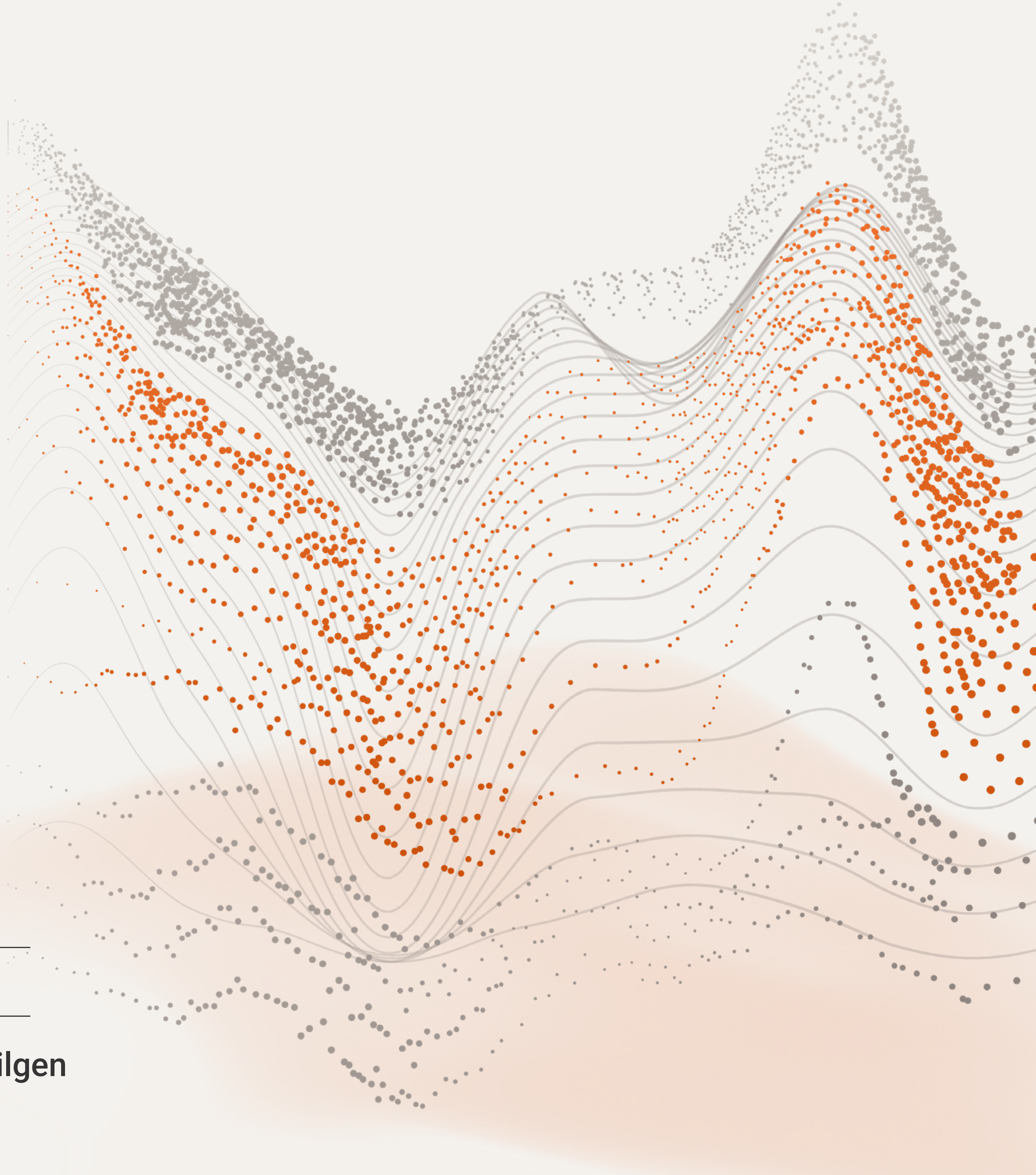
NORC Research Science

LLMs Do Not Respond like Survey Respondents

A Comparison of Synthetic and Human Responses

October 24, 2024

Lilian Huang, Brandon Sepulvado, Joshua Y. Lerner, Leah Christian, Ipek Bilgen



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

A question design throwback with the help of NLP

- Most open-ended survey questions tend to elicit short, relatively homogeneous responses, which pose challenges to NLP methods
- In the mid-20th century, surveys were largely open-ended, but researchers lacked methods to efficiently analyze text at-scale
- Methodological testing to better align open-ended question design and NLP analyses is challenging
  - Time, cost, respondent burden (e.g. cognitive testing)

Can LLMs help reduce these pain points?

Can we use LLMs to generate synthetic responses that approximate human responses for testing different approaches?

1. Do LLMs **consistently** generate the prompted behavior?
2. Is it possible to **meaningfully** guide the underlying data generating process for responses?
3. What are **effective** prompt engineering strategies?
4. Do different versions of LLMs differ in their **utility** for methodological research?

Methodological Approach: Setup

Survey Data

- AmeriSpeak Omnibus panel: 1,024 responses
- Question variants
  - (50%) Thinking about the problems facing the United States and the world today, which problems would you like the government to be working on in the next year?
  - (50%) Thinking about the problems facing the United States and the world today, in a few sentences which problems would you like the government to be working on in the next year?

Models

- GPT-3.5 Turbo and GPT-4
- No fine-tuning

Prompt engineering

- GPT-to-GPT comparisons
  - You are a respondent on a survey with an average response length of [10, 50, 100] words
- GPT-to-respondent comparisons
  - "In a few sentences" version of question

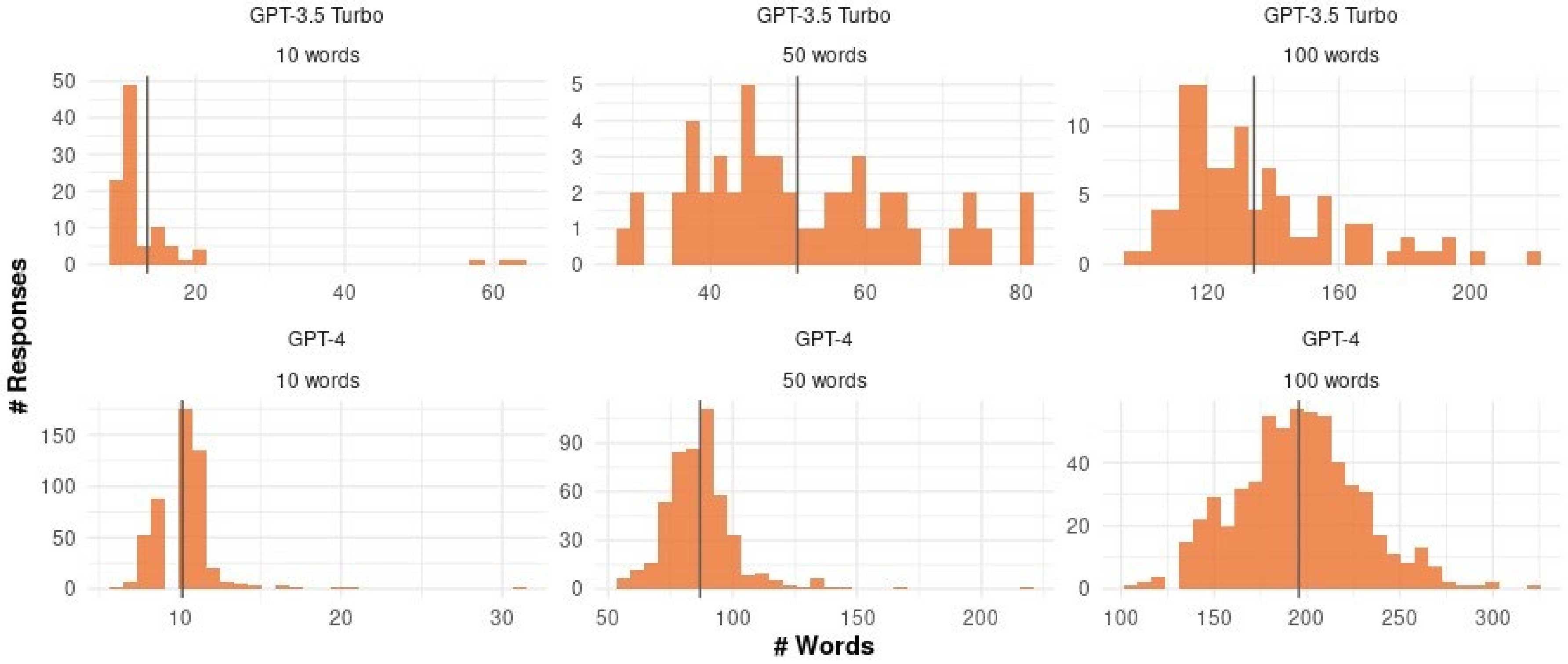
Methodological Approach: Validation

Measures

- Response length
- Readability (Flesch-Kincaid)
- Corrected type-to-token ratio (CTTR)

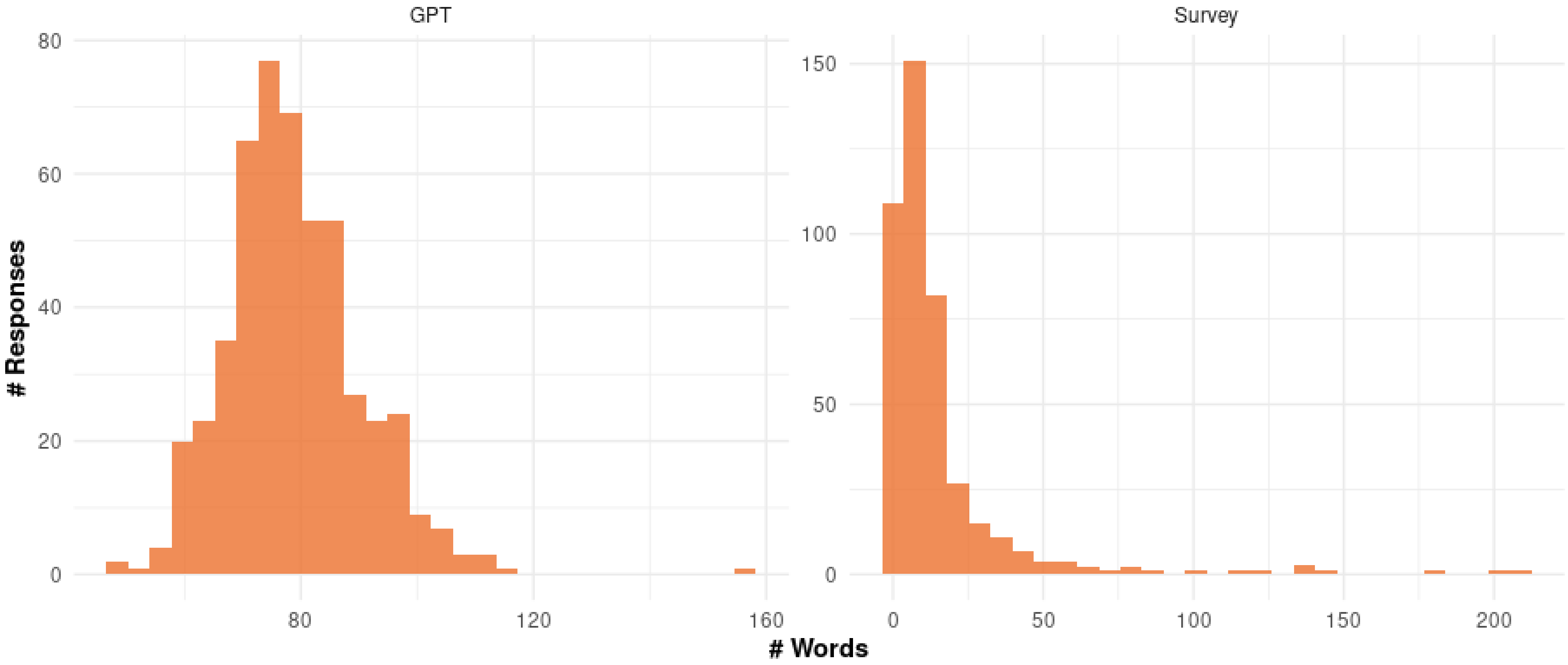
Findings

As the **prompted response length** increases, the mean response lengths in number of words tend to **overshoot the mark**.



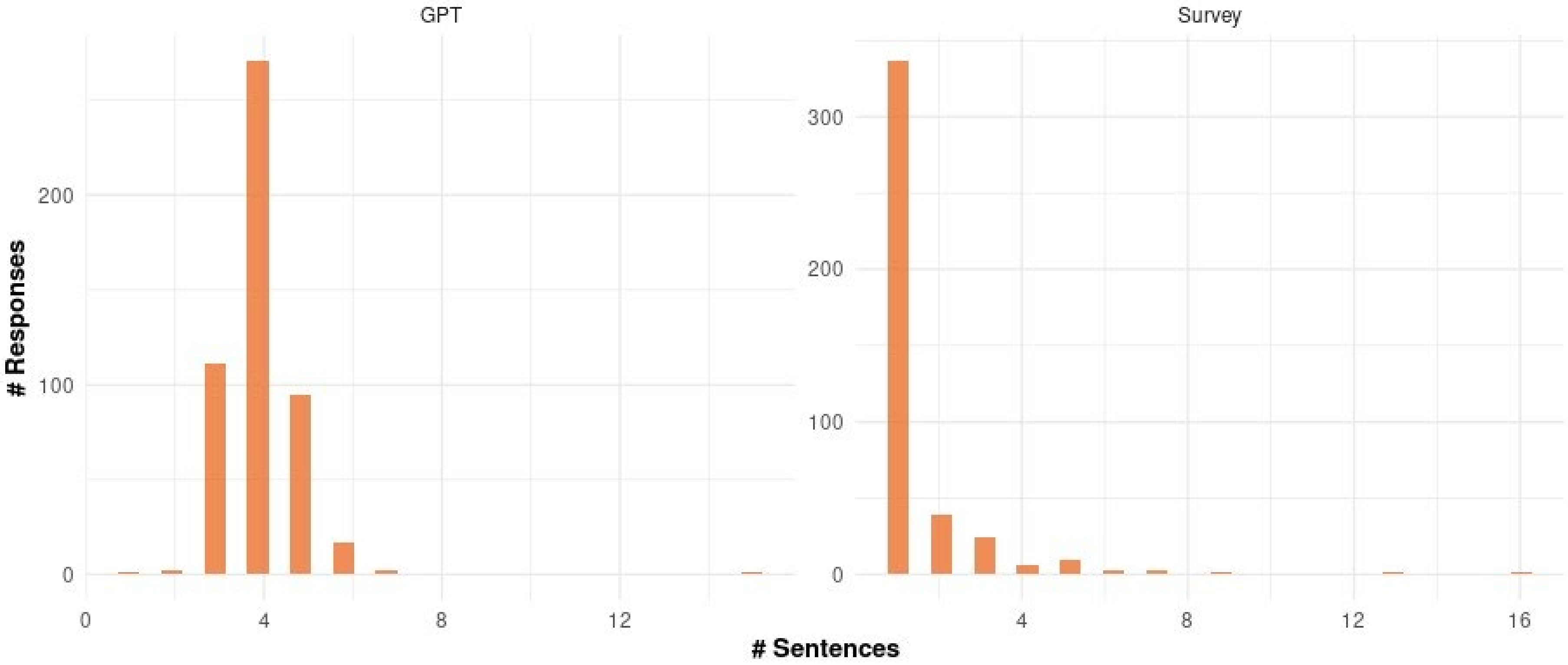
Note: Vertical lines represent means.

GPT **response lengths—in words**—differ considerably from those of survey responses, even with the same phrasing.



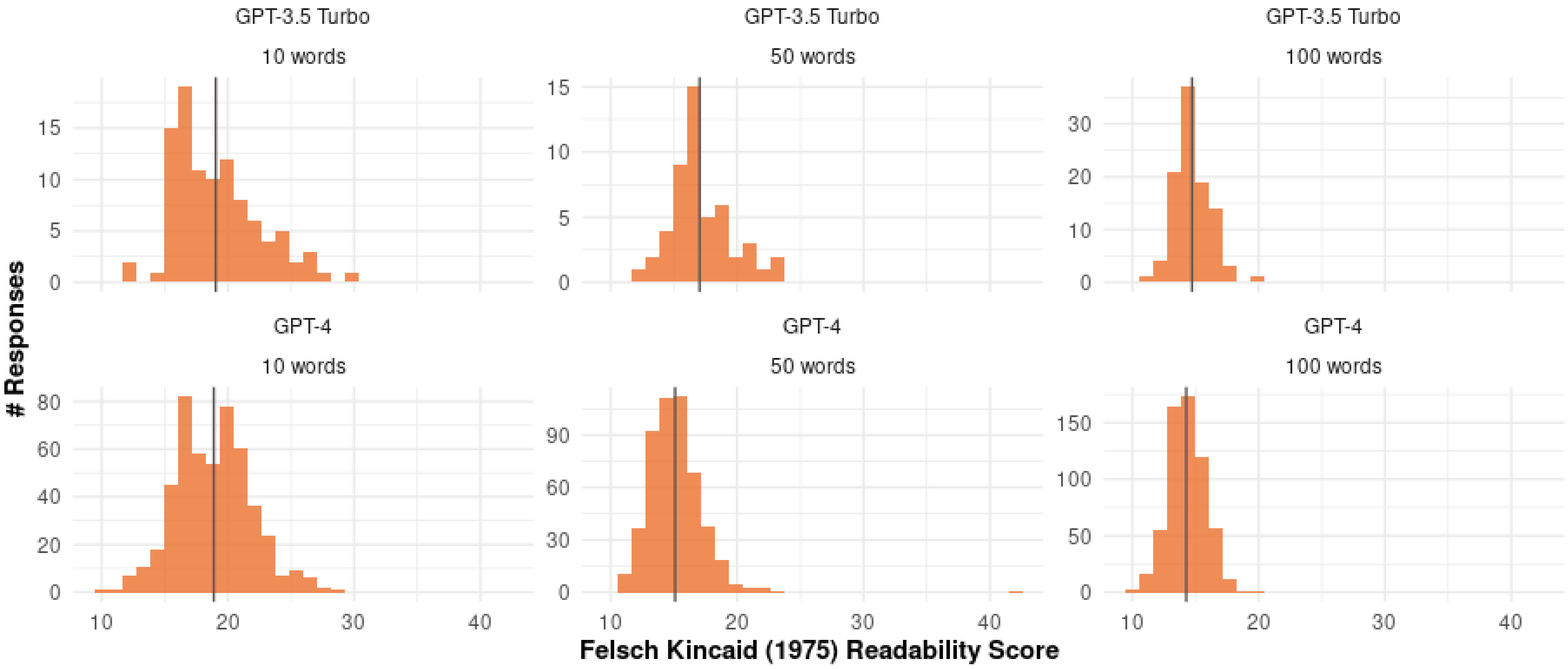
Note: Only 'in a few sentences' responses are included.

GPT **response lengths—in sentences**—also differ considerably from those of survey responses.



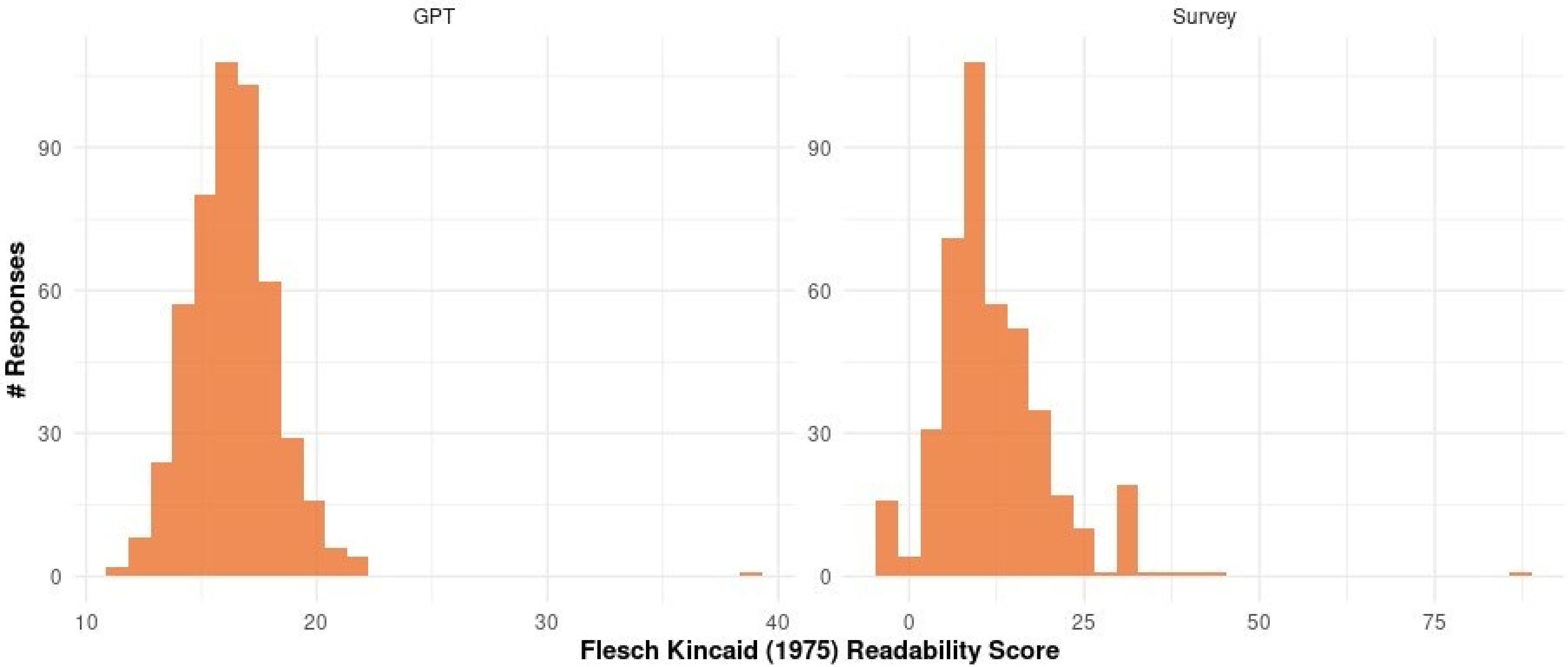
Note: Only 'in a few sentences' responses are included.

Both GPT models tend to generate similar **readability** distributions, and mean scores tend to be within a few grade levels of each other.



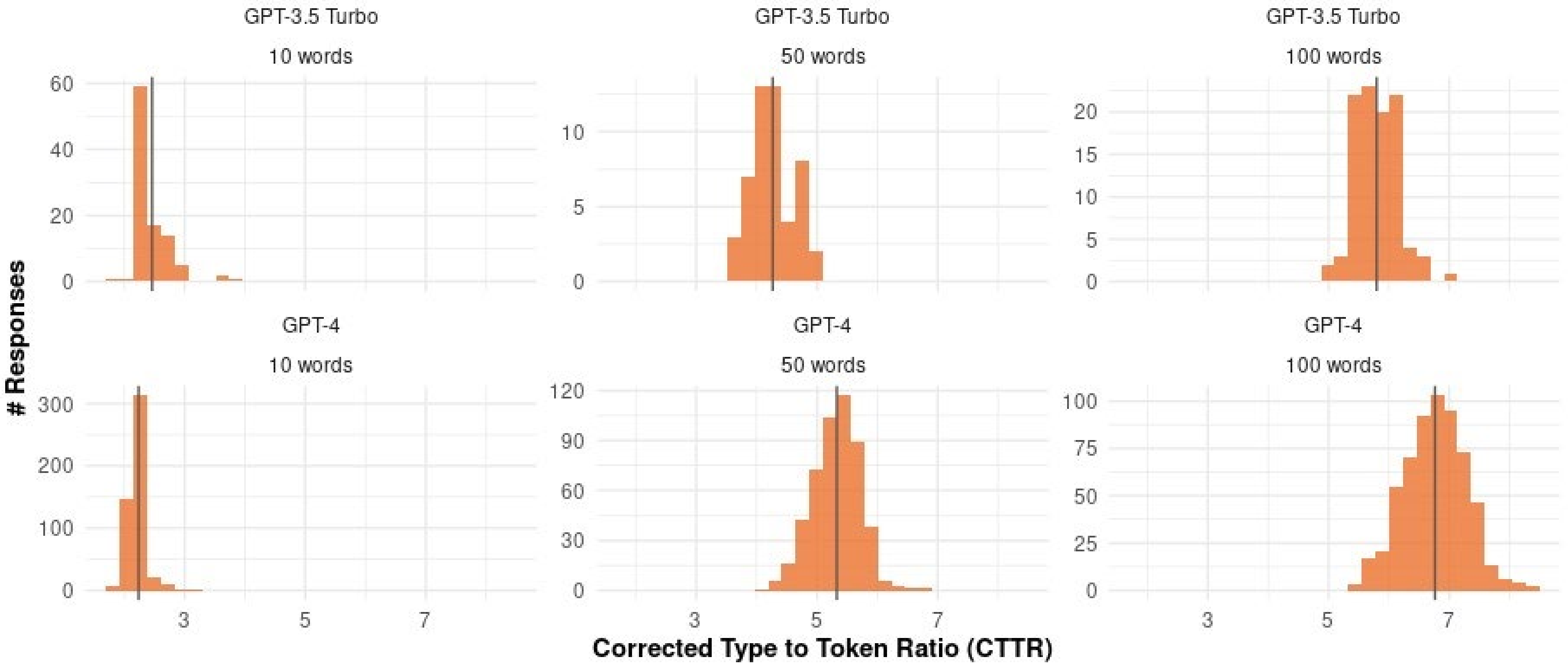
Note: Vertical lines represent means.

The mean readability of survey responses is lower than that from GPT but **varies much more** among respondents' answers.



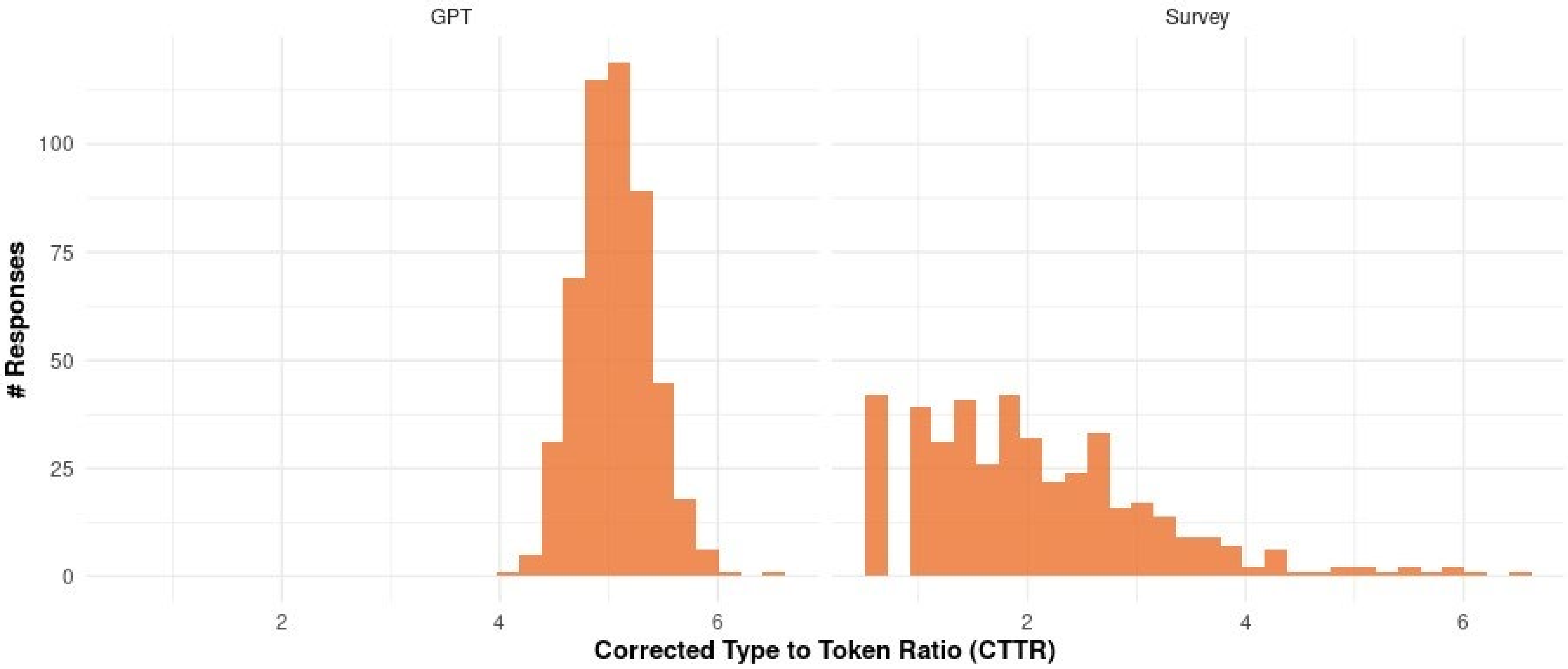
Note: Only 'in a few sentences' responses are included.

GPT appears to generate **relatively consistent CTTRs** between model versions, though with variation based upon length.



Note: Vertical lines represent means.

However, the GPT CTTR distribution **does not approximate** that from survey respondents.



Note: Only 'in a few sentences' responses are included.

What does all this mean?

- GPT generates **readable responses that might initially seem plausible...**
- But the synthetic responses are quite different from real responses.
- GPT models often—if not usually—**do not produce responses that strictly adhere to the prompt**...
- Though the system message tends to have more impact than the query.

What is the potential?

- Might LLMs still be useful for survey methodology?
  - For other tasks than generating text, e.g. Kim and Lee (2023)
  - For different question types and domains
  - With more/better prompt engineering
  - With other LLMs

Thank you!

Lilian Huang
Statistician II
Huang-Lilian@norc.org

Research You Can Trust™

NORC Research Science