pipeline, optimizing a dataset of preference without RL


Reinforcement Learning from Human Feedback (RLHF)
x: "write me a poem about the history of jazz"
$y_w$
>
$y_l$
preference data
maximum likelihood
reward model
label rewards
sample completions
LM policy
reinforcement learning
Direct Preference Optimization (DPO)
x: "write me a poem about the history of jazz"
$y_w$
>
$y_l$
preference data
maximum likelihood
final LM

$$\mathcal{L}_{\text{DPO}}(\pi_\theta; \pi_{\text{ref}}) = -\mathbb{E}_{(x, y_w, y_l) \sim \mathcal{D}} \left[ \log \sigma \left( \beta \log \frac{\pi_\theta(y_w \mid x)}{\pi_{\text{ref}}(y_w \mid x)} - \beta \log \frac{\pi_\theta(y_l \mid x)}{\pi_{\text{ref}}(y_l \mid x)} \right) \right]$$

# Our method: Quote-Tuning



Prompt Dataset

***Prompt:*** Which is older jeopardy or wheel of fortune?

**Step 1: Sample multiple responses**

Raw LLM Responses

***Response:*** Jeopardy! was created by Merv Griffin and first aired in 1964, while Wheel of Fortune was also created by Merv Griffin and first aired in 1975. Therefore, Jeopardy! is older than Wheel of Fortune.

**High-quality subset of pre-training corpus**

Measure quoting via efficient membership testing

**Step 2: Constructing preference data via rank-by-quoting**

**Pre-trained LLM**

**Step 3: Preference Optimization**

**Quote-tuned LLM**

Preference Dataset for Quoting

***Prompt:*** Which is older jeopardy or wheel of fortune?

***Chosen Response:*** Jeopardy! was created by Merv Griffin and first aired in 1964, while Wheel of Fortune was also created by Merv Griffin and first aired in 1975. Therefore, Jeopardy! is older than Wheel of Fortune.

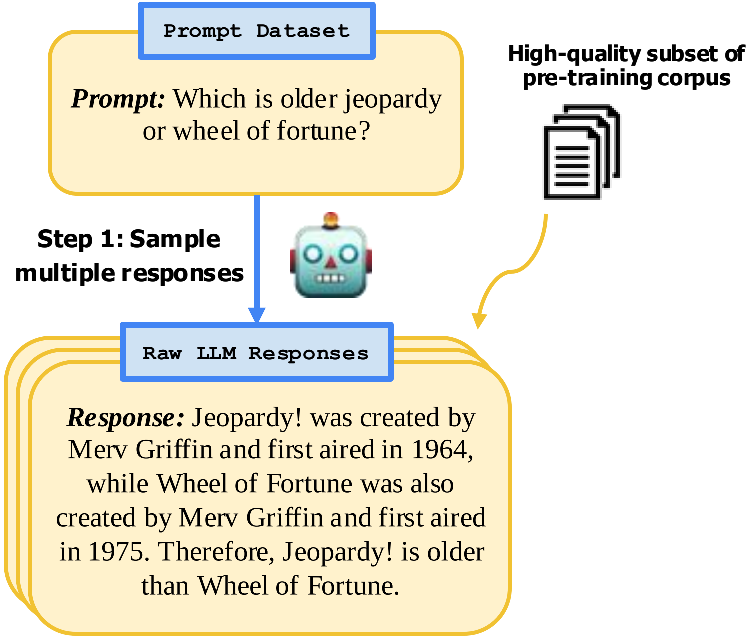
✅ **QUIP: 31.4, length: 66**

≈

***Rejected Response:*** Jeopardy! was created in 1964 by Merv Griffin, while Wheel of Fortune was created in 1975 by Merv Griffin and Roy Leonard. Therefore, Jeopardy! is older than Wheel of Fortune.

❌ **QUIP: 1.99, length: 60**

# Step 1: Sample candidate responses



Prompt Dataset

***Prompt:*** Which is older jeopardy or wheel of fortune?

High-quality subset of pre-training corpus

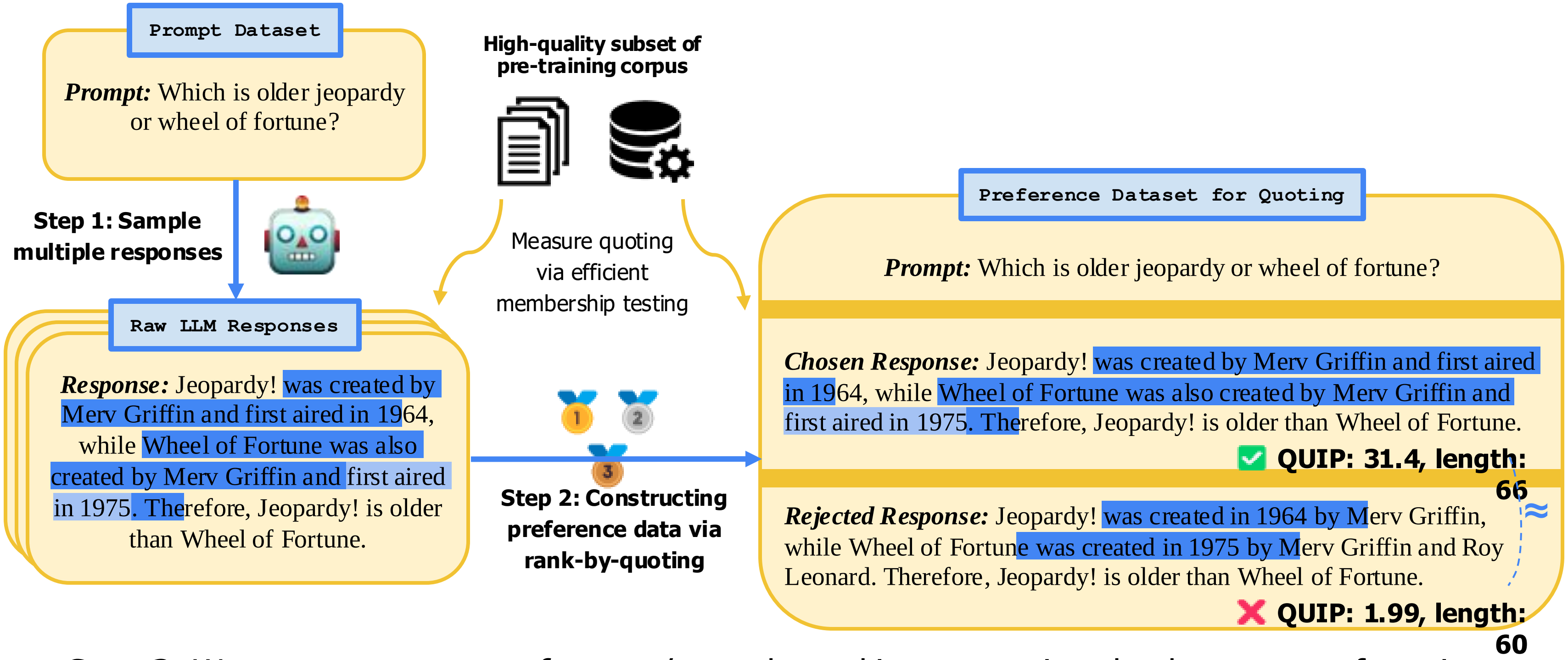
**Step 1: Sample multiple responses**

Raw LLM Responses

***Response:*** Jeopardy! was created by Merv Griffin and first aired in 1964, while Wheel of Fortune was also created by Merv Griffin and first aired in 1975. Therefore, Jeopardy! is older than Wheel of Fortune.

**Step 1.** Generate completions from an LLM (e.g. using QA pairs or text completions)

# Step 2: Synthesize Dataset by Filtering



Prompt Dataset

***Prompt:*** Which is older jeopardy or wheel of fortune?

**Step 1: Sample multiple responses**

Raw LLM Responses

***Response:*** Jeopardy! was created by Merv Griffin and first aired in 1964, while Wheel of Fortune was also created by Merv Griffin and first aired in 1975. Therefore, Jeopardy! is older than Wheel of Fortune.

**High-quality subset of pre-training corpus**

Measure quoting via efficient membership testing

**Step 2: Constructing preference data via rank-by-quoting**

Preference Dataset for Quoting

***Prompt:*** Which is older jeopardy or wheel of fortune?

***Chosen Response:*** Jeopardy! was created by Merv Griffin and first aired in 1964, while Wheel of Fortune was also created by Merv Griffin and first aired in 1975. Therefore, Jeopardy! is older than Wheel of Fortune.

✅ **QUIP: 31.4, length: 66**

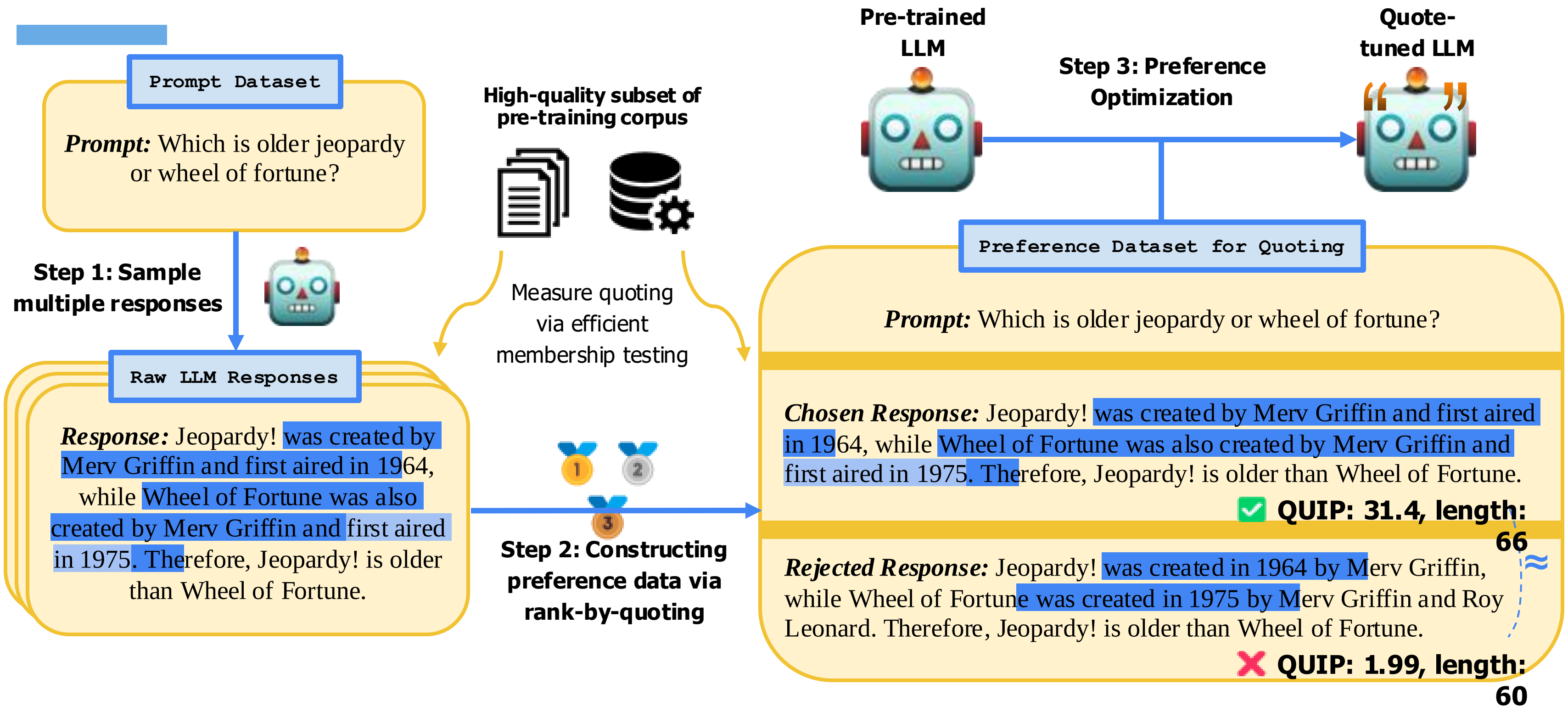
≈

***Rejected Response:*** Jeopardy! was created in 1964 by Merv Griffin, while Wheel of Fortune was created in 1975 by Merv Griffin and Roy Leonard. Therefore, Jeopardy! is older than Wheel of Fortune.

❌ **QUIP: 1.99, length: 60**

**Step 2.** We can construct a *preference dataset* by ranking generations by the amount of quoting (**QUIP-Score; Weller et al., EACL 2024**)

# Step 3: Making the model prefer more quoting!



Prompt Dataset

*Prompt:* Which is older jeopardy or wheel of fortune?

Step 1: Sample multiple responses

Raw LLM Responses

*Response:* Jeopardy! was created by Merv Griffin and first aired in 1964, while Wheel of Fortune was also created by Merv Griffin and first aired in 1975. Therefore, Jeopardy! is older than Wheel of Fortune.

High-quality subset of pre-training corpus

Measure quoting via efficient membership testing

Step 2: Constructing preference data via rank-by-quoting

Pre-trained LLM

Step 3: Preference Optimization

Quote-tuned LLM

Preference Dataset for Quoting

*Prompt:* Which is older jeopardy or wheel of fortune?

*Chosen Response:* Jeopardy! was created by Merv Griffin and first aired in 1964, while Wheel of Fortune was also created by Merv Griffin and first aired in 1975. Therefore, Jeopardy! is older than Wheel of Fortune.

✅ **QUIP: 31.4, length: 66**

≈

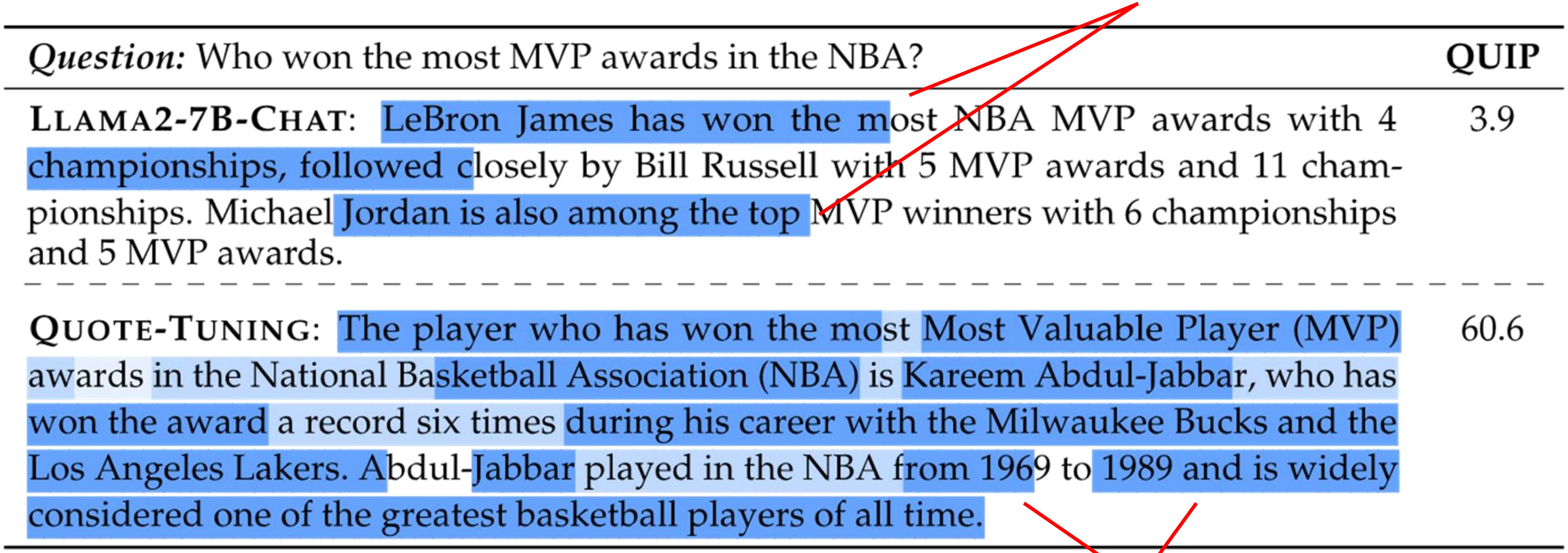
*Rejected Response:* Jeopardy! was created in 1964 by Merv Griffin, while Wheel of Fortune was created in 1975 by Merv Griffin and Roy Leonard. Therefore, Jeopardy! is older than Wheel of Fortune.

❌ **QUIP: 1.99, length: 60**

**Step 3.** Tune a model to prefer more quotes with *direct preference optimization* (DPO)!

# Quote-Tuning significantly increases amount of quoting

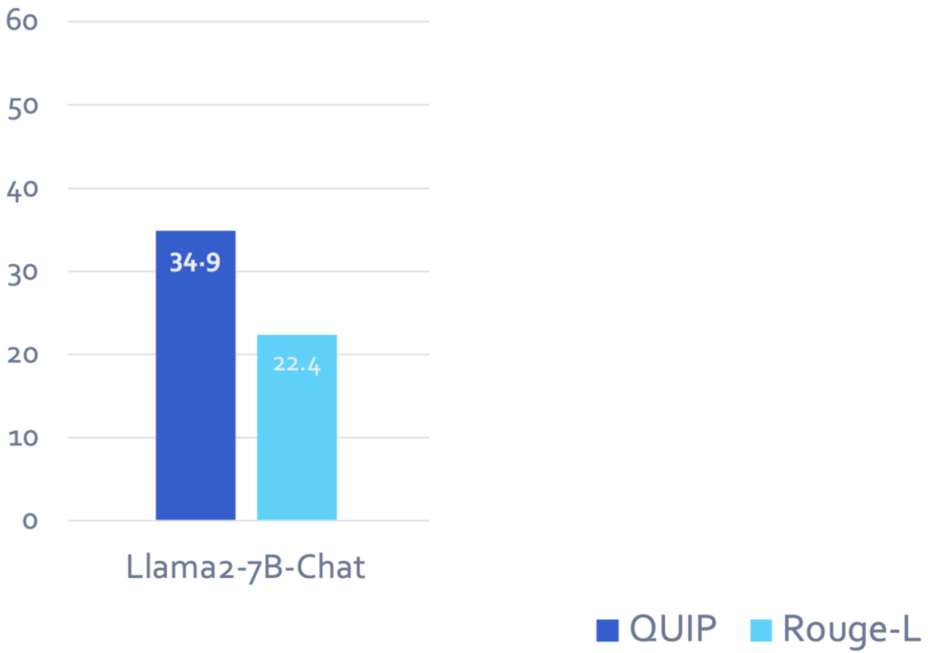
sparse quotes



| *Question:* Who won the most MVP awards in the NBA? | QUIP |
|---|---|
| LLAMA2-7B-CHAT: LeBron James has won the most NBA MVP awards with 4 championships, followed closely by Bill Russell with 5 MVP awards and 11 championships. Michael Jordan is also among the top MVP winners with 6 championships and 5 MVP awards. | 3.9 |
| QUOTE-TUNING: The player who has won the most Most Valuable Player (MVP) awards in the National Basketball Association (NBA) is Kareem Abdul-Jabbar, who has won the award a record six times during his career with the Milwaukee Bucks and the Los Angeles Lakers. Abdul-Jabbar played in the NBA from 1969 to 1989 and is widely considered one of the greatest basketball players of all time. | 60.6 |

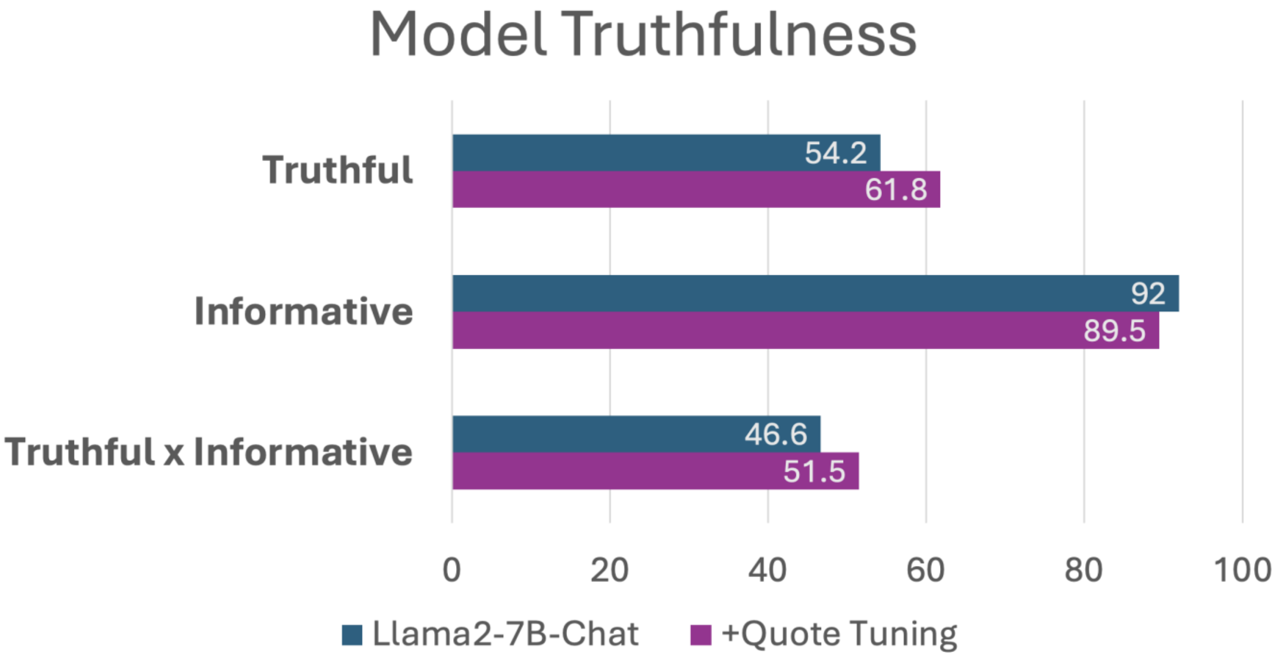
generation is densely quoted!

# Same Quality, More Quoting!



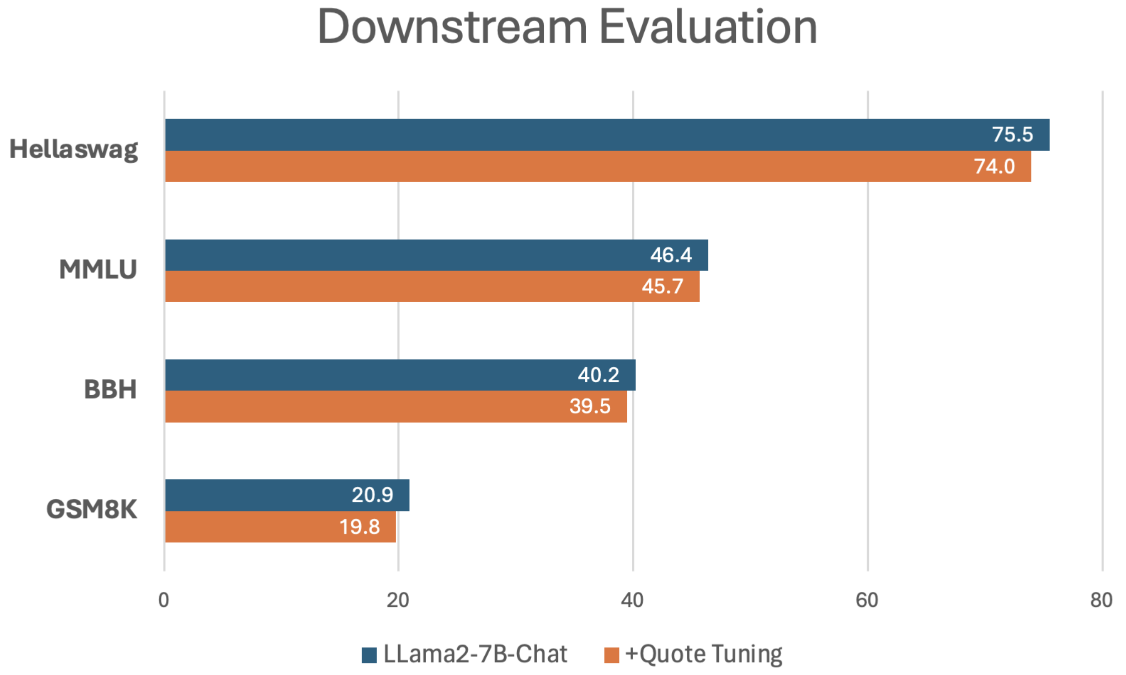
You can obtain models that provide more quoted responses, without losing accuracy.

# Quote-Tuning improves truthfulness



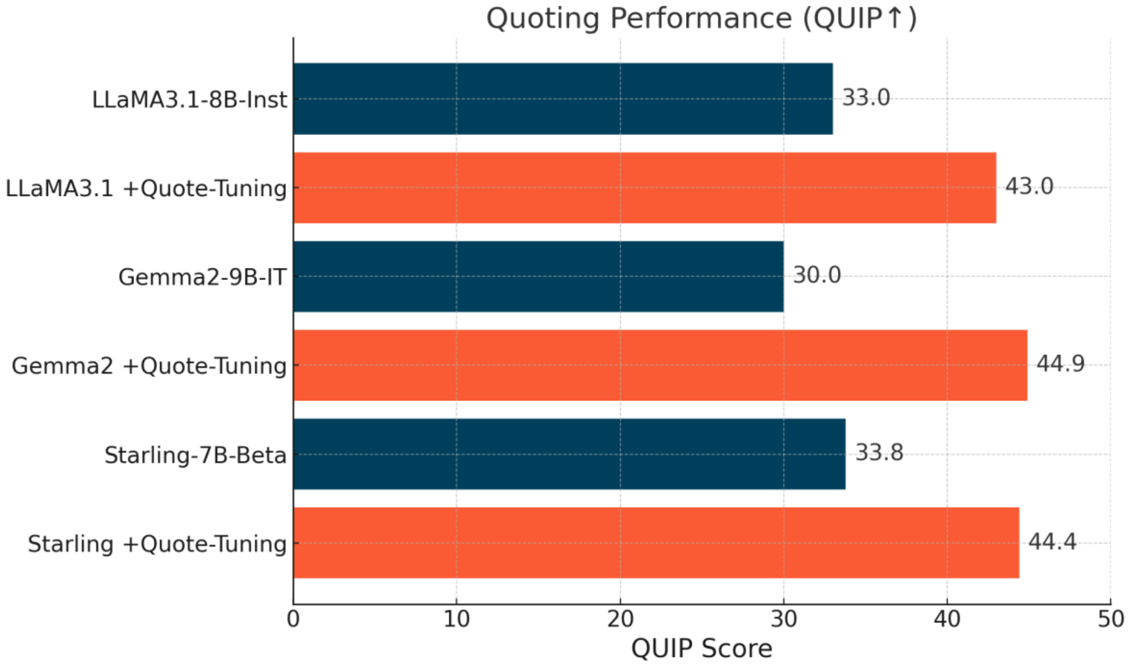
Dataset: TruthfulQA (Lin et al., 2021)

# Downstream Evaluation



Quote-Tuning significantly improves quoting with only a minor sacrifice on general performance (<2 points across the board).

# Quote-Tuning is effective across model families



# Summary and Future Directions

- LLMs can be aligned to quote from known sources observed in their pre-training data

- Quoting makes the verifiability question trivial

- Open questions and future directions:
  - How do we incentivize quoting when it matters?
  - How to ensure long, contiguous quotes?
  - How to generalize to reasoning problems, and more general settings?
  - …

# Thanks for listening!

Questions?

<-link to preprint