AmeriSpeak

CELEBRATING 10 YEARS

LLMs Don’t Respond Like Humans

But, they can.

05.14.2025

Lilian Huang, Brandon Sepulvado, Joshua Lerner

Can LLMs help speed up iteration when designing open-ended survey questions?

- Open-ended question design is challenging
  - Time, cost, respondent burden (e.g., cognitive testing)
- **Can LLMs help reduce these pain points?**
  - Question testing
  - Synthetic respondents
- **Only if:**
  - There is high fidelity between what humans say and what LLMs say
  - The distributions of LLM-generated text aligns with the latent distributions of human-generated text

Methodological Approach: Setup

Survey Data

- AmeriSpeak Omnibus panel: ~1,000 responses

Models

- Base models: GPT-3.5 Turbo, GPT-4, GPT-4o, Claude 3 Opus, Llama 3
- Fine-tuned GPT-3.5 Turbo

AI (or artificial intelligence) is a topic that has dominated news coverage of science and technology in recent years. In your own words, what does AI mean to you?

Methodological Approach: Outcomes

Measures

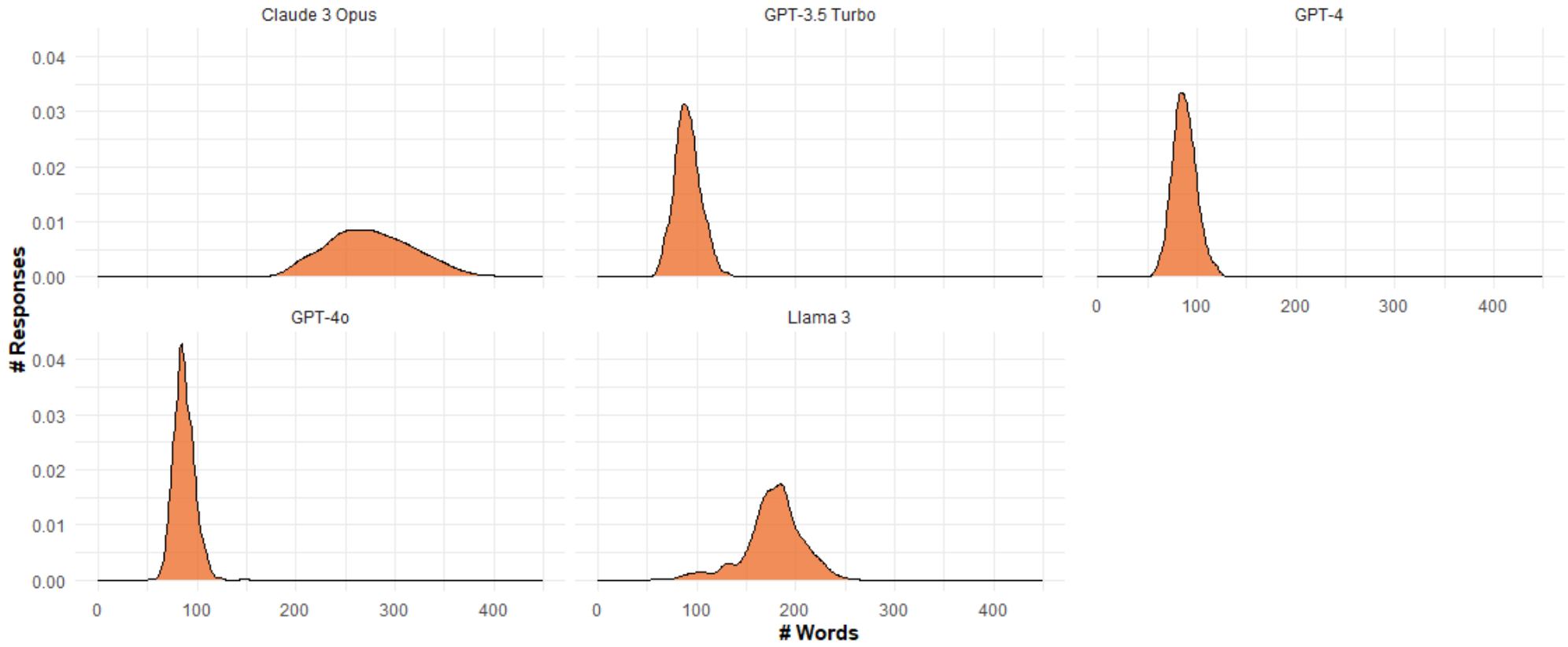
- Response length
  - Number of words
- Readability
  - Flesch-Kincaid Score
- Lexical diversity
  - Corrected type-to-token ratio (CTTR)

Off-the-Shelf Models

Do not respond like people

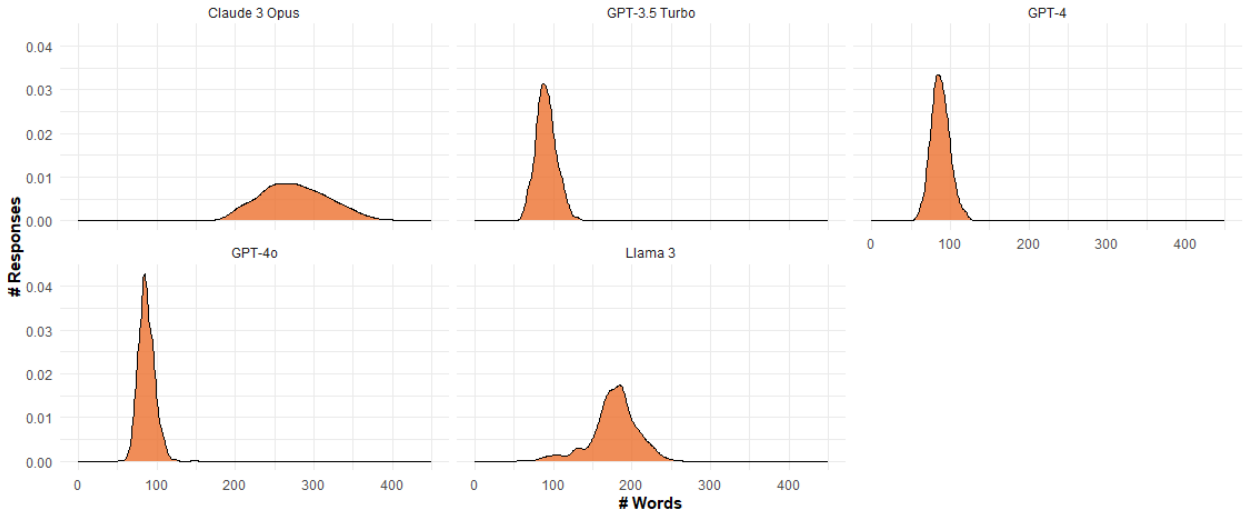
NORC Research Science

Response length—measured by the number of words—varies considerably between model types.

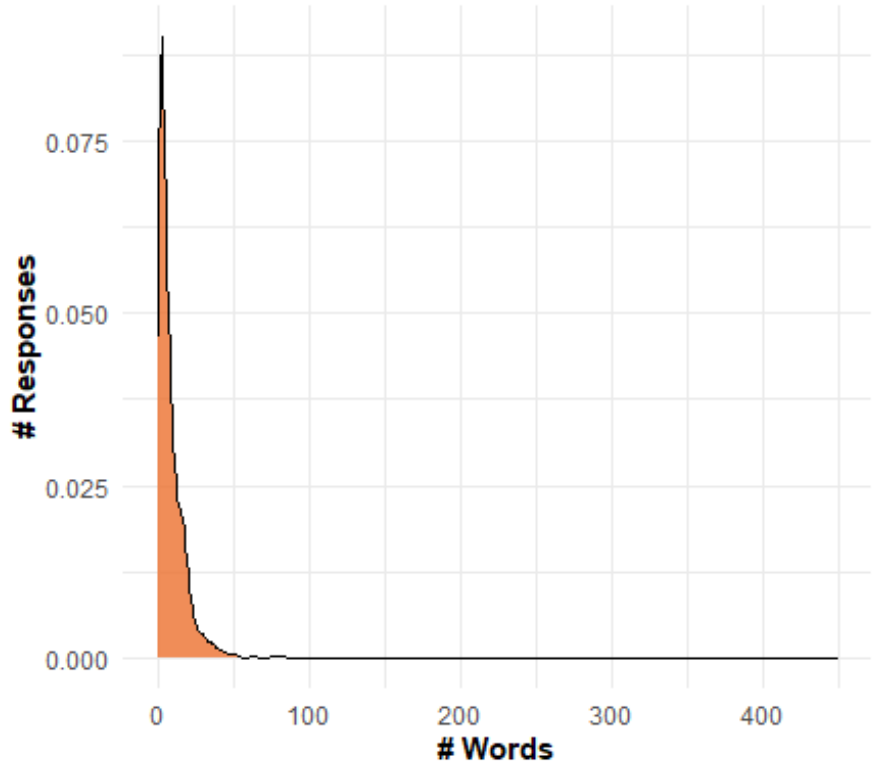


However, **none of these resemble** the response length of panelists.

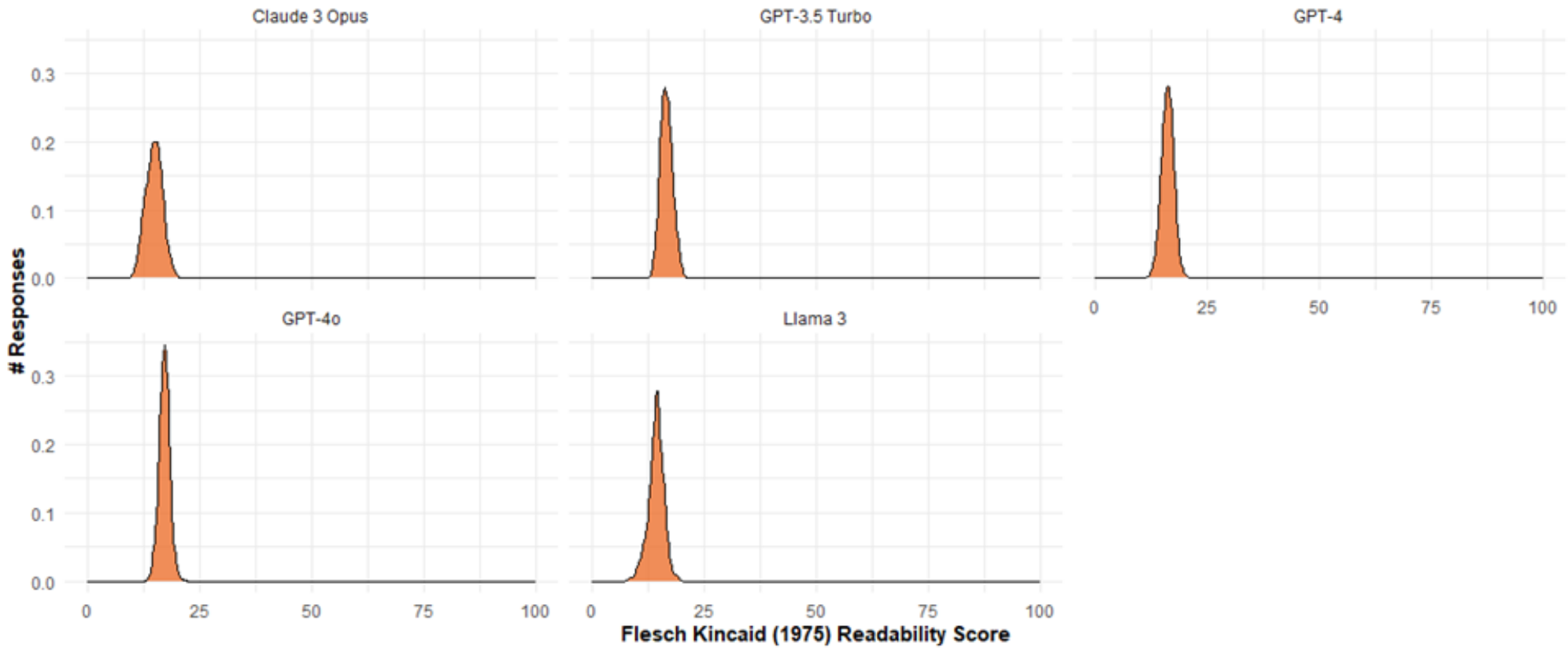
Off-the-Shelf LLMs



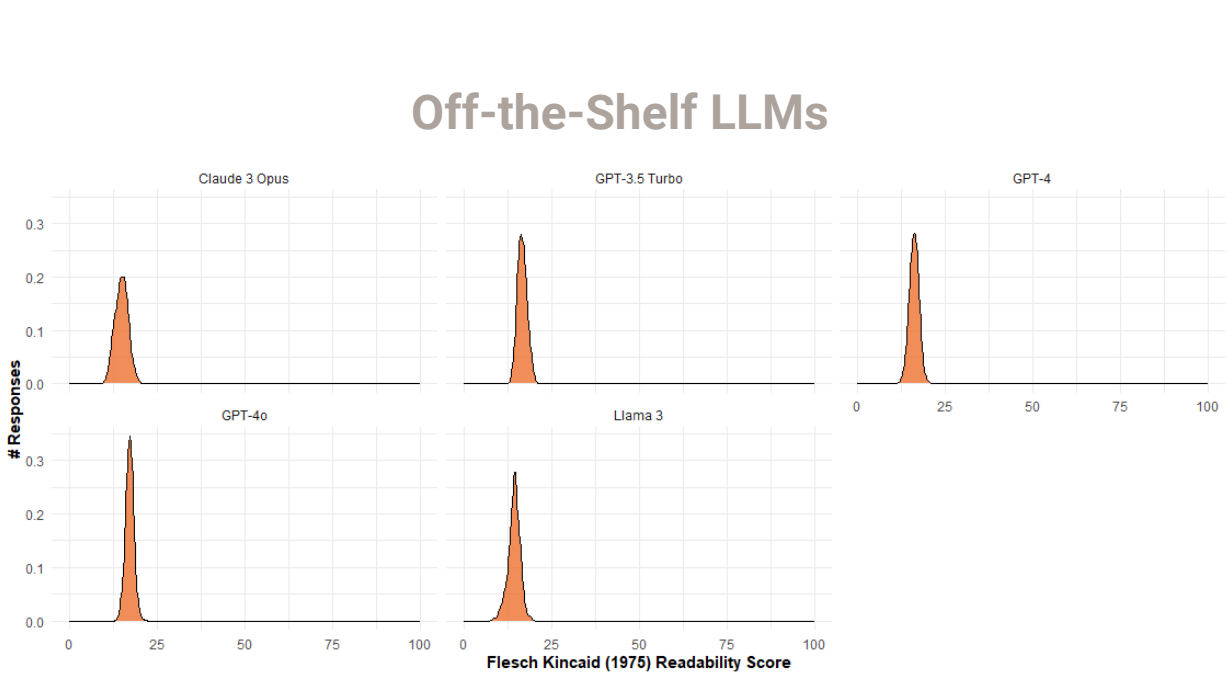
People

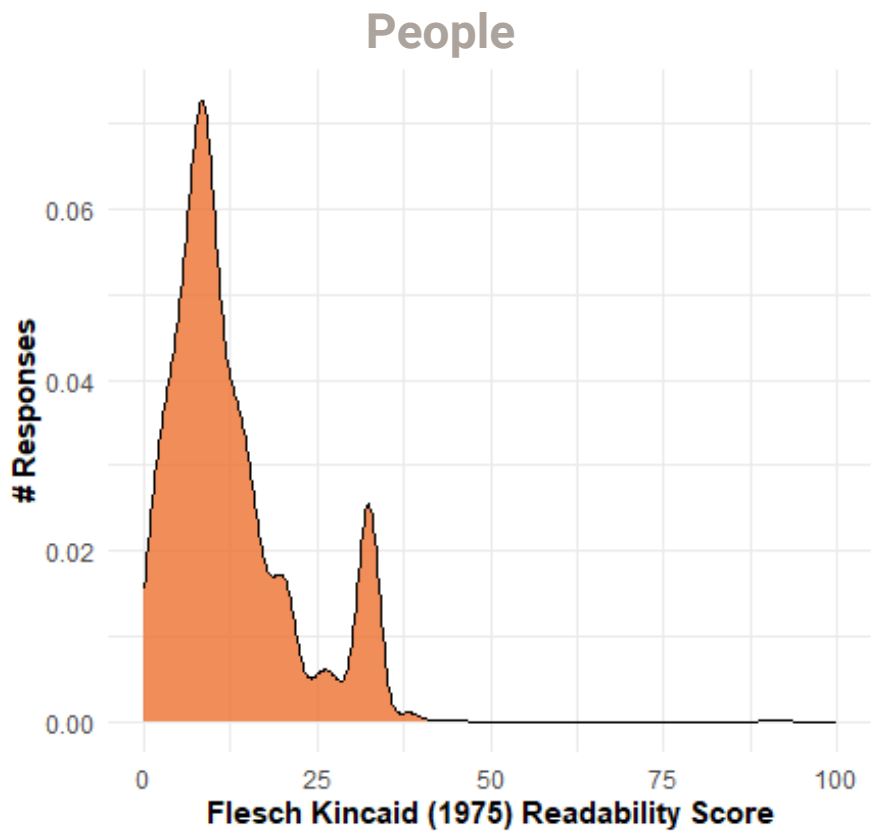


Readability scores are similar across LLMs.

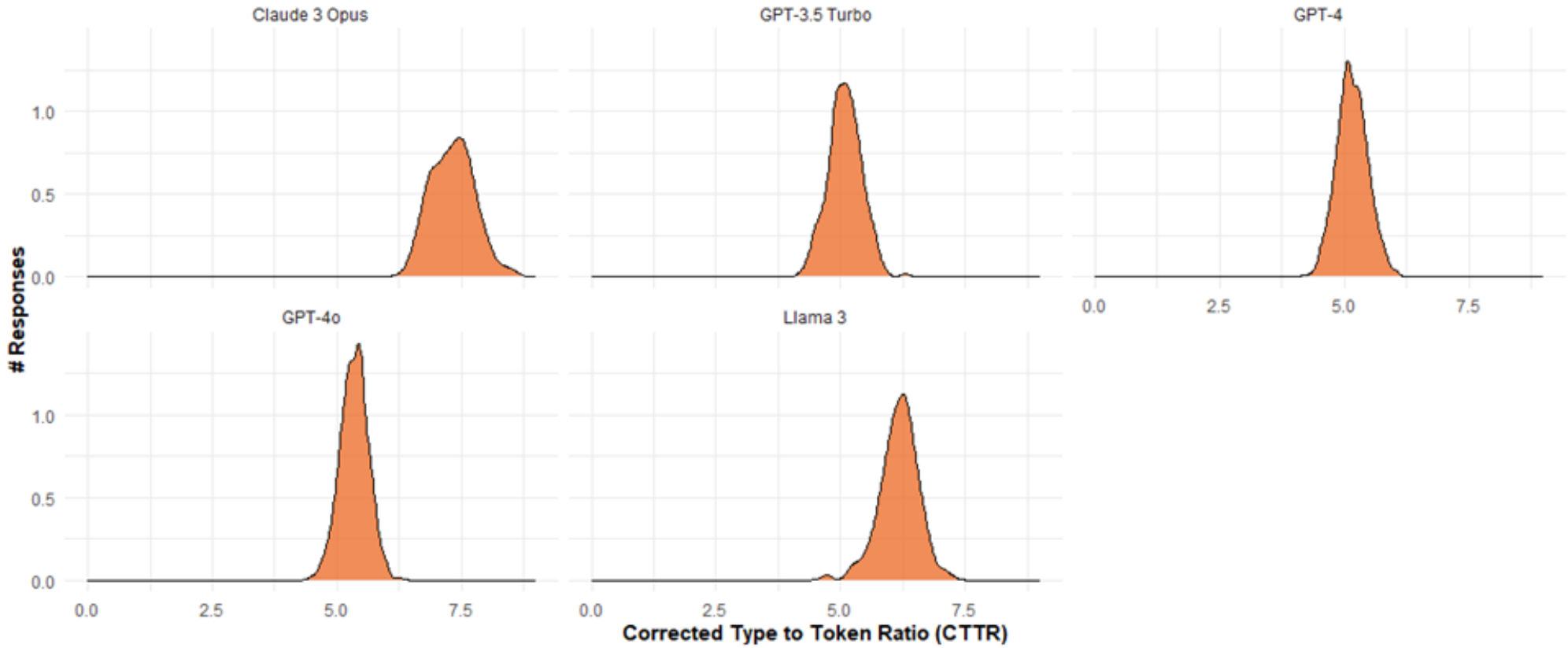


The readability of off-the-shelf LLM responses is **somewhat similar to the readability of panelists' responses**.

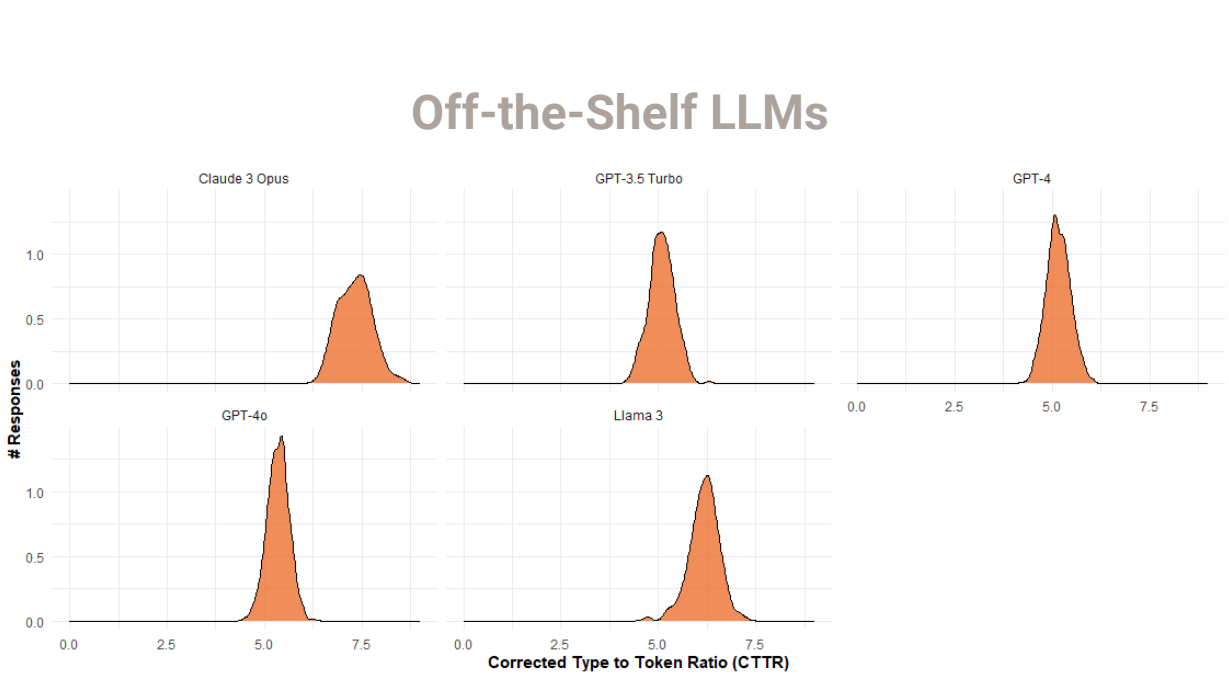


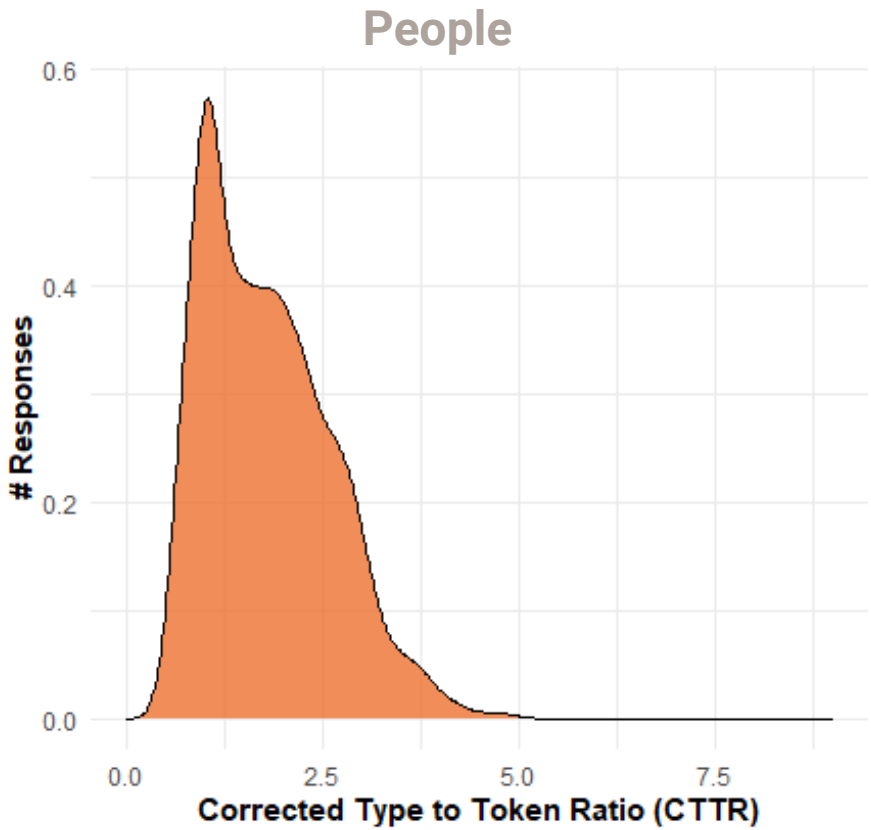


Lexical diversity is similar across LLMs.



However, the lexical diversity of panelist responses is **very different from that of LLM responses**.





Fine-Tuning

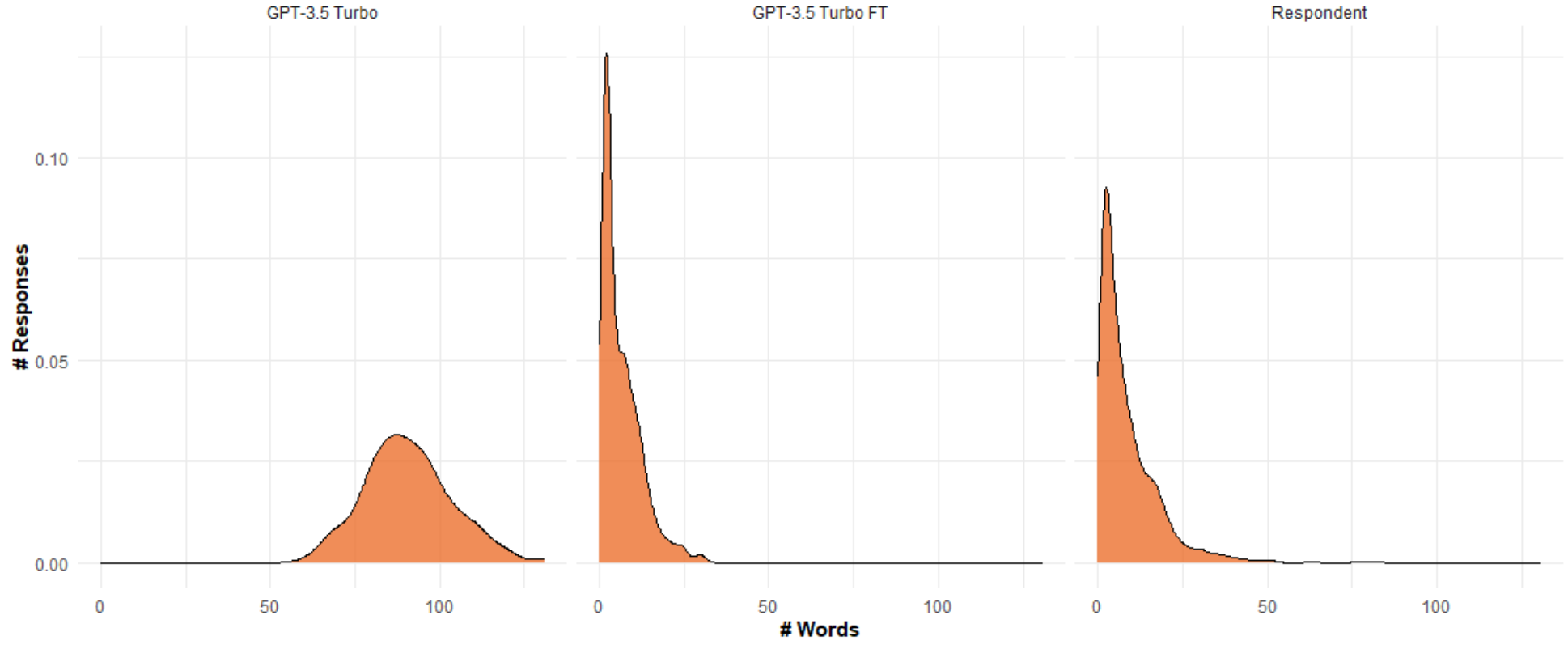
Seems to help

NORC Research Science

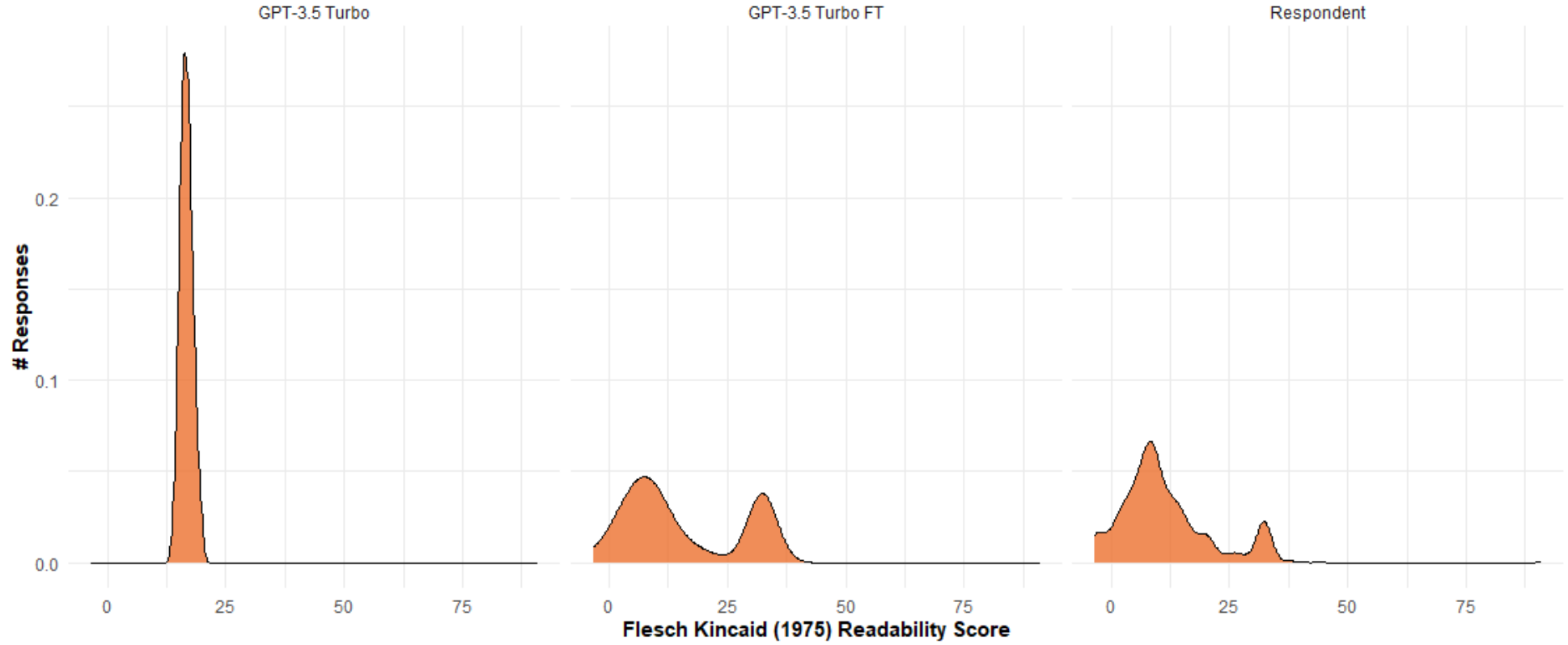
Does fine-tuning have an effect?

- Fine-tuning is the process of **further training** a pre-trained model on a **specialized dataset for a specific task**
- We fine-tuned **GPT-3.5 Turbo** with a set of **100 human responses**
- **Significant shift** in the synthetic responses produced by the fine-tuned model

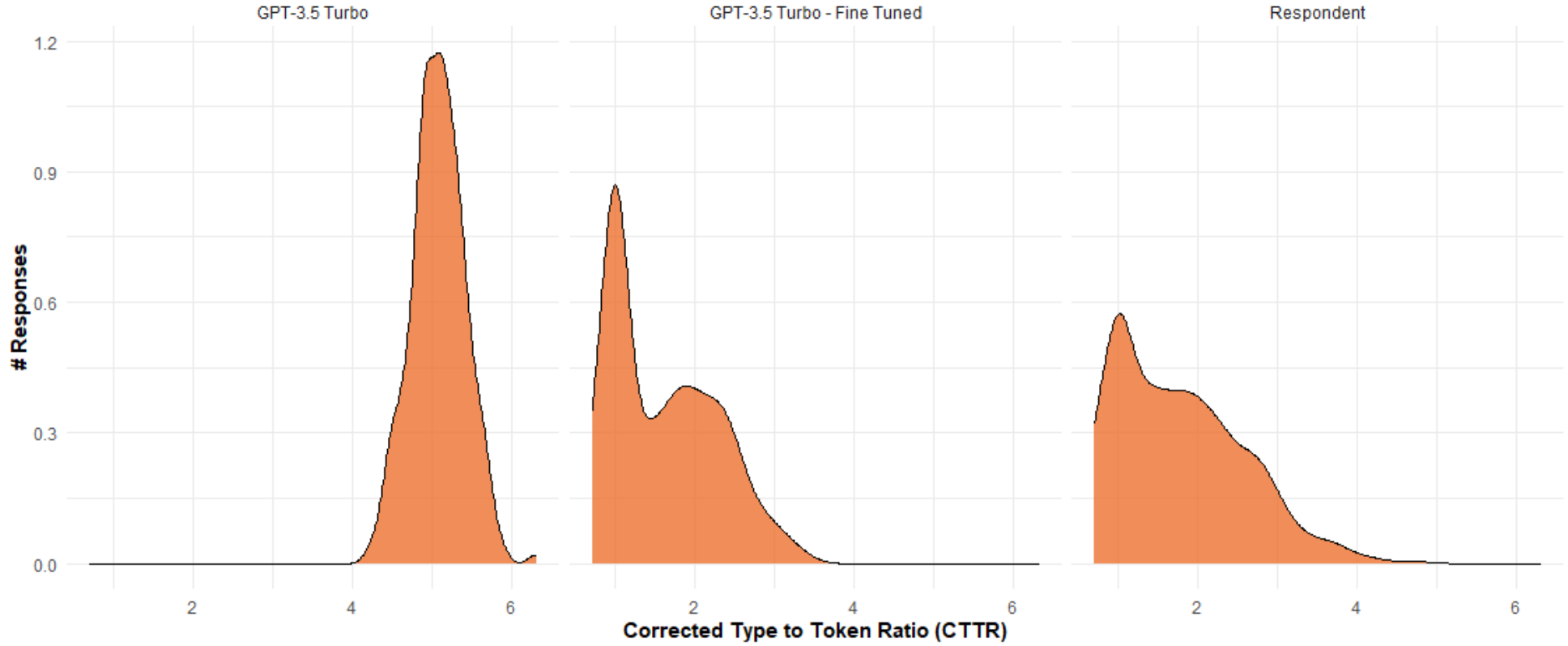
Fine-tuning produces responses whose **length better approximates that of panelists** compared to the base model.



Responses from the finetuned model **have readability similar to that of panelists**.



Fine-tuning produces responses whose **lexical diversity better approximates that of panelists** compared to the base model.



Soulless software created by the devil himself

Robotics - logic and science refining tasks.

Microsoft is a big Brother that watches everything you do

Computer program that can learn and adapt to situations.

an abstract, yet scientific technology that has demand intelligence to preformoundary tasks like driving, creating plans, making decisions,ect

It is the future and will help us in our everyday lives.

AI is the future!

The AI is a virtual assistant, it is like an assistant to the user, but through technology.

An underachieving computer system that can't even understand the language it is written in.

The ability for a machine to do things that are difficult for humans.

Computer systems designed to replace human work

Robots that can think and learn

hype

Off-the-Shelf

To me, AI or artificial intelligence refers to the development and use of machines or computer systems that can perform tasks and make decisions that typically require human intelligence. It involves creating algorithms and models that enable machines to learn from data, recognize patterns, and adapt their behavior accordingly. AI has the potential to revolutionize various industries by automating processes, improving efficiency, and enabling machines to perform complex tasks that were once limited to humans.

Fine-tuned

AI is the future!

Robotics - logic and science refining tasks.

hype

Computer program that can learn and adapt to situations.

Robots that can think and learn

It is the future and will help us in our everyday lives.

What does all this mean?

- Off-the shelf LLMs generate **readable responses that might initially seem plausible...**
- But the synthetic responses are **quite different** from real responses
- This hampers using LLMs for survey methodological purposes (e.g. question design or evaluation)
- **Fine-tuning makes a significant difference**
- Better at emulating human responses
- More similar to human responses in textual characteristics
- Reflects a variety of attitudes towards the question topic

What is the potential?

- How can we keep improving synthetic responses?
  - With **fine-tuning**
  - For **different question types** and domains
  - With more/better **prompt engineering**
  - With **other LLMs**

Thank you!

Lilian Huang
Statistician
huang-lilian@norc.org

Research You Can Trust™

NORC Research Science