

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



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

- 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](#).
- 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](#).
- 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](#).
- 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](#).
- 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](#).
- '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](#).
- 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](#).



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!



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

- 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](#).
- 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](#).
- 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](#).
- 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](#).
- 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](#).
- 'Hell's Kitchen' and 'Stereophonic' Tie for Most Tony Nominations:** The Alicia Keys semi-autobiographical musical and David Adjmi's biopic play each picked up 13 nominations [1](#).
- 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](#).

User: which NYT article support this claim???

The screenshot shows the front page of The New York Times from April 30, 2024. At the top, there's a navigation bar with links for U.S., World, Business, Arts, Lifestyle, Opinion, Audio, Games, Cooking, Wricutter, and The Athletic. Below the navigation is a large headline: "Police Banks Swarm at Columbia's Gates as Protest Standoff Continues". To the right of this headline is a photo of police officers in riot gear. Further down the page, there's another headline: "Trump, Venting About Lawyer in His Criminal Trial, Seeks More Aggression". To the right of this is a photo of Donald Trump. A large, yellow emoji of a face with a thoughtful expression is overlaid on the page, positioned between the two main headlines. The page continues with other news items, including "Columbia Protesters Erect Barricades After Taking Over Campus Building", "Europe's Economic Laggards Have Become Its Leaders", "8 Officers Are Shot, 4 Fatally, While Serving Warrant in Charlotte, N.C.", "'Hell's Kitchen' and 'Stereophonic' Tie for Most Tony Nominations", and "U.N. Court Rejects Request to Halt German Arms Sales to Israel". Each news item includes a thumbnail image and a brief description.

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

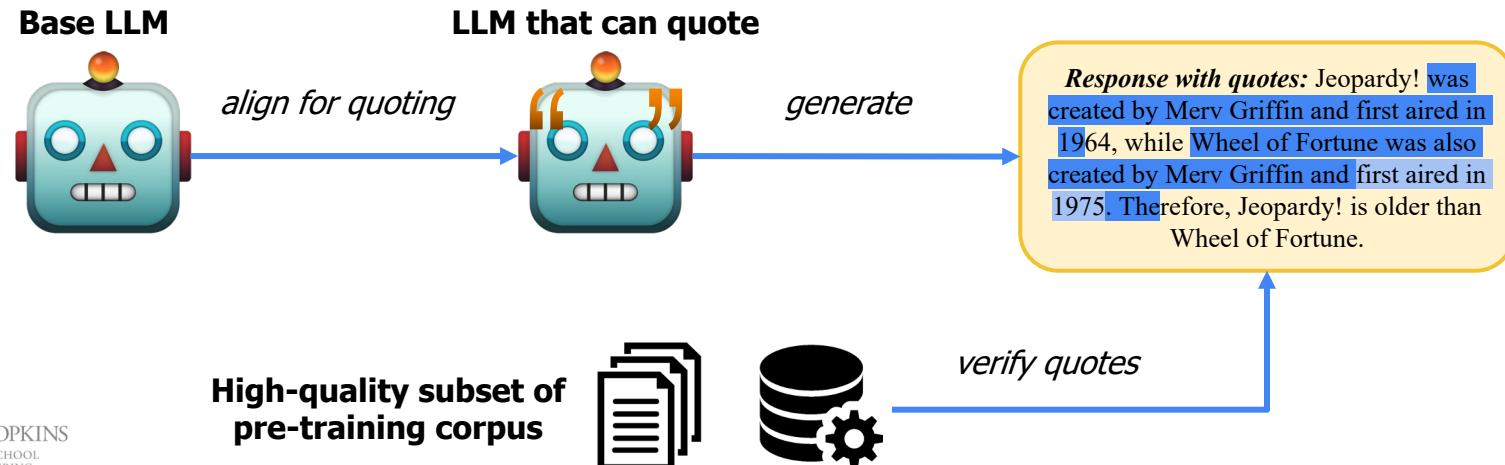
- [1]: nasa.gov (✖ citation does not support its associated statement)
[NASA's Webb Confirms Its First Exoplanet](#)
... Researchers confirmed an exoplanet, a planet that orbits another star, using NASA's James Webb Space Telescope for the first time. ...
- [2]: cnn.com (⚠ citation partially supports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

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



Background: measuring quoting at scale

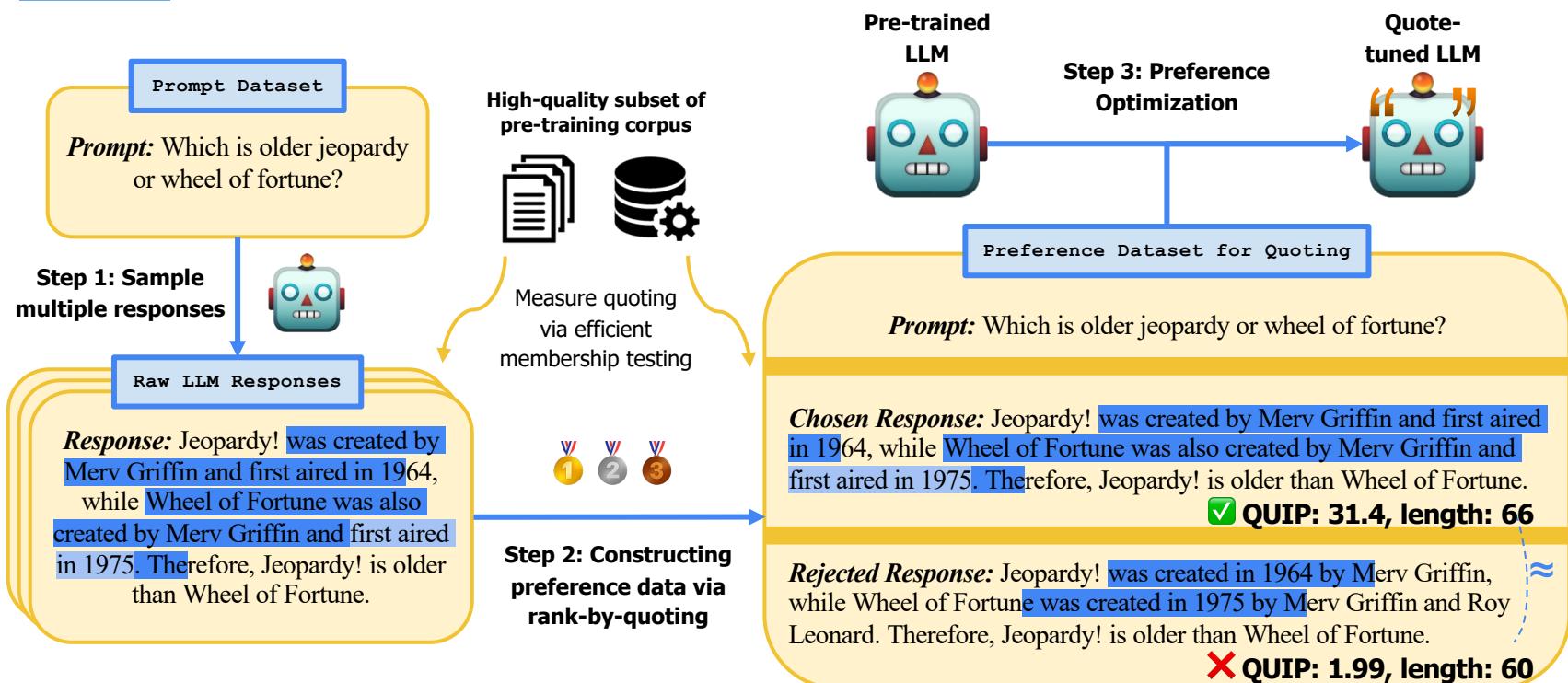
generated text

A large corpus

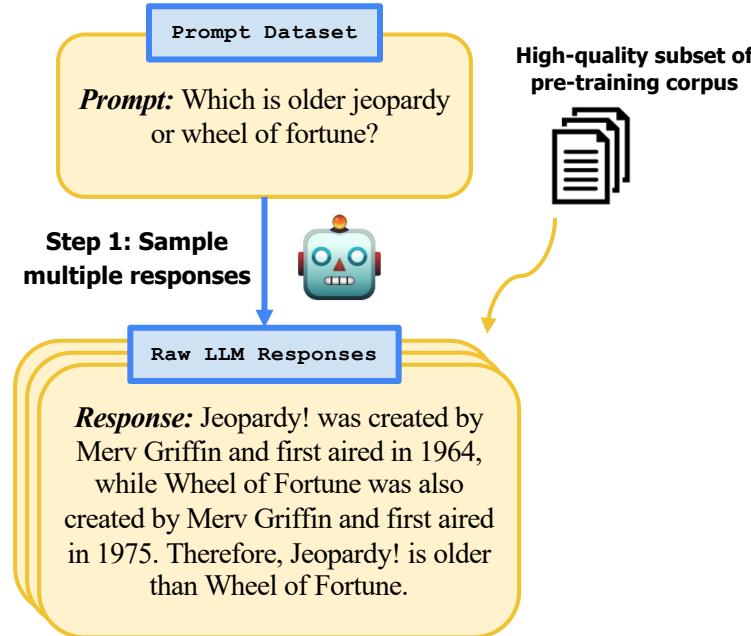
$$\text{QUIP}(Y; \text{WIKIPEDIA}) = \frac{\#\text{n-grams in } Y \text{ found in } C}{\#\text{n-grams in } Y}$$

- 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)

Our method: Quote-Tuning

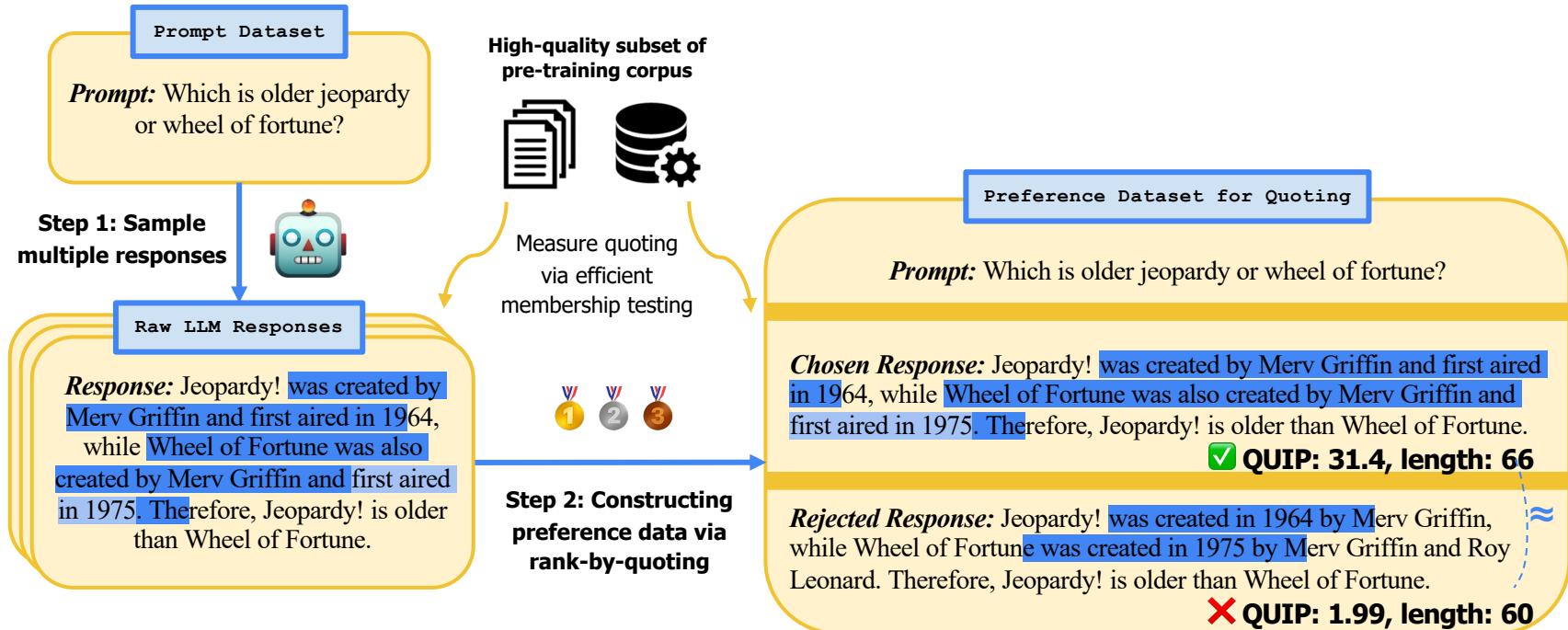


Step 1: Sample candidate responses



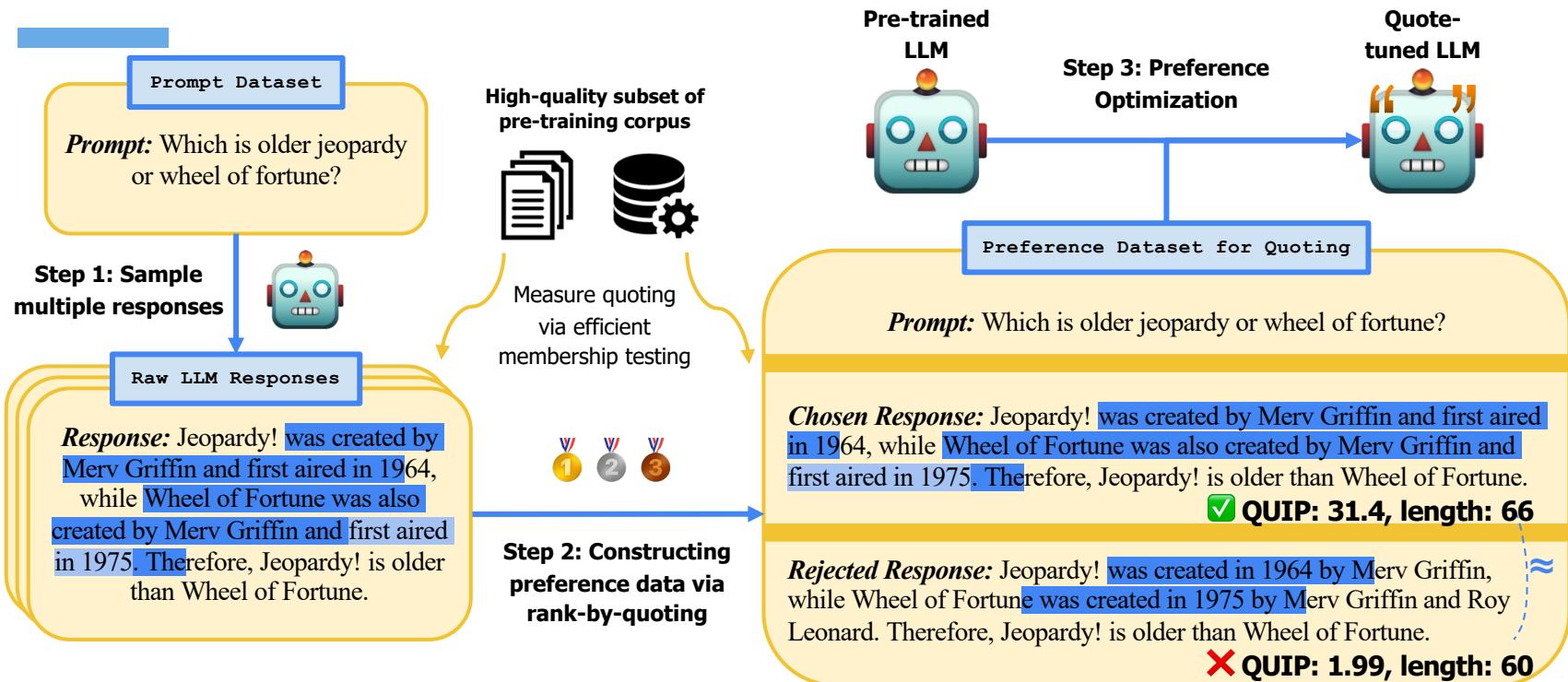
Step 1. Generate completions from an LLM (e.g. using QA pairs or text completions)

Step 2: Synthesize Dataset by Filtering



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!



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.

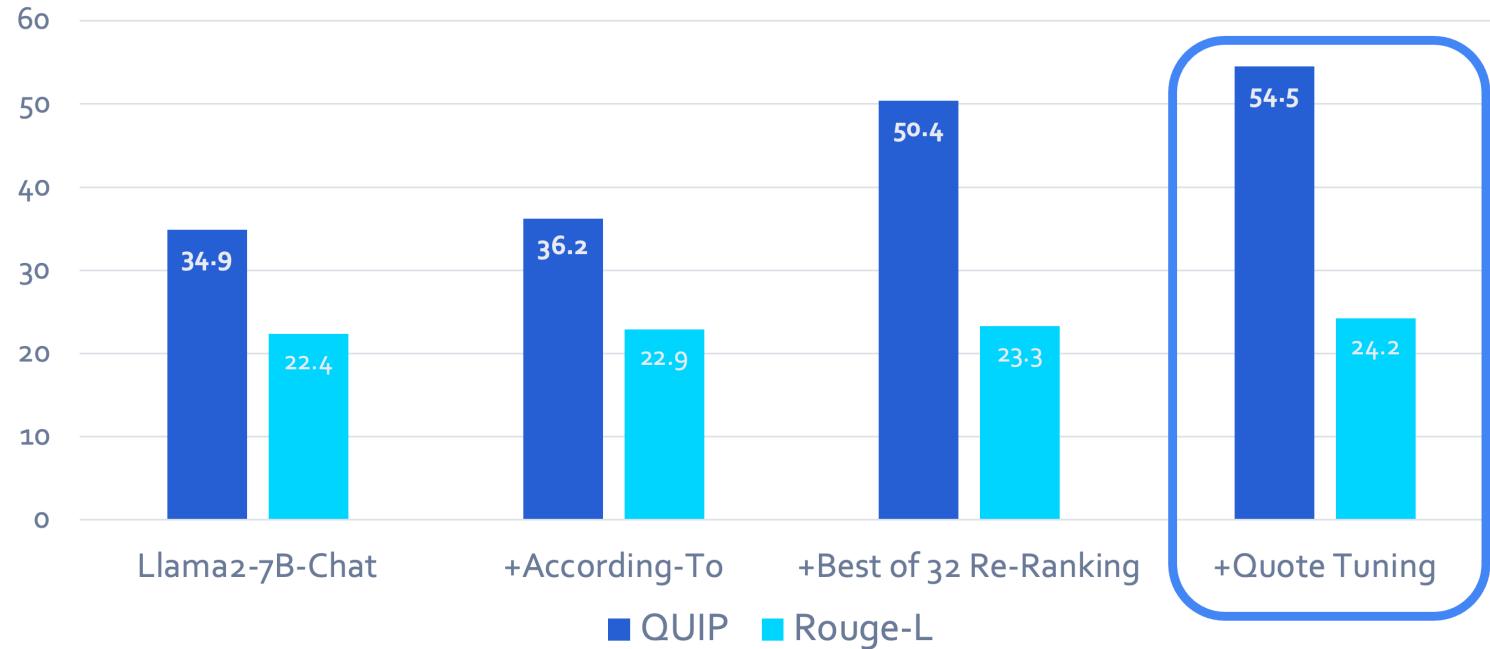
3.9

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

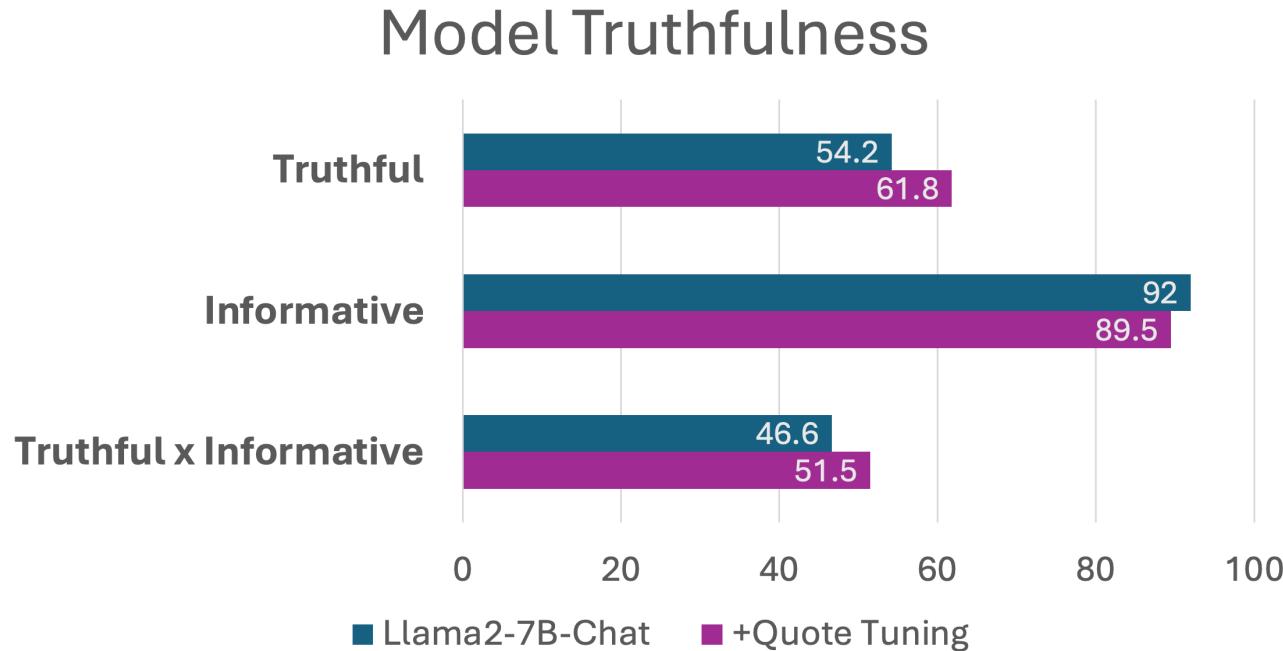
generation is densely quoted!

Same Quality, More Quoting!



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

Quote-Tuning improves truthfulness



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?
 - ...

Thank
you!



Questions?

<-link to preprint