Factuality

CS 685, Fall 2025

Advanced Natural Language Processing

https://people.cs.umass.edu/~brenocon/cs685_f25/

Brendan O'Connor

College of Information and Computer Sciences

University of Massachusetts Amherst

Quiz: which ones have errors?

Input: Tell me a bio of Bridget Moynahan.

Bridget Moynahan is a actress who were born in April 28, 1971 at Binghamton, New York. She become best known with her role as ...

Bridget Moynahan was born in Binghamton, New York. (...) Born in Springfield, Massachusetts, ...

Bridget Christie is an English stand-up comedian, actress and writer.

Bridget Moynahan is an American actress, model and producer, best known for her roles in Grey's Anatomy, I, Robot and Blue Bloods.

[slide: Sewon Min]

Quiz: which ones have errors?

Input: Tell me a bio of Bridget Moynahan.

Bridget Moynahan is ~~a~~ an actress who ~~were~~ was born ~~in~~ on April 28, 1971 ~~at~~ in Binghamton, New York. She ~~become~~ has been best known ~~as~~ for her role as ...

Bridget Moynahan was born in Binghamton, New York. (...) Born in Springfield, Massachusetts, ...
(Inconsistent!)

Bridget Christie is an English stand-up comedian, actress and writer.
(This is not about Bridget Moynahan!)

Bridget Moynahan is an Ameri
actress, mode
kn utomy, I,
d Blue Bloods.

Hardest to detect errors!

[slide: Sewon Min]

- Many ways to consider the quality of LM output

- Should we ever expect LMs to be "factual"?

- *Factuality* of LM generation is relative to *grounding documents/knowledge*

- What does it mean for language to be *supported* by a grounding text?
- How to evaluate LM factuality?

Natural language inference

- Idea: the meaning of a phrase or sentence is a set

- Meaning relationships can be described as set relationships
  - [[I saw a red cat]] ENTAILS [[I saw a cat]]
  - [[red cat]] ENTAILS [[cat]]

- "Natural Language Inference" framework of 3 meaning relationships between a pair of sentences/phrases
  - Entailment
  - Contradiction
  - Neutral

- Entailment and contradiction relationships are unidirectional

- Context (entail | contradict | neutral) Hypothesis?

- C: An Irishman won the Nobel prize for literature.
  H: An Irishman won a Nobel prize.

- C: Nucor has pioneered the first mini-mill.
  H: Nucor has pioneered a giant mini-mill in which steel is poured into continuous casting machines

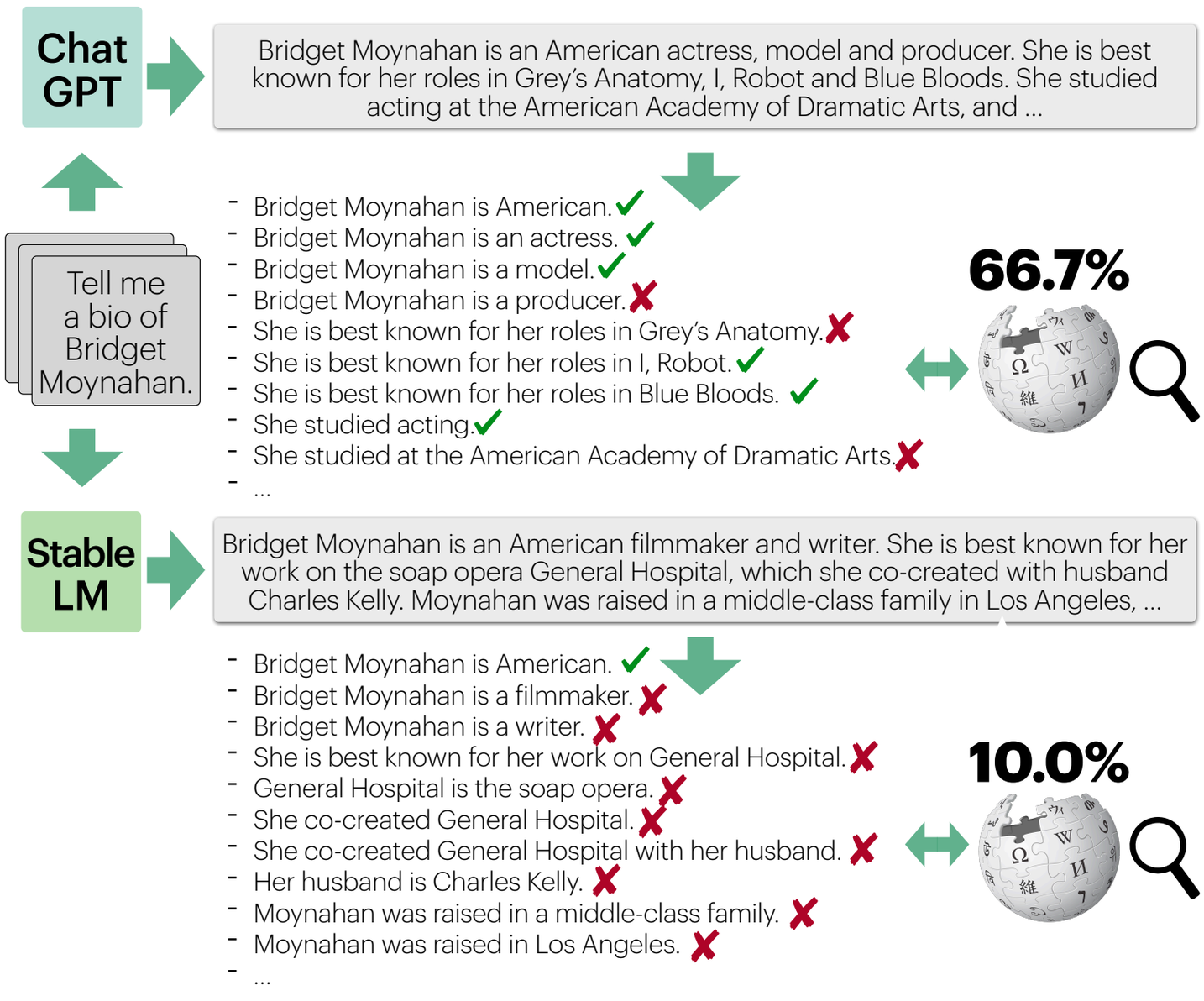


Figure 1: An overview of FACTSCORE, a fraction of *atomic facts* (pieces of information) supported by a given knowledge source. FACTSCORE allows a more fine-grained evaluation of factual precision, e.g., in the figure, the top model gets a score of 66.7% and the bottom model gets 10.0%, whereas prior work would assign 0.0 to both. FACTSCORE can either be based on human evaluation, or be automated, which allows evaluation of a large set of LMs with no human efforts.

- FACTSCORE: Fine-grained Atomic Evaluation of Factual Precision in Long Form Text Generation (Min et al., EMNLP 2023)

- Main assumptions
  - 1. Atomic fact is a unit
  - 2. Factual precision is relative to a given knowledge source

- Experiments just on evaluation
  - 1. Human-assisted evalution
  - 2. Fully automatic eval.

- Other work: use automatic factuality evaluation for RLHF and other fine-tuning

Human Evaluation of FActScore

Step 1: Preparing Prompts

Step 2: Obtaining Generations from the LMeval

Step 3: Atomic Fact Break-down

Step 4: Annotating Precision Labels

Bridget Moynahan is an American.
Bridget Moynahan is an actress.
Bridget Moynahan is a model.
Bridget Moynahan is a producer.
...
She studied acting at the American Academy of Dramatic Arts.
She began her career in the late 1990s.

upwork

Irrelevant

Supported

Not-supported

[slide: Sewon Min]

Fact decomposition

Segment the following sentence into individual facts:

Sentence: Other title changes included Lord Steven Regal and The Nasty Boys winning the World Television Championship and the World Tag Team Championship respectively.
Facts:
- Lord Steven Regal won the World Television Championship.
- The Nasty Boys won the World Tag Team Championship.

Sentence: The parkway was opened in 2001 after just under a year of construction and almost two decades of community requests.
Facts:
- The parkway was opened in 2001.
- The parkway was opened after just under a year of construction.
- The parkway was opened after two decades of community requests.

Sentence: Touring began in Europe in April-June with guitarist Paul Gilbert as the opening act, followed by Australia and New Zealand in July, Mexico and South America in late July-August, and concluding in North America in October-November.
Facts:
- Touring began in Europe in April-June.
- The opening act of the tour was guitarist Paul Gilbert.
- The tour was in Australia and New Zealand in July.
- The tour was in Mexico and South America in late July-August. - The tour was concluded in North America in October-November.

Sentence: In March 2018, the company partnered With Amazon Web Services (AWS) to offer Al-enabled conversational solutions to customers in India.
Facts:
- The company partnered with Amazon Web Services (AWS) in March 2018.
- The two companies partnered to offer Al-enabled conversational solutions to customers in India.

Sentence: The most significant of these is in Germany, which now has a Yazidi community of more than 200,000 living primarily in Hannover, Bielefeld, Celle, Bremen, Bad Oeynhausen, Pforzheim and Oldenburg.
Facts:
- The most significant of these is in Germany.
- Germany now has a Yazidi community of more than 200,000.
- Yazidi community in Germany lives primarily in Hannover, Bielefeld, Celle, Bremen, Bad Oeynhausen, Pforzheim and Oldenburg.

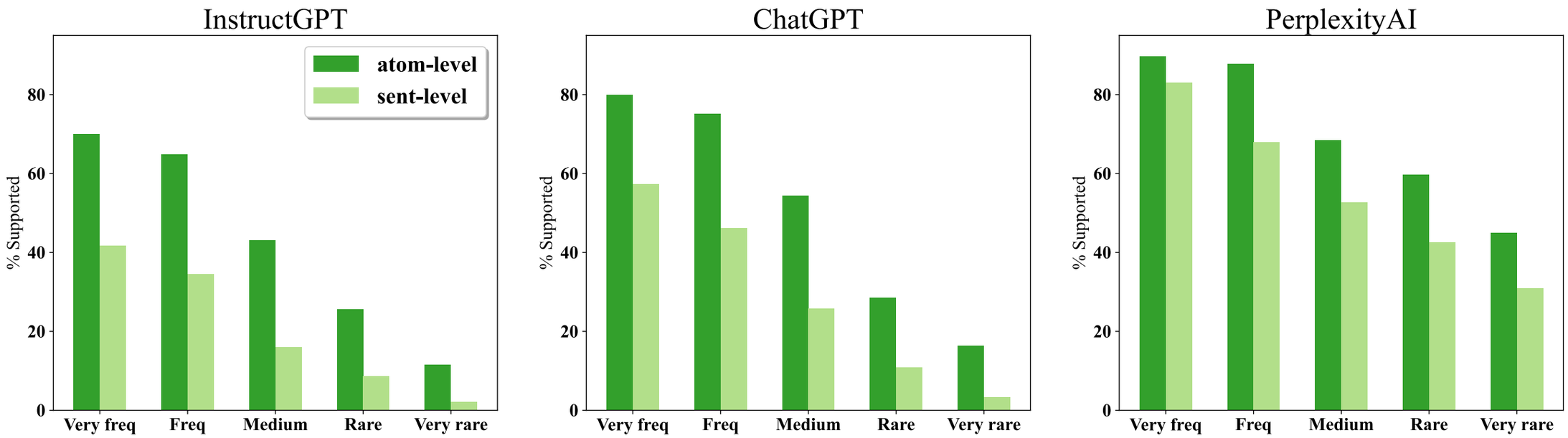
Sentence: A previous six-time winner of the Nations' Cup, Sebastian Vettel became Champion of Champions for the first time, defeating Tom Kristensen, who made the final for the fourth time, 2-0.
Facts:
- Sebastian Vettel is a previous six-time winner of the Nations' Cup.
- Sebastian Vettel became Champion of Champions for the first time, defeating Tom Kristensen, 2-0. - Tom Kristensen made the final for the fourth time.

Sentence: [SENTENCE]
Facts:

- From "MiniCheck" (Tang et al., EMNLP 2024)

- (Results circa 2023; scores much higher now...)

Results



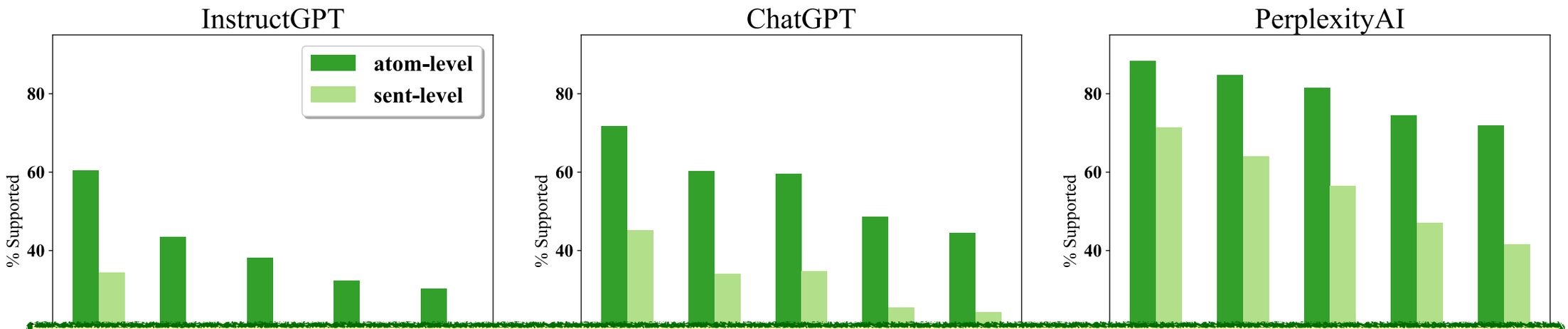
More errors as the rarity of entities increases

In agreement with QA results in Kandpal et al. (2022) and Mallen et al. (2022) in models w/o search
Different from QA, same trend in models w/ search

[slide: Sewon Min]

- (Results circa 2023; scores much higher now...)

Results



Systems that generate longer outputs have more rooms for errors

More errors at a later part of generation

Reason #1: Error propagation

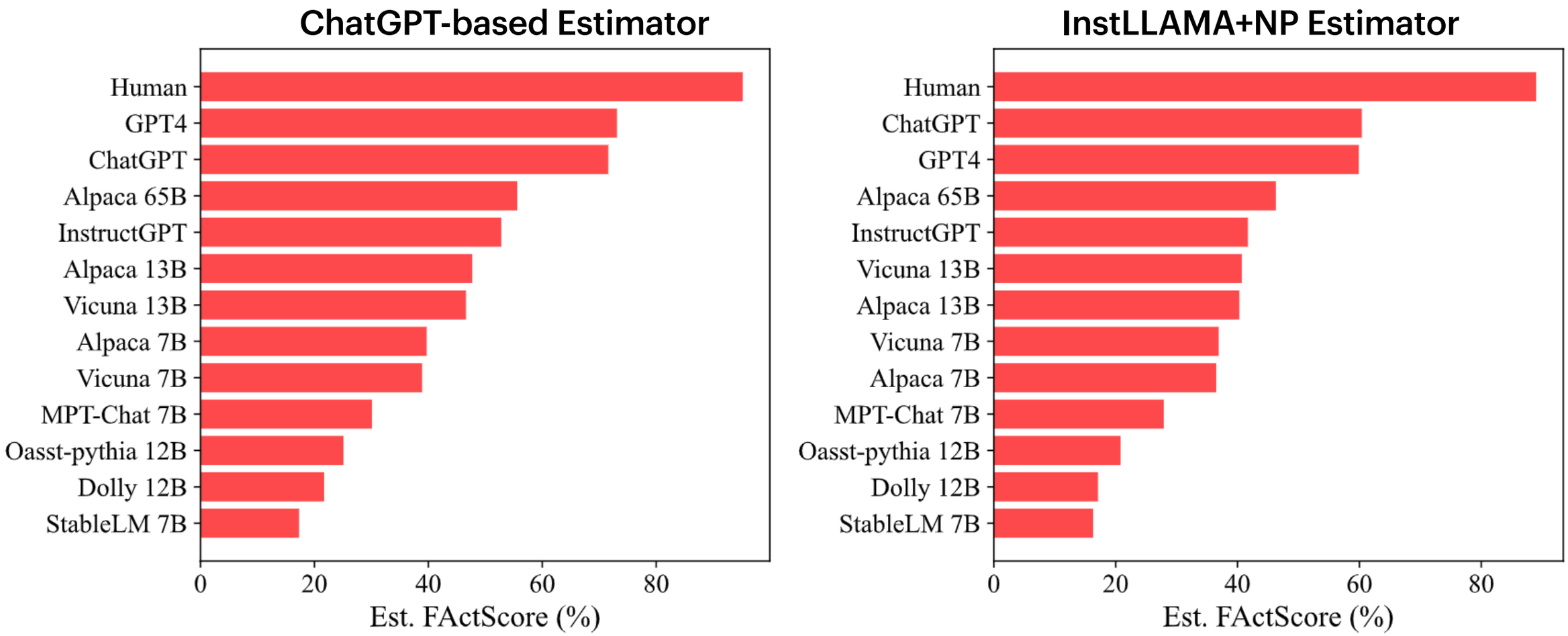
Reason #2: Earlier facts are more representative than the later ones

26

[slide: Sewon Min]

- (Results circa 2023; scores much higher now...)

Results



Pearson's r = 0.99

39

[slide: Sewon Min]

- Retrieval of relevant documents
  - Retrieval-Augmented Generation

- What documents to use? How to select?

- Incorporating factuality scoring as RLHF reward