FActScore: Fine-grained Atomic Evaluation of Factual Precision in Long Form Text Generation

Sewon Min*, Kalpesh Krishna*,
Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Wei Koh,
Mohit Iyyer, Luke Zettlemoyer, Hannaneh Hajishirzi

* = core contributors

Long-form Text Generation

Input: Tell me a bio of Bridget Moynahan.

Long-form Text Generation

Input: Tell me a bio of Bridget Moynahan.

Output:

Bridget Moynahan is an American actress, model and producer. She is best known for her roles in Grey's Anatomy, I, Robot and Blue Bloods. She has had a successful career in both film and television, and has been nominated for many awards, including a Golden Globe. She studied acting at the American Academy of Dramatic Arts, and ...

Long-form Text Generation

Input: Tell me a bio of Bridget Moynahan.

Output:

Bridget Moynahan is an American actress, model and producer. She is best known for her roles in Grey's Anatomy, I, Robot and Blue Bloods. She has had a successful career in both film and television, and has been nominated for many awards, including a Golden Globe. She studied acting at the American Academy of Dramatic Arts, and ...

Long-form Text Generation

Fluency

Long-form Text Generation

Fluency

Relevance

Long-form Text Generation

Fluency

Relevance

Coherence

Long-form Text Generation

Fluency

Relevance

Coherence

Factuality

Long-form Text Generation

Fluency

Relevance

Coherence

Factuality

Long-form Text Generation

Fluency

Relevance

Coherence

Factuality

The current SoTA LMs suffer most on this.

Long-form Text Generation

Fluency

Relevance

Coherence

Factuality

The current SoTA LMs suffer most on this.
Most difficult & expensive to evaluate.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Hardest to detect errors!

Why is evaluation challenging?

Input: Tell me a bio of Bridget Moynahan.

Bridget Moynahan is an American actress, model and producer. She is best known for her roles in Grey's Anatomy, The Recruit, I, Robot and Blue Bloods. She studied acting at the American Academy of Dramatic Arts, and ...

Why is evaluation challenging?

Input: Tell me a bio of Bridget Moynahan.

This is true

Bridget Moynahan is an American actress, model and producer. She is best known for her roles in Grey's Anatomy, The Recruit, I, Robot and Blue Bloods. She studied acting at the American Academy of Dramatic Arts, and ...

Why is evaluation challenging?

Input: Tell me a bio of Bridget Moynahan.

This is true

This is not

Bridget Moynahan is an American actress, model and producer. She is best known for her roles in Grey's Anatomy, The Recruit, I, Robot and Blue Bloods. She studied acting at the American Academy of Dramatic Arts, and ...

Why is evaluation challenging?

Input: Tell me a bio of Bridget Moynahan.

This is true

This is not

Bridget Moynahan is an American actress, model and producer. She is best known for her roles in Grey's Anatomy, The Recruit, I, Robot and Blue Bloods. She studied acting at the American Academy of Dramatic Arts, and ...

Again, false

Why is evaluation challenging?

Input: Tell me a bio of Bridget Moynahan.

This is true

Bridget Moynahan is an American actress, model and

This is not

producer. She is best known for her roles in Grey's Anatomy,

Again, false

but these are all correct

The Recruit, I, Robot and Blue Bloods. She studied acting at

the American Academy of Dramatic Arts, and ...

Why is evaluation challenging?

Input: Tell me a bio of Bridget Moynahan.

This is true

Bridget Moynahan is an American actress, model and

This is not

producer. She is best known for her roles in Grey's Anatomy,

Again, false

but these are all correct

The Recruit, I, Robot and Blue Bloods. She studied acting at

the American Academy of Dramatic Arts, and ...

She did study acting

Why is evaluation challenging?

Input: Tell me a bio of Bridget Moynahan.

This is true

Bridget Moynahan is an American actress, model and

This is not

producer. She is best known for her roles in Grey's Anatomy,

Again, false

but these are all correct

The Recruit, I, Robot and Blue Bloods. She studied acting at

the American Academy of Dramatic Arts, and ...

She did study acting

but at the different institution...

Why is evaluation challenging?

Input: Tell me a bio of Bridget Moynahan.

This is true

Bridget Moynahan is an American actress, model and

This is not

producer. She is best known for her roles in Grey's Anatomy,

Again, false

but these are all correct

The Recruit, I, Robot and Blue Bloods. She studied acting at

the American Academy of Dramatic Arts, and ...

She did study acting

but at the different institution...

Why is evaluation challenging?

Input: Tell me a bio of Bridget Moynahan.

This is true

Bridget Moynahan is an American actress, model and

This is not

producer. She is best known for her roles in Grey's Anatomy, Again, false

but these are all correct

The Recruit, I, Robot and Blue Bloods. She studied acting at

the American Academy of Dramatic Arts, and ... She did study acting

but at the different institution...

Key challenge 1: Mixture of true and false info

Why is evaluation challenging?

Input: Tell me a bio of Bridget Moynahan.

This is true

This is not

Bridget Moynahan is an American actress, model and producer. She is best known for her roles in Grey's Anatomy, The Recruit, I, Robot and Blue Bloods. She studied acting at the American Academy of Dramatic Arts, and ...

Again, false

but these are all correct

She did study acting

but at the different institution...

Key challenge 1: Mixture of true and false info

Key challenge 2: Many pieces of info to verify (e.g. ~40)

→ Evaluation is costly (e.g., 10min/generation)

Why is evaluation challenging?

Key challenge 1: Mixture of true and false info

Key challenge 2: Many pieces of info to verify (e.g. ~40)

→ Evaluation is costly (e.g., 10min/generation)

Why is evaluation challenging?

- Define a "**Partial support**" label — Manakul et al. 2023, Liu et al. 2023
- Strictest definition of "**Support**" — Rashkin et al. 2021, Gao et al. 2022

Key challenge 1: Mixture of true and false info

Key challenge 2: Many pieces of info to verify (e.g. ~40)

→ Evaluation is costly (e.g., 10min/generation)

Why is evaluation challenging?

- Define a "**Partial support**" label — Manakul et al. 2023, Liu et al. 2023

Less well-defined

- Strictest definition of "**Support**" — Rashkin et al. 2021, Gao et al. 2022

Key challenge 1: Mixture of true and false info

Key challenge 2: Many pieces of info to verify (e.g. ~40)

→ Evaluation is costly (e.g., 10min/generation)

Why is evaluation challenging?

- Define a "**Partial support**" label — Manakul et al. 2023, Liu et al. 2023

Less well-defined

- Strictest definition of "**Support**" — Rashkin et al. 2021, Gao et al. 2022

Ignores partial support

Key challenge 1: Mixture of true and false info

Key challenge 2: Many pieces of info to verify (e.g. ~40)

→ Evaluation is costly (e.g., 10min/generation)

Why is evaluation challenging?

- Define a "**Partial support**" label — Manakul et al. 2023, Liu et al. 2023

Less well-defined

- Strictest definition of "**Support**" — Rashkin et al. 2021, Gao et al. 2022

Ignores partial support

Input: Tell me a bio of Bridget Moynahan.

This is true

Bridget Moynahan is an American actress, model and

This is not

producer. She is best known for her roles in Grey's Anatomy,

Again, false

but these are all correct

The Recruit, I, Robot and Blue Bloods. She studied acting at

the American Academy of Dramatic Arts, and ...

She did study acting

but at the different institution...

Why is evaluation challenging?

- Define a "**Partial support**" label — Manakul et al. 2023, Liu et al. 2023

Less well-defined

- Strictest definition of "**Support**" — Rashkin et al. 2021, Gao et al. 2022

Ignores partial support

Input: Tell me a bio of Bridget Moynahan.

This is true

Bridget Moynahan is an American actress, model and

This is not

producer. She is best known for her roles in Grey's Anatomy,

Again, false

but these are all correct

The Recruit, I, Robot and Blue Bloods. She studied acting at

the American Academy of Dramatic Arts, and ...

She did study acting

but at the different institution...

0.0

FActScore: Factual precision in Atomicity Score

FActScore: Factual precision in Atomicity Score

Bridget Moynahan is an American actress, model and producer. She is best known for her roles in Grey's Anatomy, I, Robot and Blue Bloods. She studied acting at the American Academy of Dramatic Arts, and ...

FActScore: Factual precision in Atomicity Score

Bridget Moynahan is an American actress, model and producer. She is best known for her roles in Grey's Anatomy, I, Robot and Blue Bloods. She studied acting at the American Academy of Dramatic Arts, and ...

Atomic facts

- Bridget Moynahan is American.
- Bridget Moynahan is an actress.
- Bridget Moynahan is a model.
- Bridget Moynahan is a producer.
- She is best known for her roles in Grey's Anatomy.
- She is best known for her roles in I, Robot.
- She is best known for her roles in Blue Bloods.
- She studied acting.
- She studied at the American Academy of Dramatic Arts.
- ...

FActScore: Factual precision in Atomicity Score

Bridget Moynahan is an American actress, model and producer. She is best known for her roles in Grey's Anatomy, I, Robot and Blue Bloods. She studied acting at the American Academy of Dramatic Arts, and ...

Atomic facts

- Bridget Moynahan is American.
- Bridget Moynahan is an actress.
- Bridget Moynahan is a model.
- Bridget Moynahan is a producer.
- She is best known for her roles in Grey's Anatomy.
- She is best known for her roles in I, Robot.
- She is best known for her roles in Blue Bloods.
- She studied acting.
- She studied at the American Academy of Dramatic Arts.
- ...

FActScore: Factual precision in Atomicity Score

Bridget Moynahan is an American actress, model and producer. She is best known for her roles in Grey's Anatomy, I, Robot and Blue Bloods. She studied acting at the American Academy of Dramatic Arts, and ...

Atomic facts

- Bridget Moynahan is American.
- Bridget Moynahan is an actress.
- Bridget Moynahan is a model.
- Bridget Moynahan is a producer.
- She is best known for her roles in Grey's Anatomy.
- She is best known for her roles in I, Robot.
- She is best known for her roles in Blue Bloods.
- She studied acting.
- She studied at the American Academy of Dramatic Arts.
- ...

FActScore: Factual precision in Atomicity Score

Bridget Moynahan is an American actress, model and producer. She is best known for her roles in Grey's Anatomy, I, Robot and Blue Bloods. She studied acting at the American Academy of Dramatic Arts, and ...

Atomic facts

- Bridget Moynahan is American. ✓
- Bridget Moynahan is an actress. ✓
- Bridget Moynahan is a model. ✓
- Bridget Moynahan is a producer. ✗
- She is best known for her roles in Grey's Anatomy. ✗
- She is best known for her roles in I, Robot. ✓
- She is best known for her roles in Blue Bloods. ✓
- She studied acting. ✓
- She studied at the American Academy of Dramatic Arts. ✗
- ...

FActScore: Factual precision in Atomicity Score

Bridget Moynahan is an American actress, model and producer. She is best known for her roles in Grey's Anatomy, I, Robot and Blue Bloods. She studied acting at the American Academy of Dramatic Arts, and ...

Atomic facts

- Bridget Moynahan is American. ✓
- Bridget Moynahan is an actress. ✓
- Bridget Moynahan is a model. ✓
- Bridget Moynahan is a producer. ✗
- She is best known for her roles in Grey's Anatomy. ✗
- She is best known for her roles in I, Robot. ✓
- She is best known for her roles in Blue Bloods. ✓
- She studied acting. ✓
- She studied at the American Academy of Dramatic Arts. ✗
- ...

66.7%

FActScore: Factual precision in Atomicity Score

Bridget Moynahan is an American filmmaker and writer. She is best known for her work on the soap opera General Hospital, which she co-created with husband Charles Kelly. Moynahan was raised in a middle-class family in Los Angeles, ...

FActScore: Factual precision in Atomicity Score

Bridget Moynahan is an American filmmaker and writer. She is best known for her work on the soap opera General Hospital, which she co-created with husband Charles Kelly. Moynahan was raised in a middle-class family in Los Angeles, ...

Atomic facts

- Bridget Moynahan is American.
- Bridget Moynahan is a filmmaker.
- Bridget Moynahan is a writer.
- She is best known for her work on General Hospital.
- General Hospital is the soap opera.
- She co-created General Hospital.
- She co-created General Hospital with her husband.
- Her husband is Charles Kelly.
- Moynahan was raised in a middle-class family.
- Moynahan was raised in Los Angeles.
- ...

FActScore: Factual precision in Atomicity Score

Bridget Moynahan is an American filmmaker and writer. She is best known for her work on the soap opera General Hospital, which she co-created with husband Charles Kelly. Moynahan was raised in a middle-class family in Los Angeles, ...

Atomic facts

- Bridget Moynahan is American.
- Bridget Moynahan is a filmmaker.
- Bridget Moynahan is a writer.
- She is best known for her work on General Hospital.
- General Hospital is the soap opera.
- She co-created General Hospital.
- She co-created General Hospital with her husband.
- Her husband is Charles Kelly.
- Moynahan was raised in a middle-class family.
- Moynahan was raised in Los Angeles.
- ...

FActScore: Factual precision in Atomicity Score

Bridget Moynahan is an American filmmaker and writer. She is best known for her work on the soap opera General Hospital, which she co-created with husband Charles Kelly. Moynahan was raised in a middle-class family in Los Angeles, ...

Atomic facts

- Bridget Moynahan is American. ✓
- Bridget Moynahan is a filmmaker. ✗
- Bridget Moynahan is a writer. ✗
- She is best known for her work on General Hospital. ✗
- General Hospital is the soap opera. ✗
- She co-created General Hospital. ✗
- She co-created General Hospital with her husband. ✗
- Her husband is Charles Kelly. ✗
- Moynahan was raised in a middle-class family. ✗
- Moynahan was raised in Los Angeles. ✗
- ...

FActScore: Factual precision in Atomicity Score

Bridget Moynahan is an American filmmaker and writer. She is best known for her work on the soap opera General Hospital, which she co-created with husband Charles Kelly. Moynahan was raised in a middle-class family in Los Angeles, ...

Atomic facts

- Bridget Moynahan is American. ✓
- Bridget Moynahan is a filmmaker. ✗
- Bridget Moynahan is a writer. ✗
- She is best known for her work on General Hospital. ✗
- General Hospital is the soap opera. ✗
- She co-created General Hospital. ✗
- She co-created General Hospital with her husband. ✗
- Her husband is Charles Kelly. ✗
- Moynahan was raised in a middle-class family. ✗
- Moynahan was raised in Los Angeles. ✗
- ...

10.0%

FActScore: Factual precision in Atomicity Score

	Previous work	FActScore
Bridget Moynahan is an American actress, model and producer. She is best known for her roles in Grey's Anatomy ...	0.0	66.7
Bridget Moynahan is an American filmmaker and writer. She is best known for her work on the soap opera General Hospital, ...	0.0	10.0

FActScore: Definition

$\mathcal{X}$: a set of prompts

$\mathcal{C}$: a knowledge source

FActScore: Definition

$\mathcal{X}$: a set of prompts

$\mathcal{C}$: a knowledge source

$y = \mathcal{M}_x$: a response from the model $\mathcal{M}$ for $x \in \mathcal{X}$

FActScore: Definition

$\mathcal{X}$: a set of prompts

$\mathcal{C}$: a knowledge source

$y = \mathcal{M}_x$: a response from the model $\mathcal{M}$ for $x \in \mathcal{X}$

$\mathcal{A}_y$: a set of atomic facts in y

FActScore: Definition

$\mathcal{X}$: a set of prompts

$\mathcal{C}$: a knowledge source

$y = \mathcal{M}_x$: a response from the model $\mathcal{M}$ for $x \in \mathcal{X}$

$\mathcal{A}_y$: a set of atomic facts in y

$$f(y) = \frac{1}{|\mathcal{A}_y|} \sum_{a \in \mathcal{A}_y} [a \text{ is supported by } \mathcal{C}]$$

$$\textsc{FactScore}(\mathcal{M}) = \mathbb{E}_{x \in \mathcal{X}}[f(\mathcal{M}_x) | \mathcal{M}_x \text{ responds}]$$

FActScore: Definition

$\mathcal{X}$: a set of prompts

$\mathcal{C}$: a knowledge source

$y = \mathcal{M}_x$: a response from the model $\mathcal{M}$ for $x \in \mathcal{X}$

$\mathcal{A}_y$: a set of atomic facts in y

$$f(y) = \frac{1}{|\mathcal{A}_y|} \sum_{a \in \mathcal{A}_y} [a \text{ is supported by } \mathcal{C}]$$

$$\textsc{FactScore}(\mathcal{M}) = \mathbb{E}_{x \in \mathcal{X}}[f(\mathcal{M}_x) | \mathcal{M}_x \text{ responds}]$$

We'll exclude if the model abstains from responding

FActScore

In the rest of the presentation:

People biographies &

FActScore

Facts are objective & specific
(No subjective, vague)

In the rest of the presentation:

People biographies &

FActScore

Facts are objective & specific
(No subjective, vague)

In the rest of the presentation:

People biographies &

Wikipedia has high coverage
(& consistent)

FActScore

Facts are objective & specific
(No subjective, vague)

In the rest of the presentation:

People biographies &

Wikipedia has high coverage
(& consistent)

Good-enough coverage

FActScore

Facts are objective & specific
(No subjective, vague)

In the rest of the presentation:

People biographies &

Wikipedia has high coverage
(& consistent)

Good-enough coverage

We later discuss which types of generations FActScore can & cannot handle

FActScore

Facts are objective & specific
(No subjective, vague)

In the rest of the presentation:

People biographies &

Wikipedia has high coverage
(& consistent)

Good-enough coverage

Prompting LM to generate bios is actually a good way of evaluating factuality of a new LM

We later discuss which types of generations FActScore can & cannot handle

Human Evaluation
(experts)

Automatic Evaluation
(model-based)

Human Evaluation (experts)

Automatic Evaluation (model-based)

Key challenge 1: Mixture of true and false info

Key challenge 2: Many pieces of info to verify → costly

Address challenge 1

Human Evaluation (experts)

Automatic Evaluation (model-based)

Key challenge 1: Mixture of true and false info

Key challenge 2: Many pieces of info to verify → costly

Address challenge 1

Address challenge 1 & 2

Human Evaluation (experts)

Automatic Evaluation (model-based)

Key challenge 1: Mixture of true and false info

Key challenge 2: Many pieces of info to verify → costly

Human Evaluation
(experts)

Automatic Evaluation
(model-based)

Human Evaluation of FActScore

Step 1:
Preparing
Prompts

WIKIDATA

- Rarity of entities
- Country of citizenship

Human Evaluation of FActScore

Step 1:
Preparing
Prompts

Step 2:
Obtaining Generations
from the LMeval

InstructGPT

ChatGPT

Perplexity

Human Evaluation of FActScore

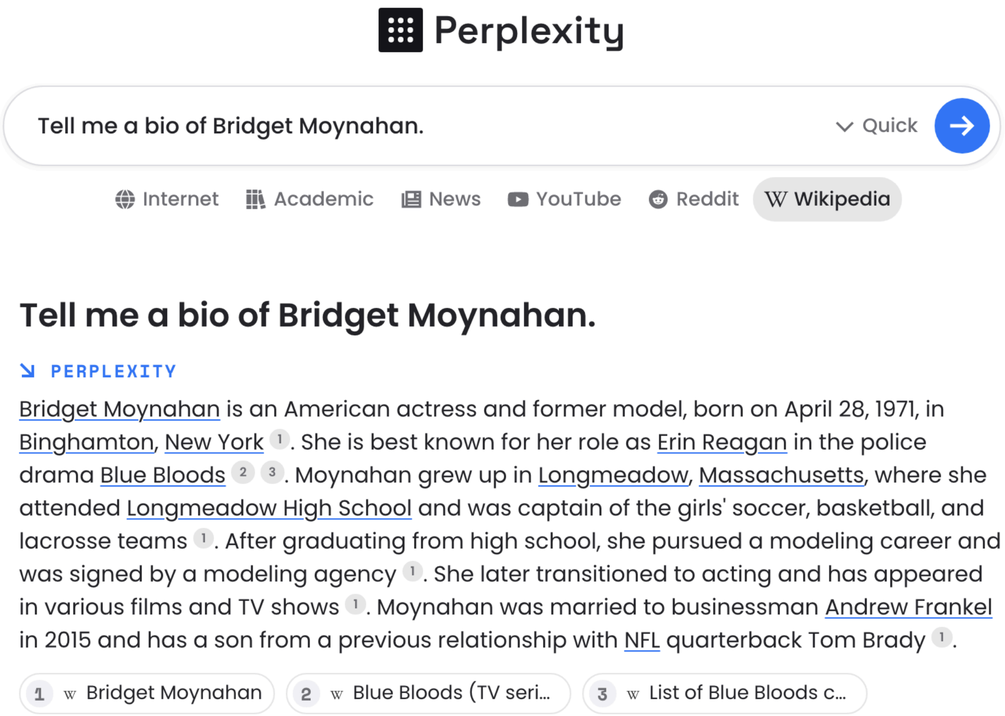
Step 1:
Preparing
Prompts

Step 2:
Obtaining Generations
from the LMeval

InstructGPT

ChatGPT

Perplexity



Human Evaluation of FActScore

Step 1:
Preparing
Prompts

Step 2:
Obtaining Generations
from the LMeval

Step 3:
Atomic Fact Break-
down

InstructGPT

upwork

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.

Human Evaluation of FActScore

Step 1:
Preparing
Prompts

Step 2:
Obtaining Generations
from the LMeval

Step 3:
Atomic Fact Break-
down

InstructGPT

upwork

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.

26—41 atomic facts

Human Evaluation of FActScore

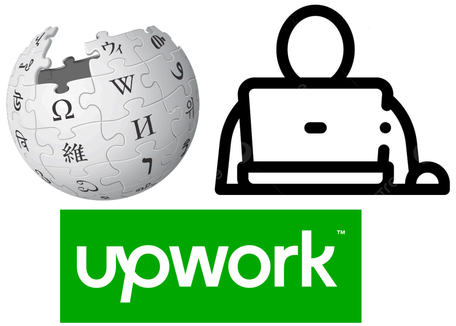
Step 1:
Preparing
Prompts

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.



Irrelevant

Supported

Not-supported

Human Evaluation of FActScore

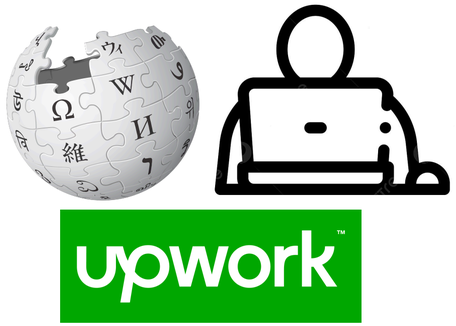
Step 1: Preparing Prompts

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.



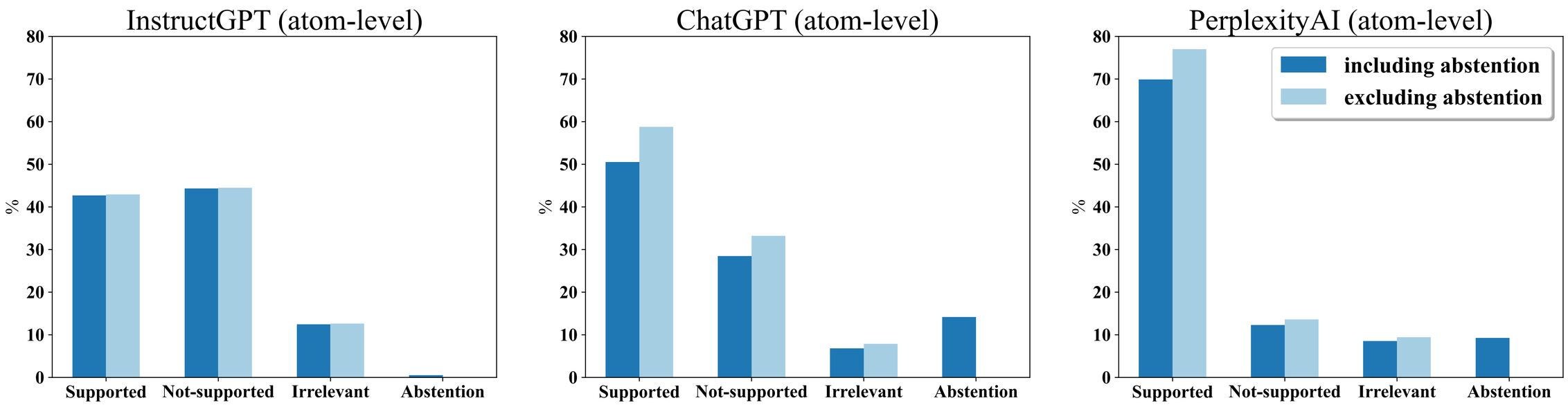
Irrelevant

Supported

Not-supported

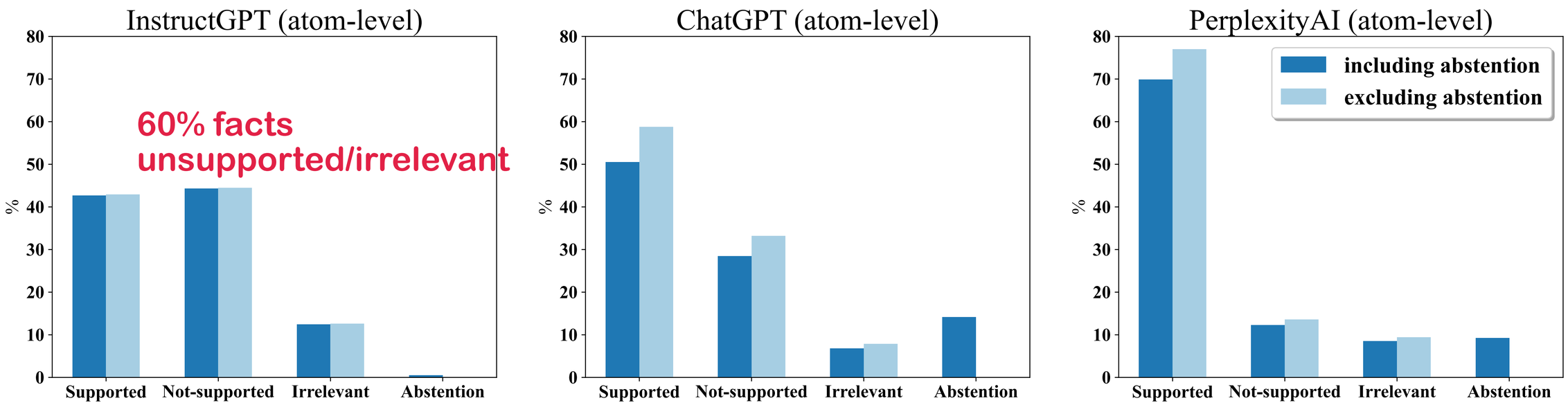
~25min for three model generations

Results



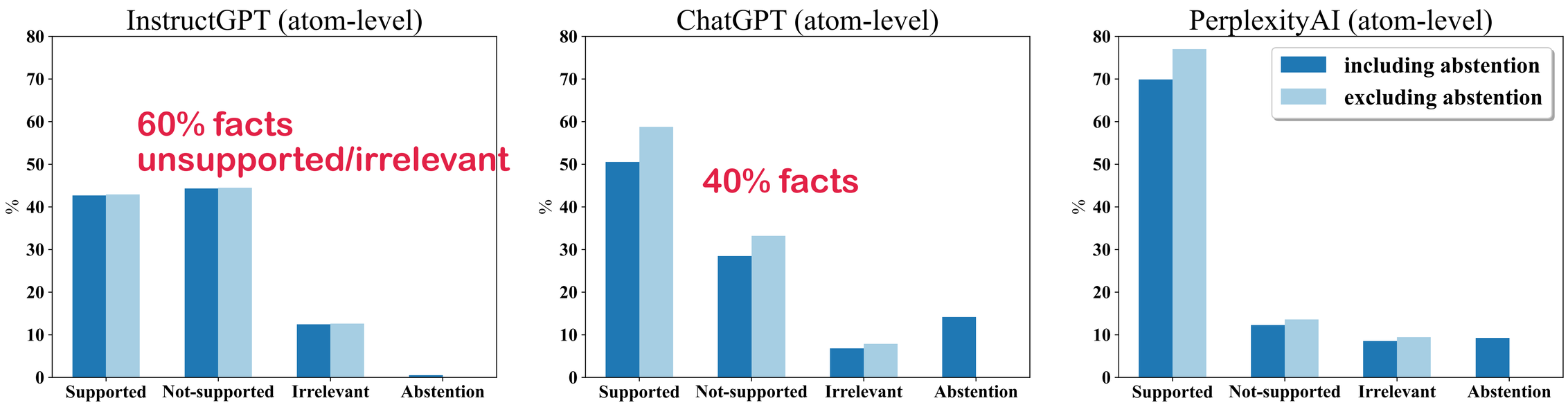
All models make substantial precision errors

Results



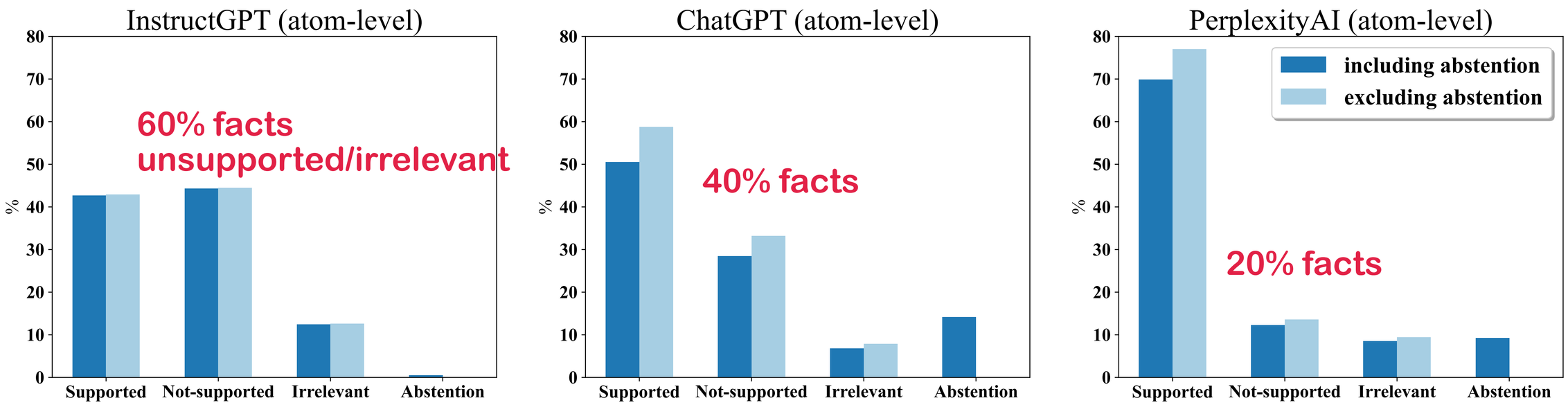
All models make substantial precision errors

Results



All models make substantial precision errors

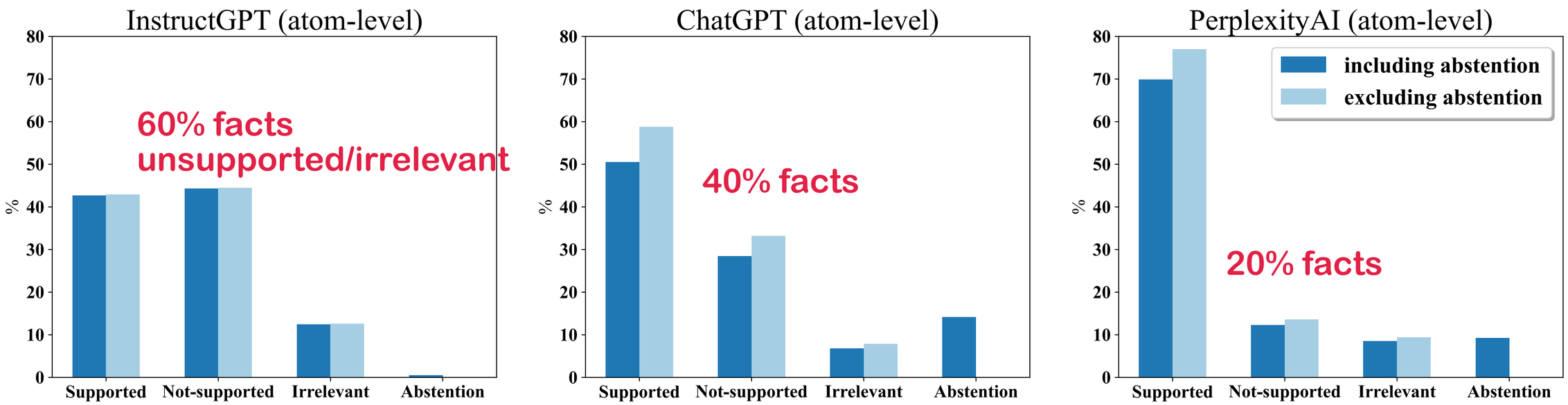
Results



All models make substantial precision errors

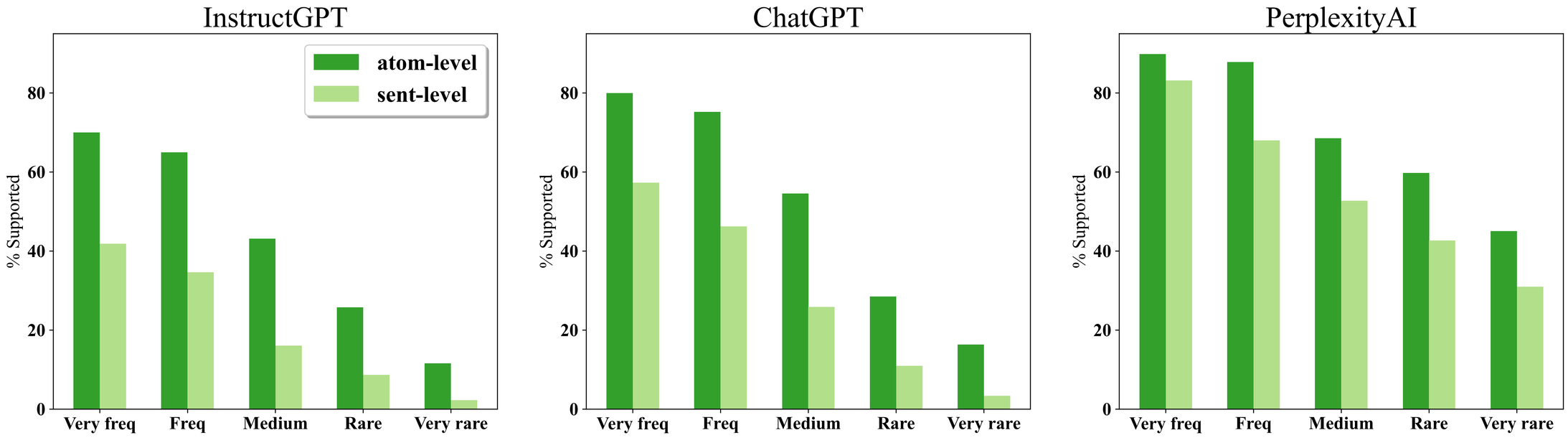
Results

Incorporating search helps!
But…



All models make substantial precision errors

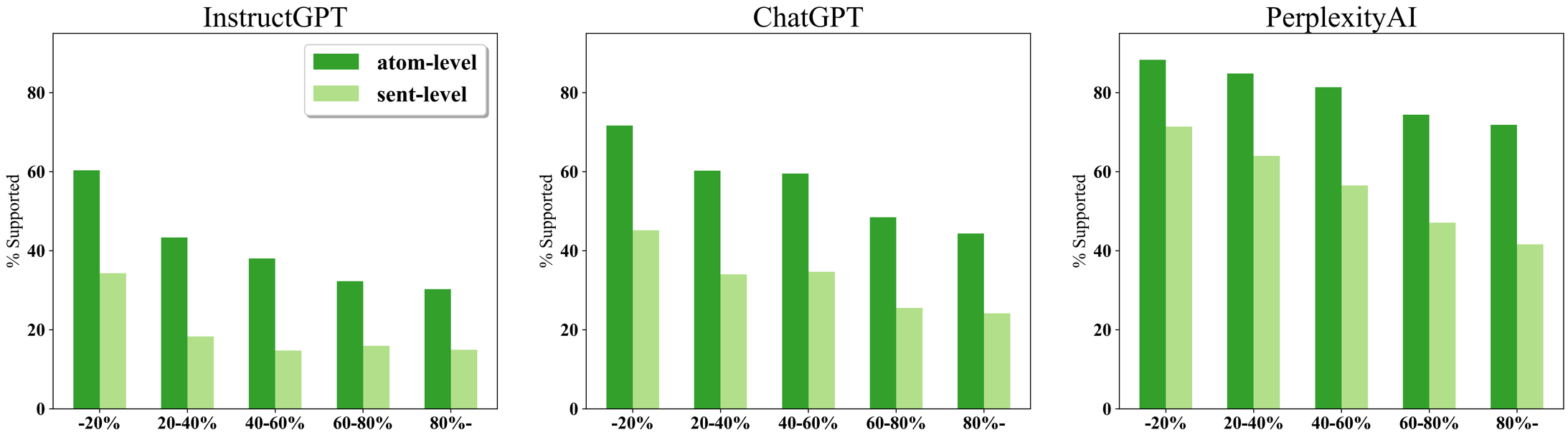
Results



More errors as the rarity of entities increases

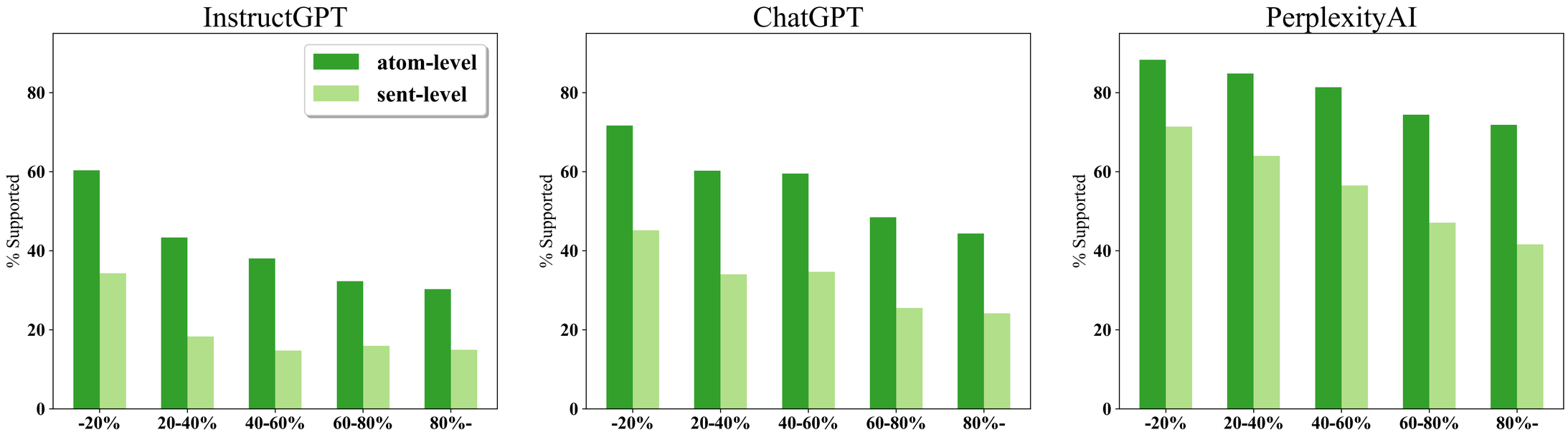
In agreement with QA results in Kandpal et al. (2022) and Mallen et al. (2022) in models w/o search
Different from QA, same trend in models w/ search

Results



More errors at a later part of generation

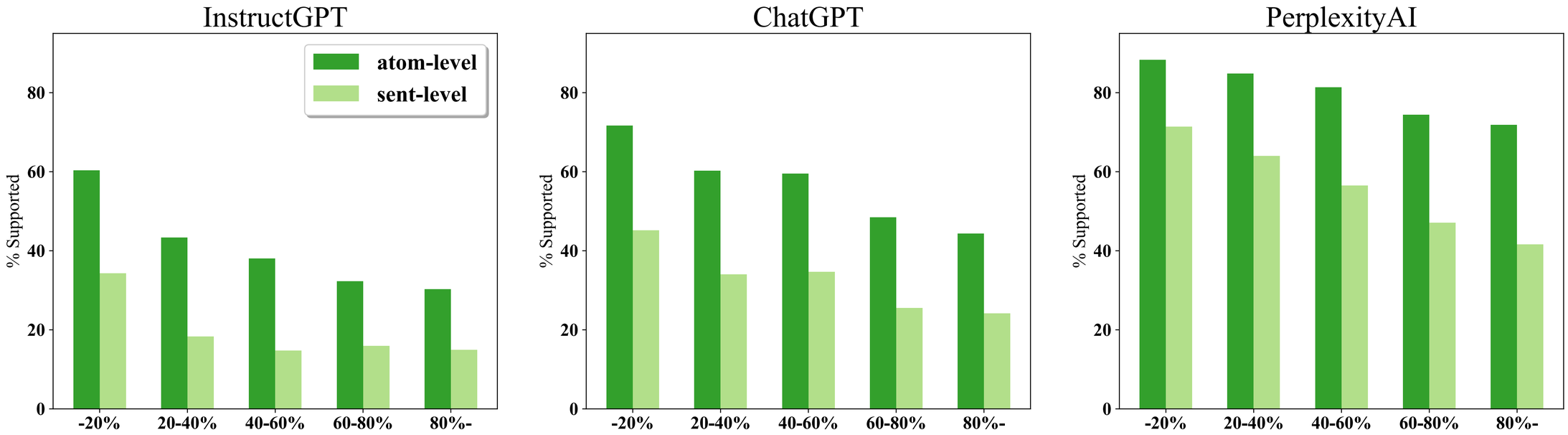
Results



More errors at a later part of generation

Reason #1: Error propagation

Results

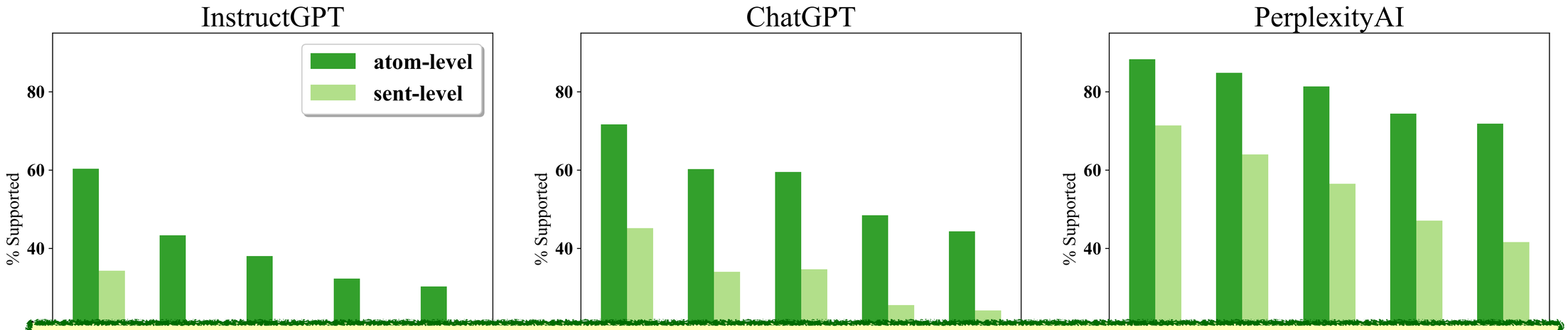


More errors at a later part of generation

Reason #1: Error propagation

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

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

Human Evaluation
(experts)

Automatic Evaluation
(model-based)

Can we estimate FActScore?

Bridget Moynahan studied acting at the American Academy of Dramatic Arts.

Can we estimate FActScore?

Bridget Moynahan studied acting at the American Academy of Dramatic Arts.

Retrieval

Can we estimate FActScore?

Bridget Moynahan studied acting at the American Academy of Dramatic Arts.

Retrieval

Bridget Moynahan

(…) in addition to taking acting and art classes. She studied acting at the Caymichael Patten Studio in New York, and in 1999, made her TV debut (…)

Can we estimate FActScore?

Bridget Moynahan studied acting at the American Academy of Dramatic Arts.

Retrieval

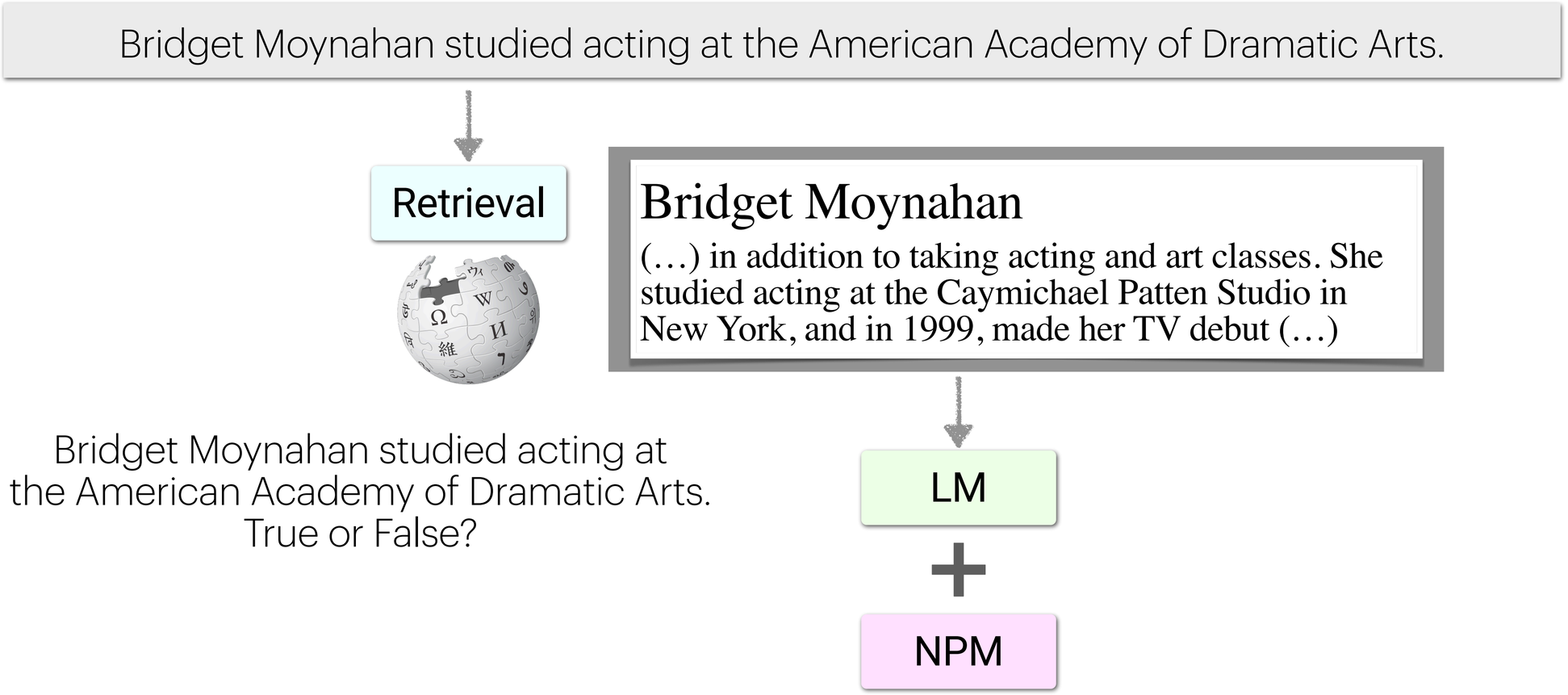
Bridget Moynahan

(…) in addition to taking acting and art classes. She studied acting at the Caymichael Patten Studio in New York, and in 1999, made her TV debut (…)

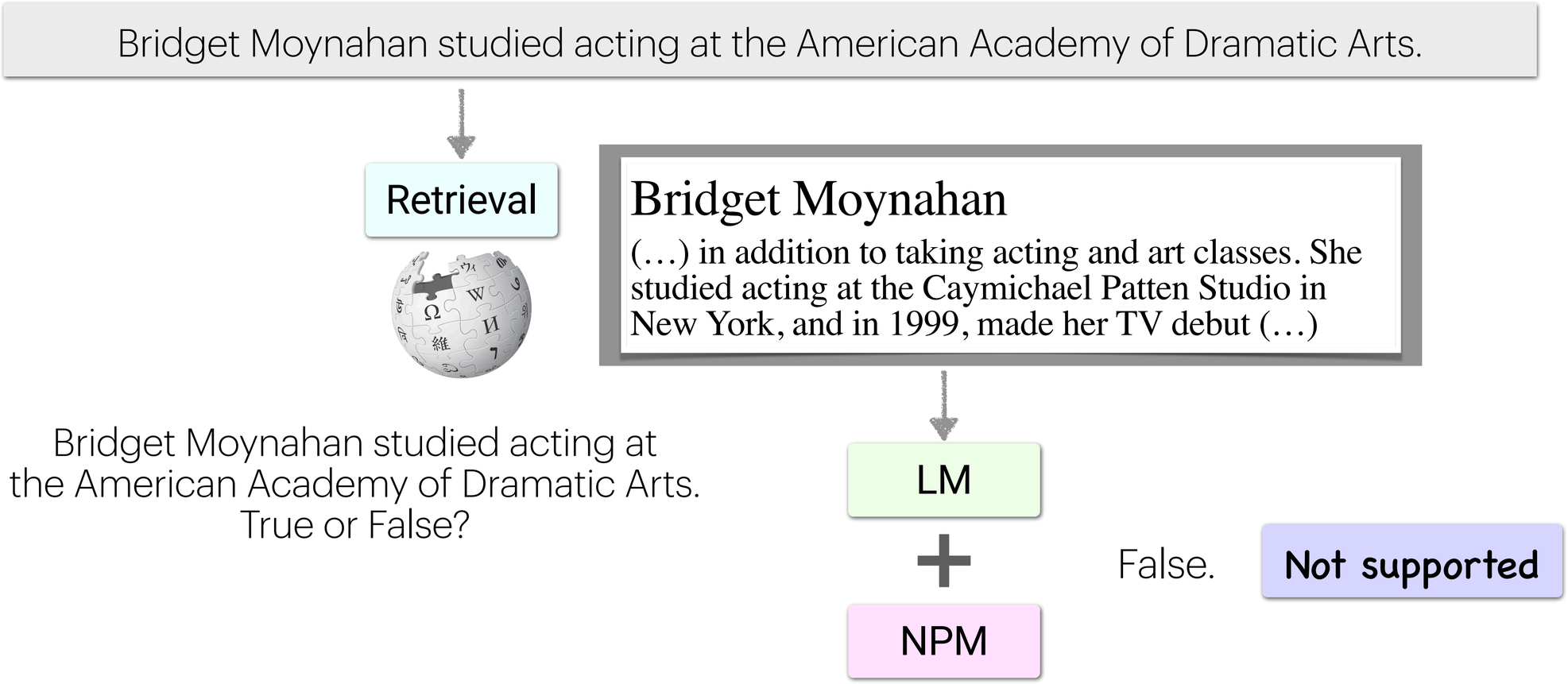
LM

Bridget Moynahan studied acting at the American Academy of Dramatic Arts. True or False?

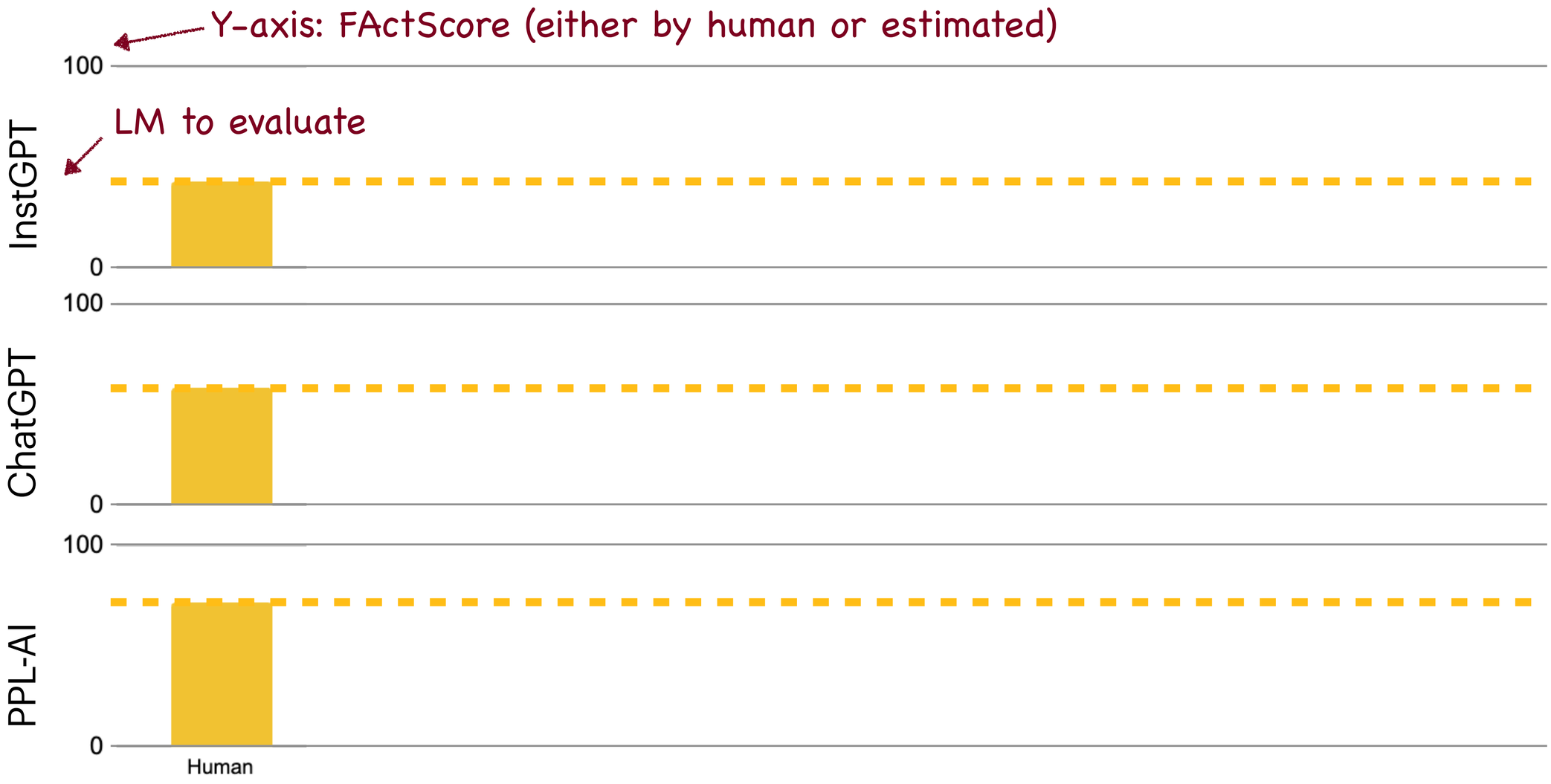
Can we estimate FActScore?



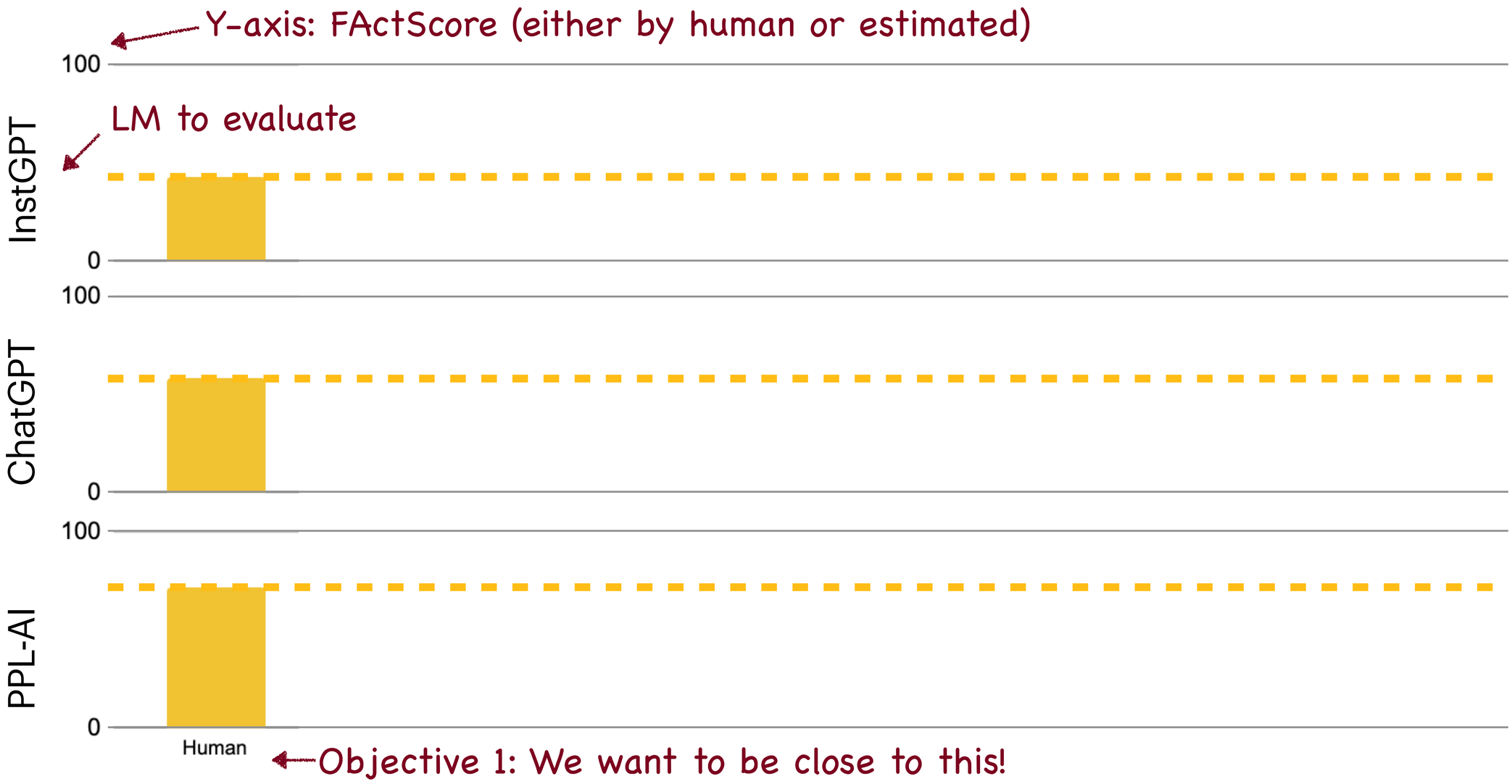
Can we estimate FActScore?



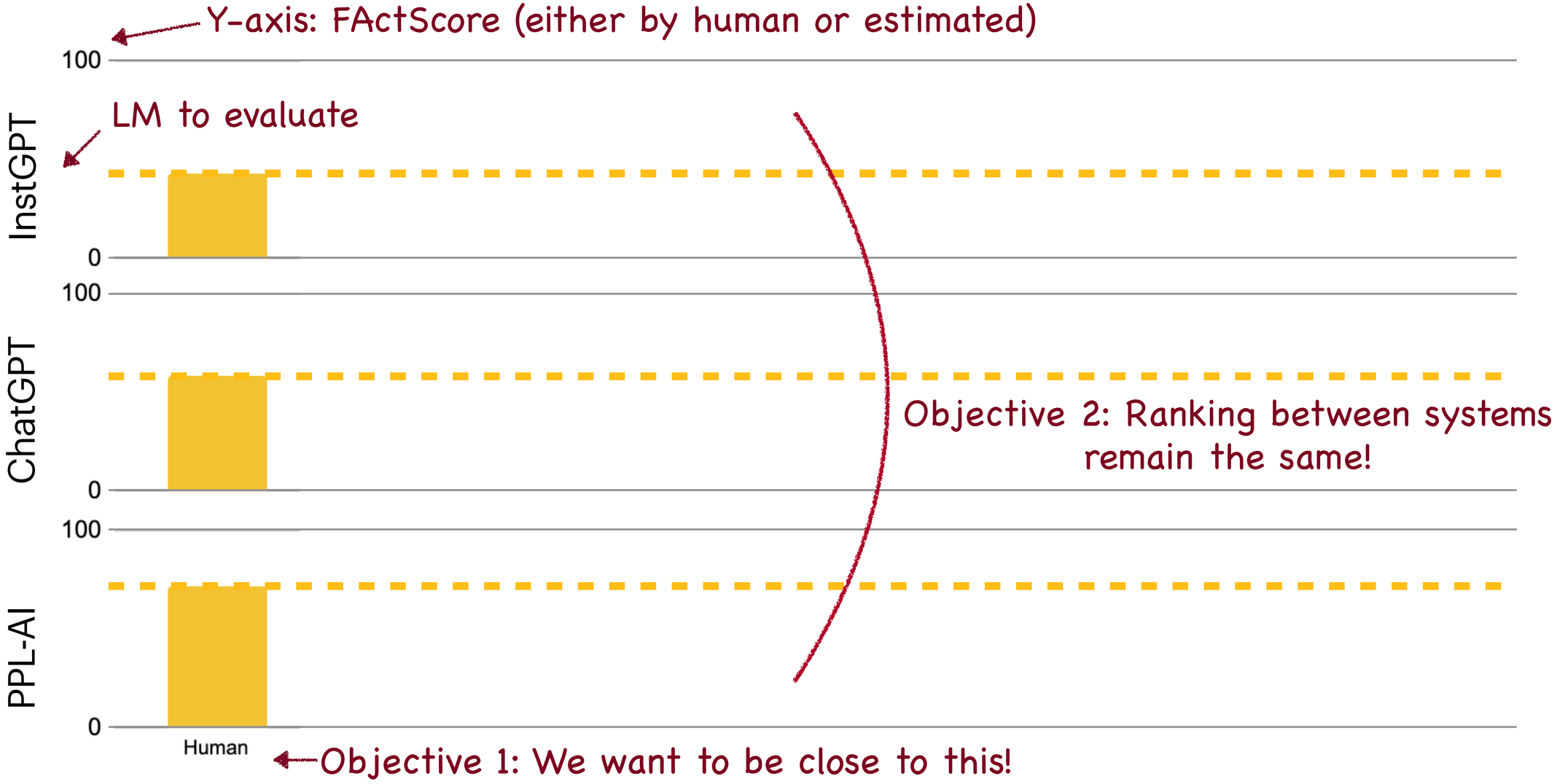
Results


Y-axis: FActScore (either by human or estimated)
LM to evaluate
100
0
InstGPT
100
0
ChatGPT
100
0
PPL-AI
Human

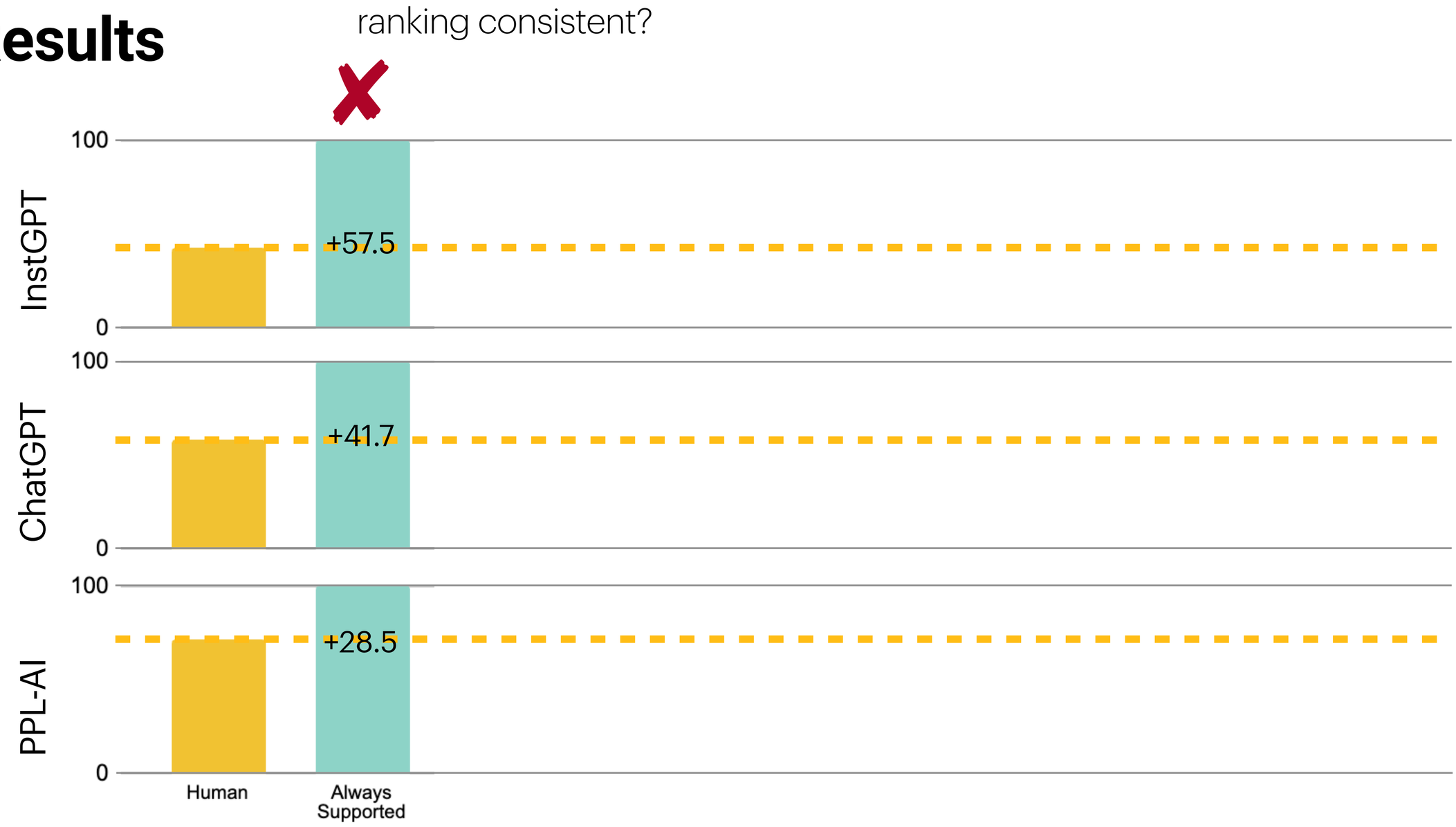
Results


Y-axis: FActScore (either by human or estimated)
LM to evaluate
InstGPT
ChatGPT
PPL-AI
100
0
100
0
100
0
Human
Objective 1: We want to be close to this!

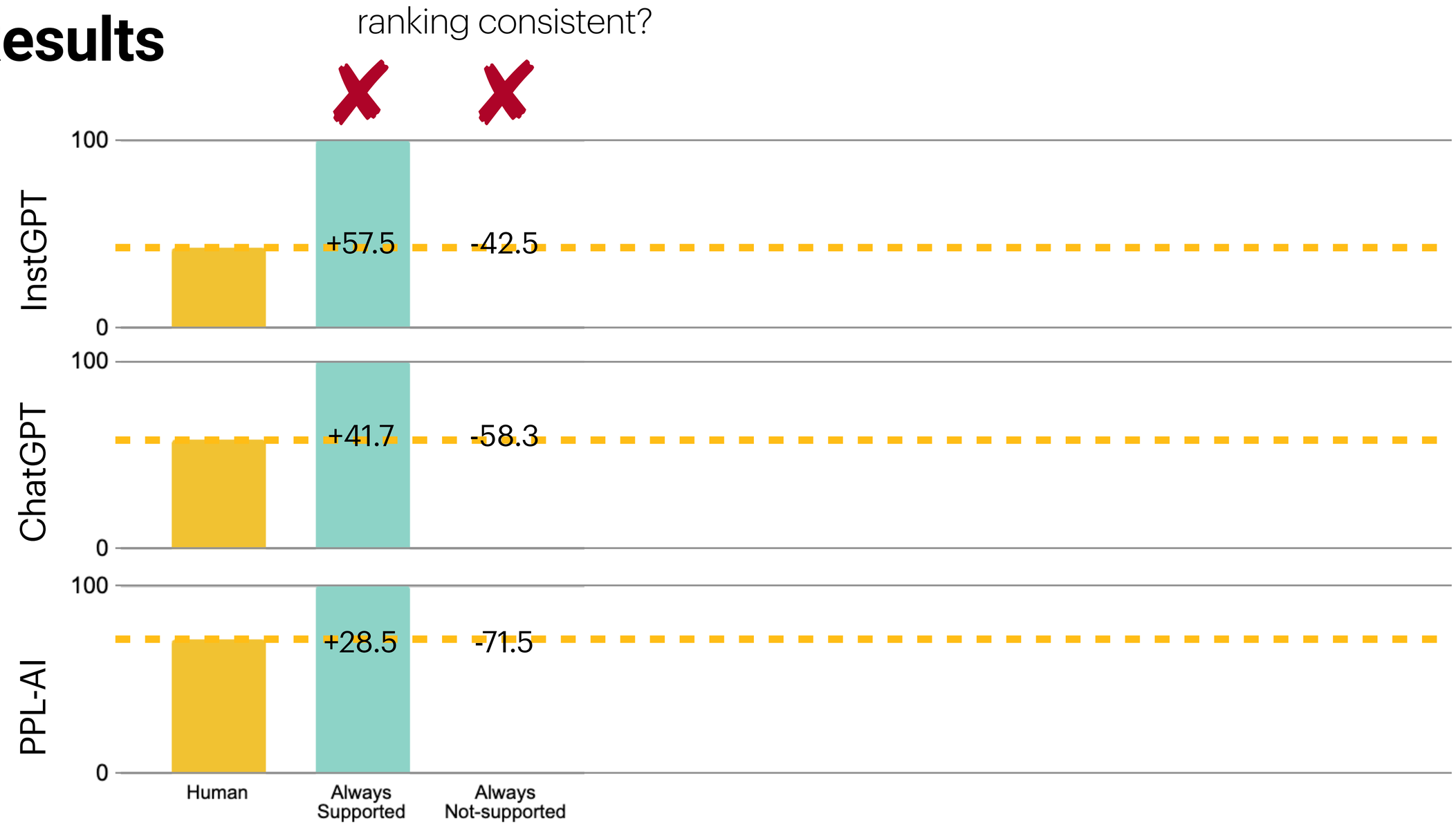
Results


Y-axis: FActScore (either by human or estimated)
100
LM to evaluate
InstGPT
0
100
ChatGPT
0
100
PPL-AI
0
Human
Objective 2: Ranking between systems remain the same!
Objective 1: We want to be close to this!

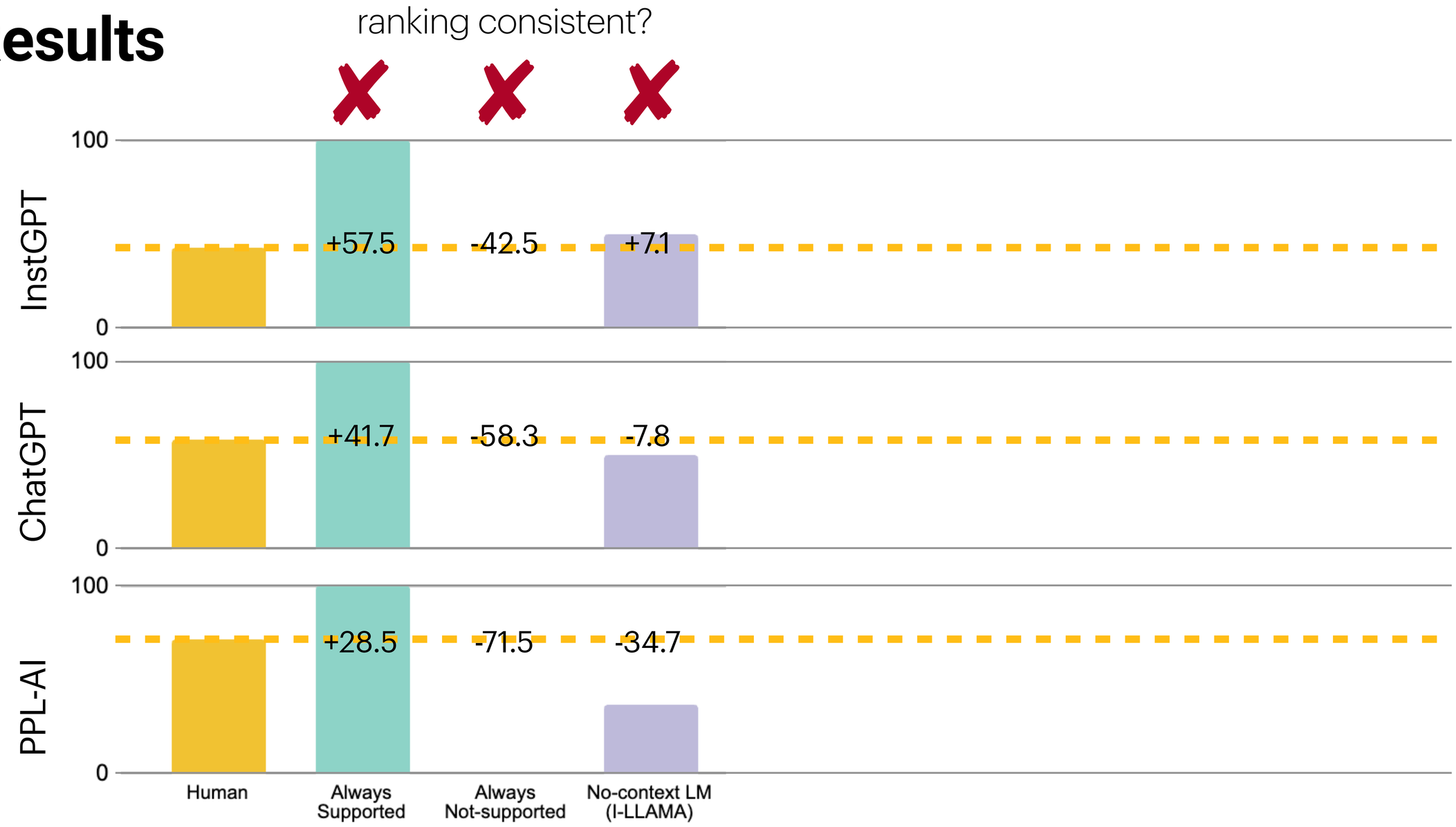
Results


ranking consistent?
100
0
InstGPT
+57.5
100
0
ChatGPT
+41.7
100
0
PPL-AI
+28.5
Human
Always Supported

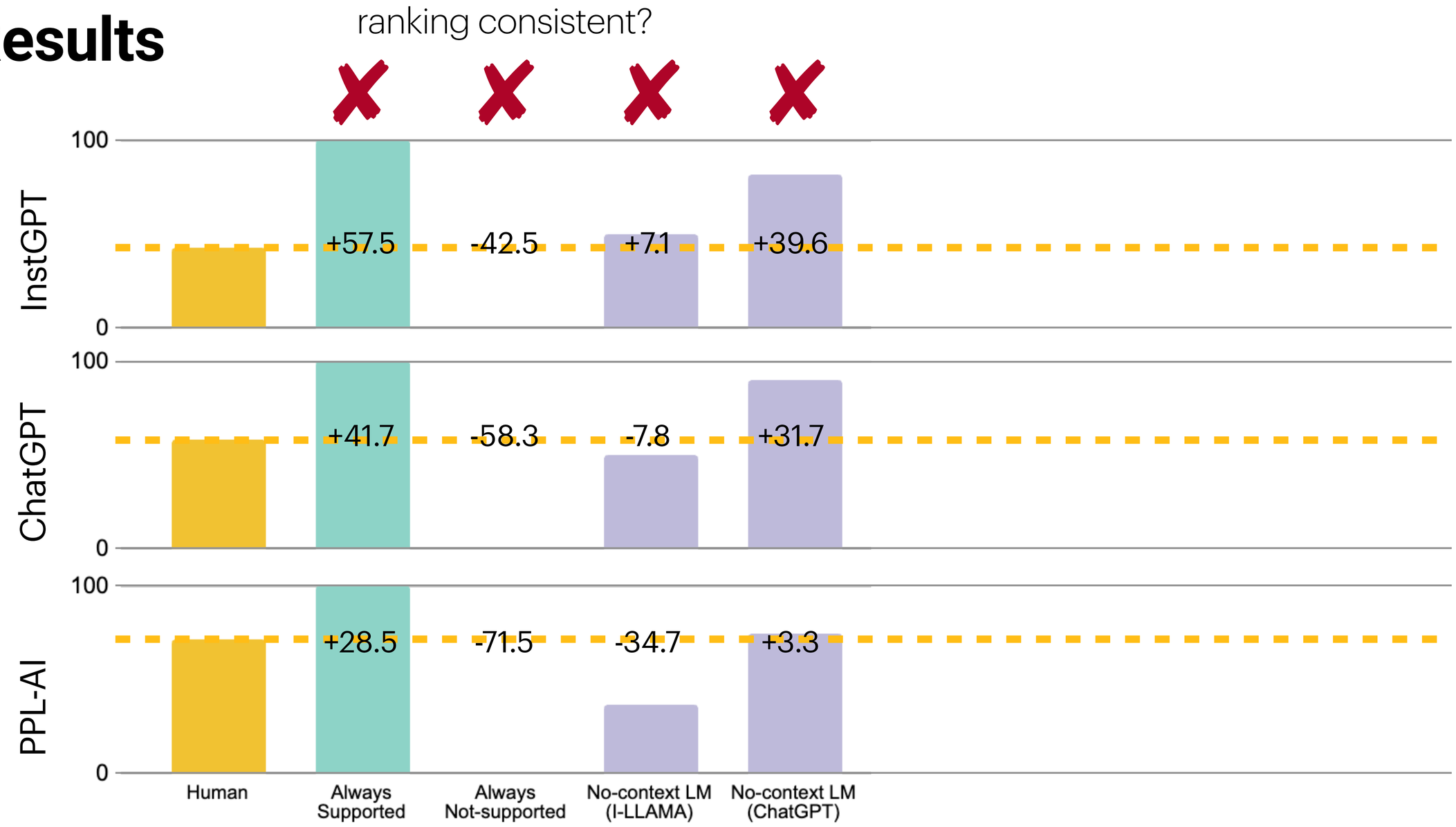
Results


ranking consistent?
100
0
InstGPT
+57.5
-42.5
100
0
ChatGPT
+41.7
-58.3
100
0
PPL-AI
+28.5
-71.5
Human
Always Supported
Always Not-supported

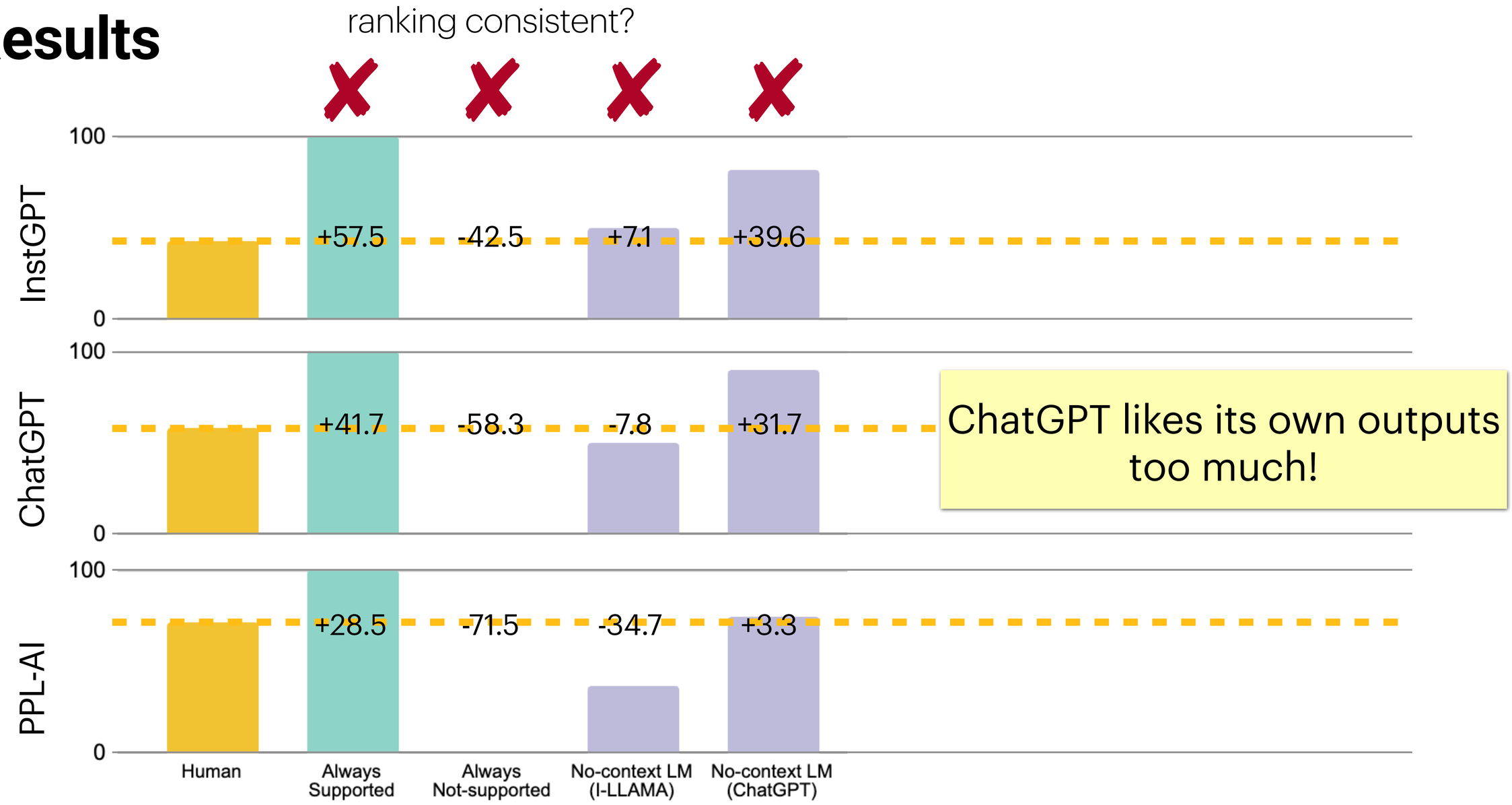
Results


ranking consistent?
InstGPT
ChatGPT
PPL-AI
100
0
+57.5
-42.5
+7.1
+41.7
-58.3
-7.8
+28.5
-71.5
-34.7
Human
Always Supported
Always Not-supported
No-context LM (I-LLAMA)

Results

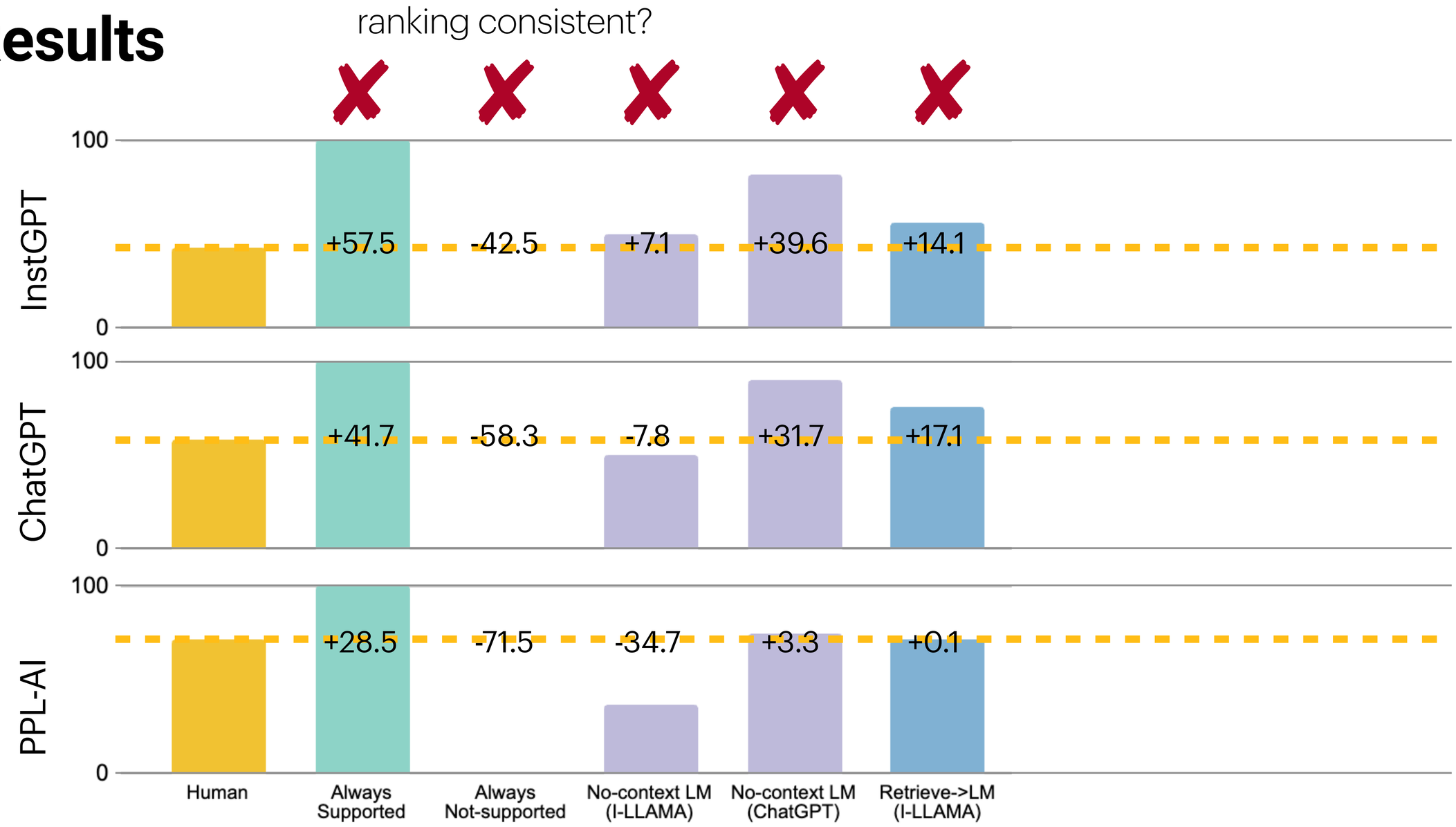

ranking consistent?
InstGPT
ChatGPT
PPL-AI
100
0
+57.5
-42.5
+7.1
+39.6
+41.7
-58.3
-7.8
+31.7
+28.5
-71.5
-34.7
+3.3
Human
Always Supported
Always Not-supported
No-context LM (I-LLAMA)
No-context LM (ChatGPT)

Results

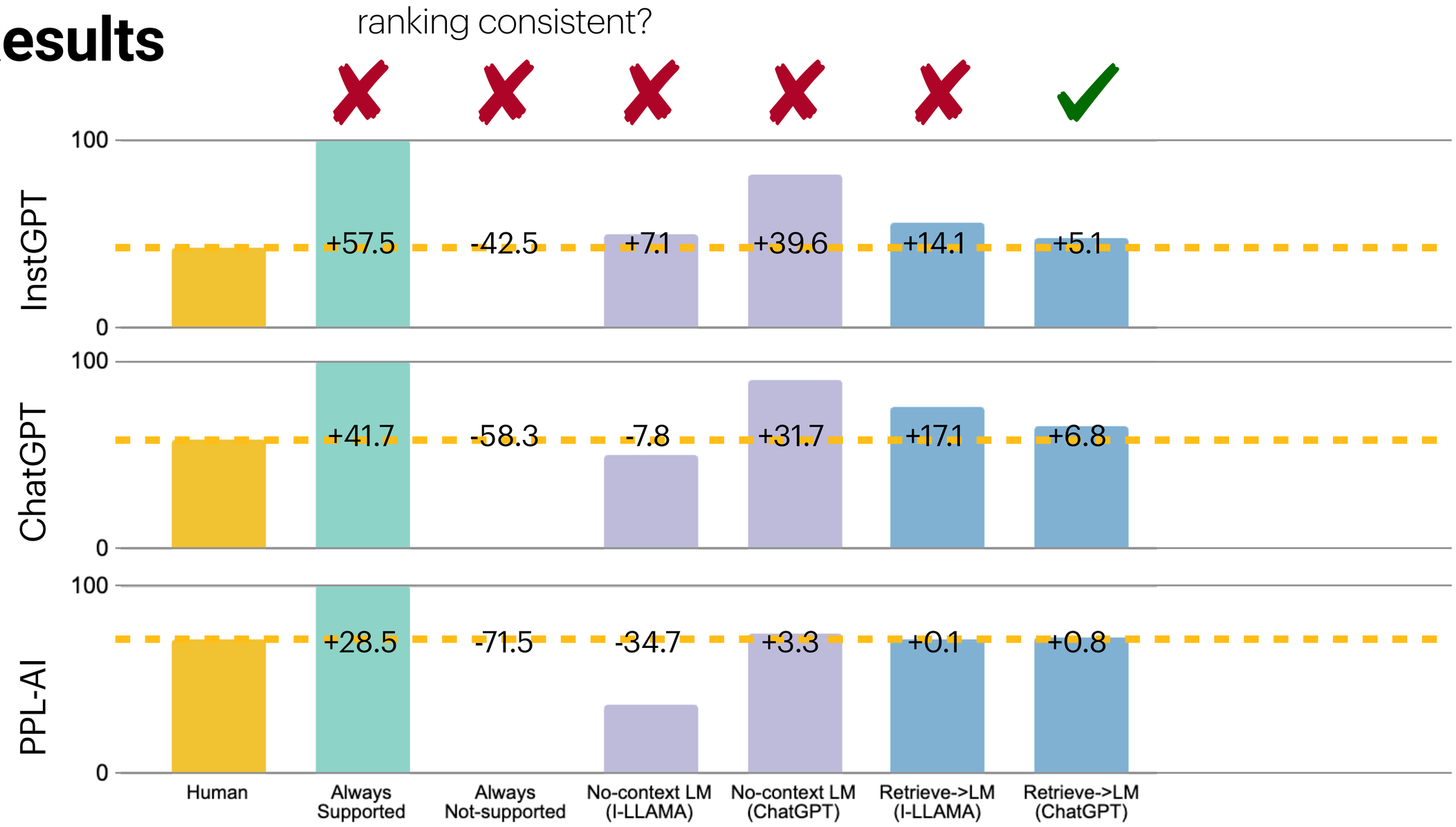


ChatGPT likes its own outputs too much!

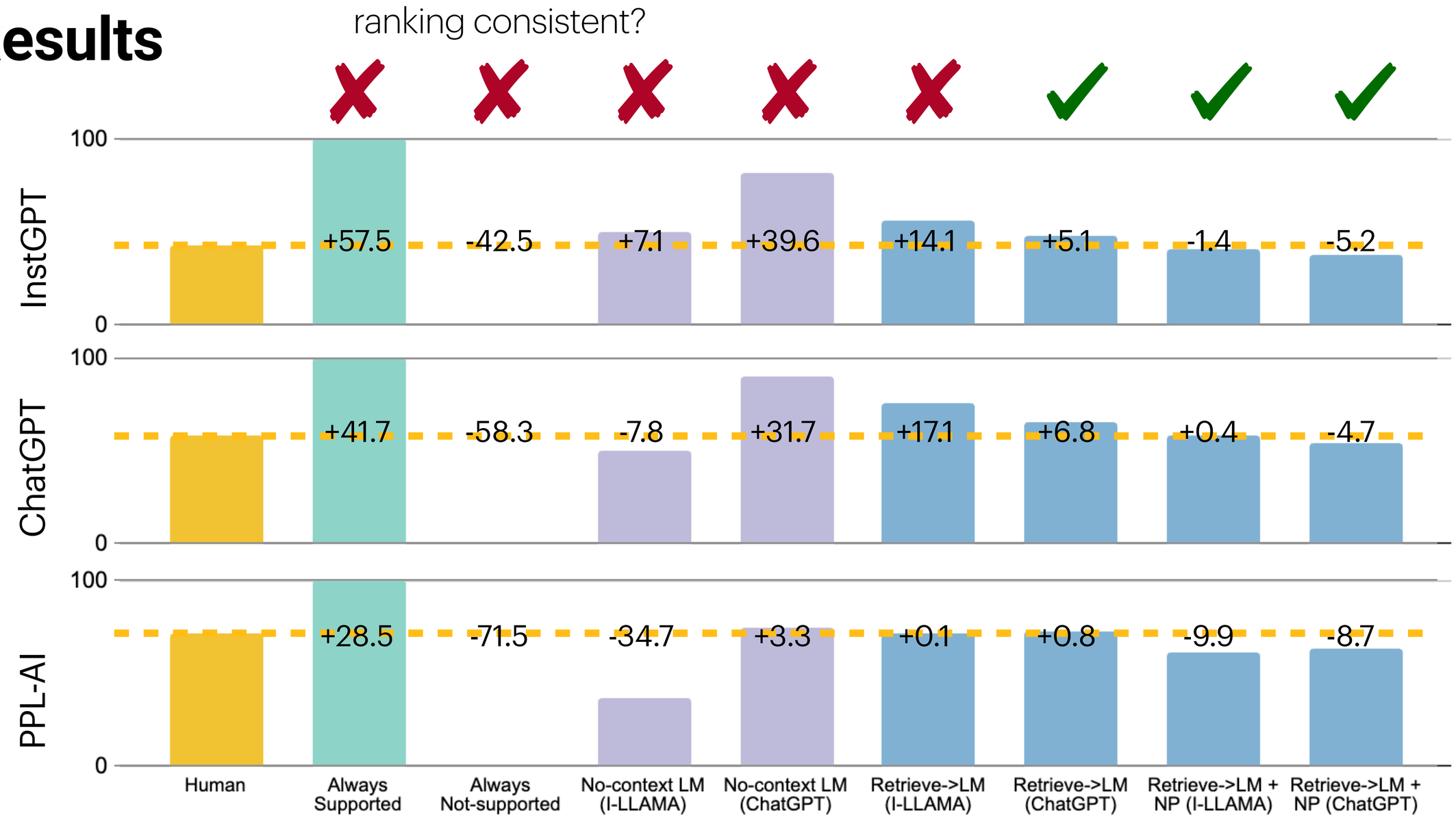
Results


ranking consistent?
InstGPT
ChatGPT
PPL-AI
100
0
+57.5
-42.5
+7.1
+39.6
+14.1
+41.7
-58.3
-7.8
+31.7
+17.1
+28.5
-71.5
-34.7
+3.3
+0.1
Human
Always Supported
Always Not-supported
No-context LM (I-LLAMA)
No-context LM (ChatGPT)
Retrieve->LM (I-LLAMA)

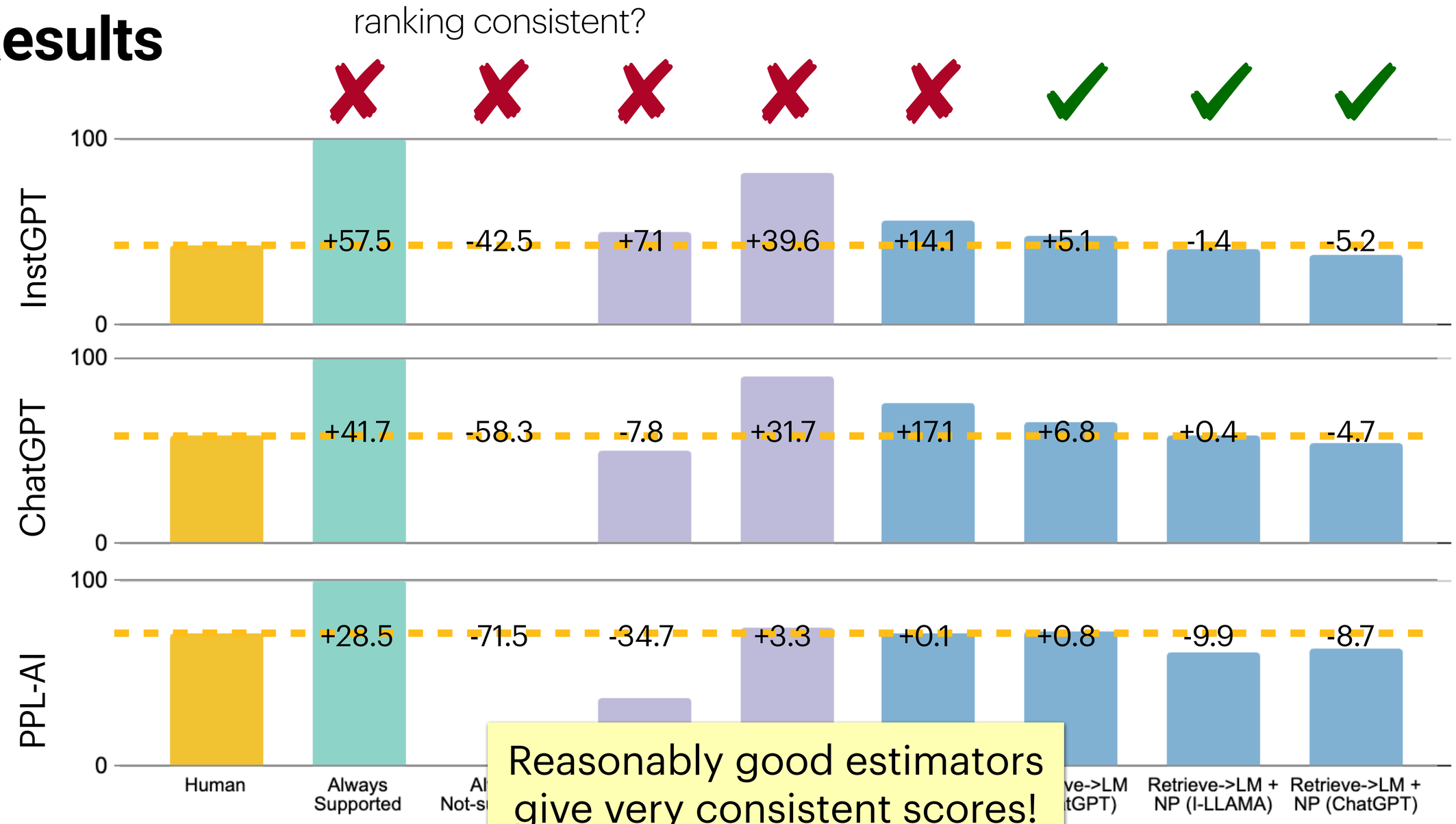
Results


ranking consistent?
InstGPT
ChatGPT
PPL-AI
100
0
+57.5
-42.5
+7.1
+39.6
+14.1
+5.1
+41.7
-58.3
-7.8
+31.7
+17.1
+6.8
+28.5
-71.5
-34.7
+3.3
+0.1
+0.8
Human
Always Supported
Always Not-supported
No-context LM (I-LLAMA)
No-context LM (ChatGPT)
Retrieve->LM (I-LLAMA)
Retrieve->LM (ChatGPT)

Results


ranking consistent?
InstGPT
ChatGPT
PPL-AI
100
0
+57.5
-42.5
+7.1
+39.6
+14.1
+5.1
-1.4
-5.2
+41.7
-58.3
-7.8
+31.7
+17.1
+6.8
+0.4
-4.7
+28.5
-71.5
-34.7
+3.3
+0.1
+0.8
-9.9
-8.7
Human
Always Supported
Always Not-supported
No-context LM (I-LLAMA)
No-context LM (ChatGPT)
Retrieve->LM (I-LLAMA)
Retrieve->LM (ChatGPT)
Retrieve->LM + NP (I-LLAMA)
Retrieve->LM + NP (ChatGPT)

Results



Reasonably good estimators give very consistent scores!
(We'll see very soon!)

Let’s apply FActScore to LMs in the wild

Let’s apply FActScore to LMs in the wild

	Base LM	Data	Release
InstructGPT	?	?	November 2022
ChatGPT	?	?	November 2022
GPT-4	?	?	March 2023
Alpaca {7B, 13B, 65B}	LLAMA	InstuctGPT outputs	March 2023
Vicuna {7B, 13B}	LLAMA	ShareGPT (ChatGPT)	March 2023
Dolly 12B	Pythia 12B	Databricks Dolly (human)	March 2023
Oasst-pythia 12B	Pythia 12B	Open Assistant (human)	March 2023
StableLM-tuned 7B	StableLM-base	Alpaca, Dolly, ShareGPT, GPT4All, Anthropic HH	April 2023
MPT Chat 7B	MPT 7B	Alpaca, ShareGPT, Anthropic HH, HC3, Evol-Instruct	May 2023
Human (DBPedia)	-	-	-

Let’s apply FActScore to LMs in the wild

	Base LM	Data	Release
InstructGPT	?	?	November 2022
ChatGPT	?	?	November 2022
GPT-4	?	?	March 2023
Alpaca {7B, 13B, 65B}	LLAMA	InstuctGPT outputs	March 2023
Vicuna {7B, 13B}	LLAMA	ShareGPT (ChatGPT)	March 2023
Dolly 12B	Pythia 12B	Databricks Dolly (human)	March 2023
Oasst-pythia 12B	Pythia 12B	Open Assistant (human)	March 2023
StableLM-tuned 7B	StableLM-base	Alpaca, Dolly, ShareGPT, GPT4All, Anthropic HH	April 2023
MPT Chat 7B	MPT 7B	Alpaca, ShareGPT, Anthropic HH, HC3, Evol-Instruct	May 2023
Human (DBPedia)	-	-	-

Diverse base LMs

Let’s apply FActScore to LMs in the wild

	Base LM	Data	Release
InstructGPT	?	?	November 2022
ChatGPT	?	?	November 2022
GPT-4	?	?	March 2023
Alpaca {7B, 13B, 65B}	LLAMA	InstuctGPT outputs	March 2023
Vicuna {7B, 13B}	LLAMA	ShareGPT (ChatGPT)	March 2023
Dolly 12B	Pythia 12B	Databricks Dolly (human)	March 2023
Oasst-pythia 12B	Pythia 12B	Open Assistant (human)	March 2023
StableLM-tuned 7B	StableLM-base	Alpaca, Dolly, ShareGPT, GPT4All, Anthropic HH	April 2023
MPT Chat 7B	MPT 7B	Alpaca, ShareGPT, Anthropic HH, HC3, Evol-Instruct	May 2023
Human (DBPedia)	-	-	-

Diverse base LMs **Diverse data**

Let’s apply FActScore to LMs in the wild

	Base LM	Data	Release
InstructGPT	?	?	November 2022
ChatGPT	?	?	November 2022
GPT-4	?	?	March 2023
Alpaca {7B, 13B, 65B}	LLAMA	InstuctGPT outputs	March 2023
Vicuna {7B, 13B}	LLAMA	ShareGPT (ChatGPT)	March 2023
Dolly 12B	Pythia 12B	Databricks Dolly (human)	March 2023
Oasst-pythia 12B	Pythia 12B	Open Assistant (human)	March 2023
StableLM-tuned 7B	StableLM-base	Alpaca, Dolly, ShareGPT, GPT4All, Anthropic HH	April 2023
MPT Chat 7B	MPT 7B	Alpaca, ShareGPT, Anthropic HH, HC3, Evol-Instruct	May 2023
Human (DBPedia)	-	-	-

Diverse base LMs **Diverse data** **All very recent!**

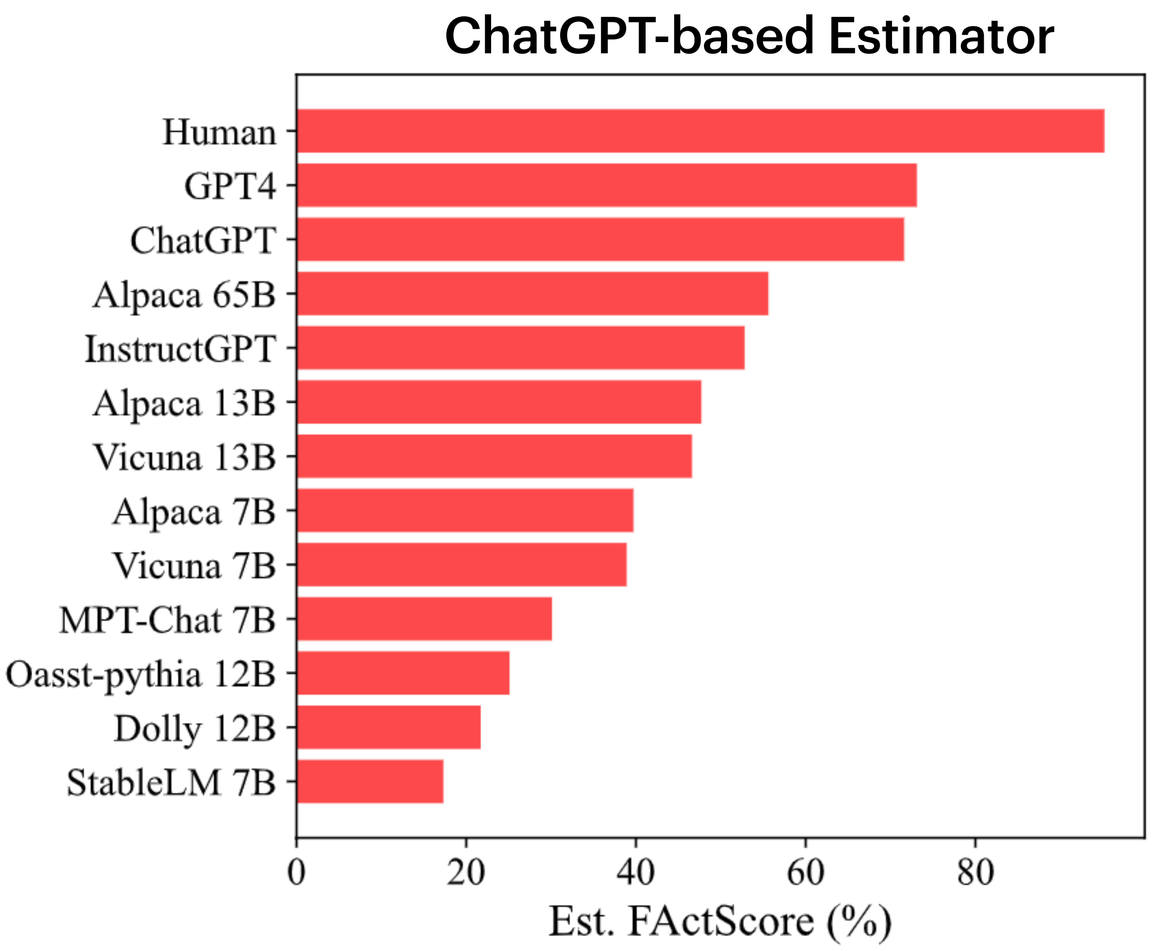
Let's apply FActScore to LMs in the wild

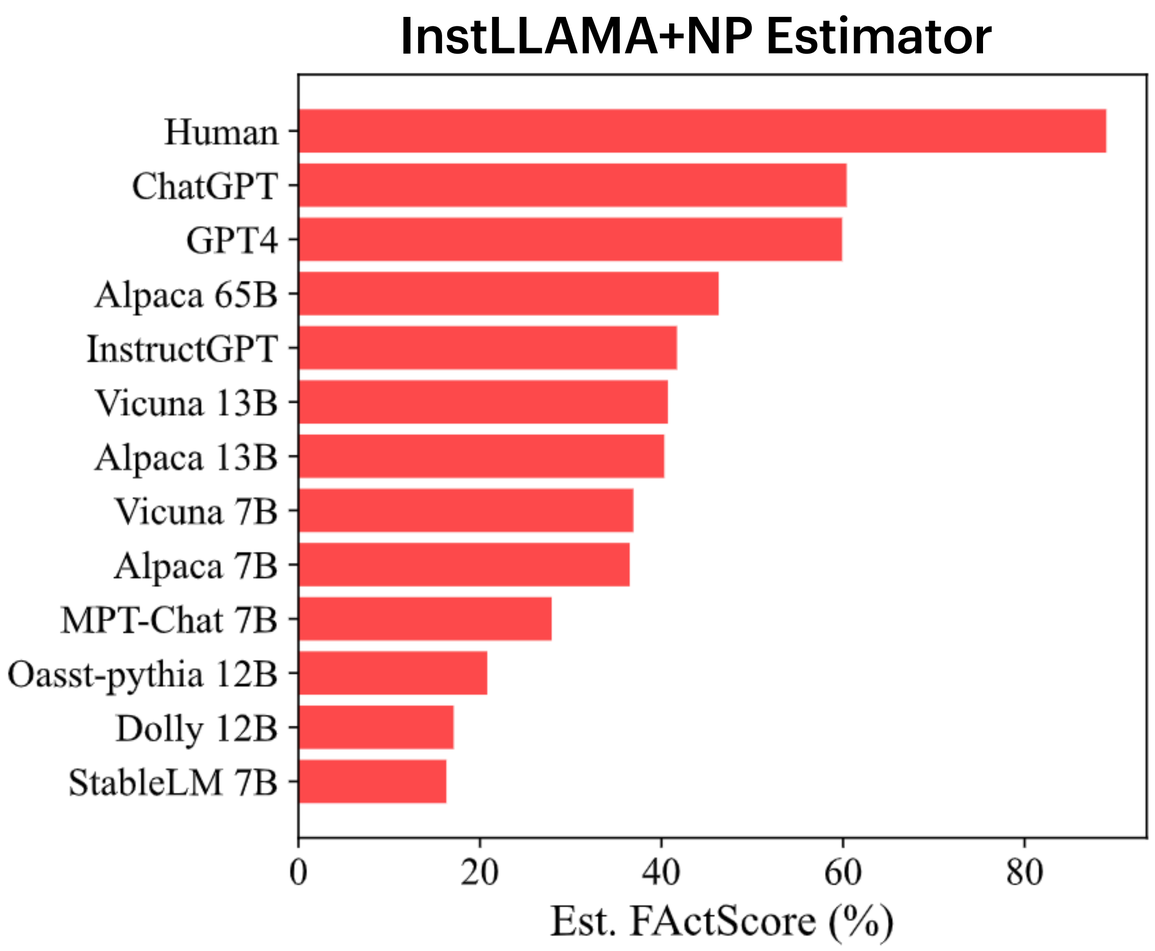
Could have been $65,000

	Base LM	Data	Release
InstructGPT	?	?	November 2022
ChatGPT	?	?	November 2022
GPT-4	?	?	March 2023
Alpaca {7B, 13B, 65B}	LLAMA	InstuctGPT outputs	March 2023
Vicuna {7B, 13B}	LLAMA	ShareGPT (ChatGPT)	March 2023
Dolly 12B	Pythia 12B	Databricks Dolly (human)	March 2023
Oasst-pythia 12B	Pythia 12B	Open Assistant (human)	March 2023
StableLM-tuned 7B	StableLM-base	Alpaca, Dolly, ShareGPT, GPT4All, Anthropic HH	April 2023
MPT Chat 7B	MPT 7B	Alpaca, ShareGPT, Anthropic HH, HC3, Evol-Instruct	May 2023
Human (DBPedia)	-	-	-

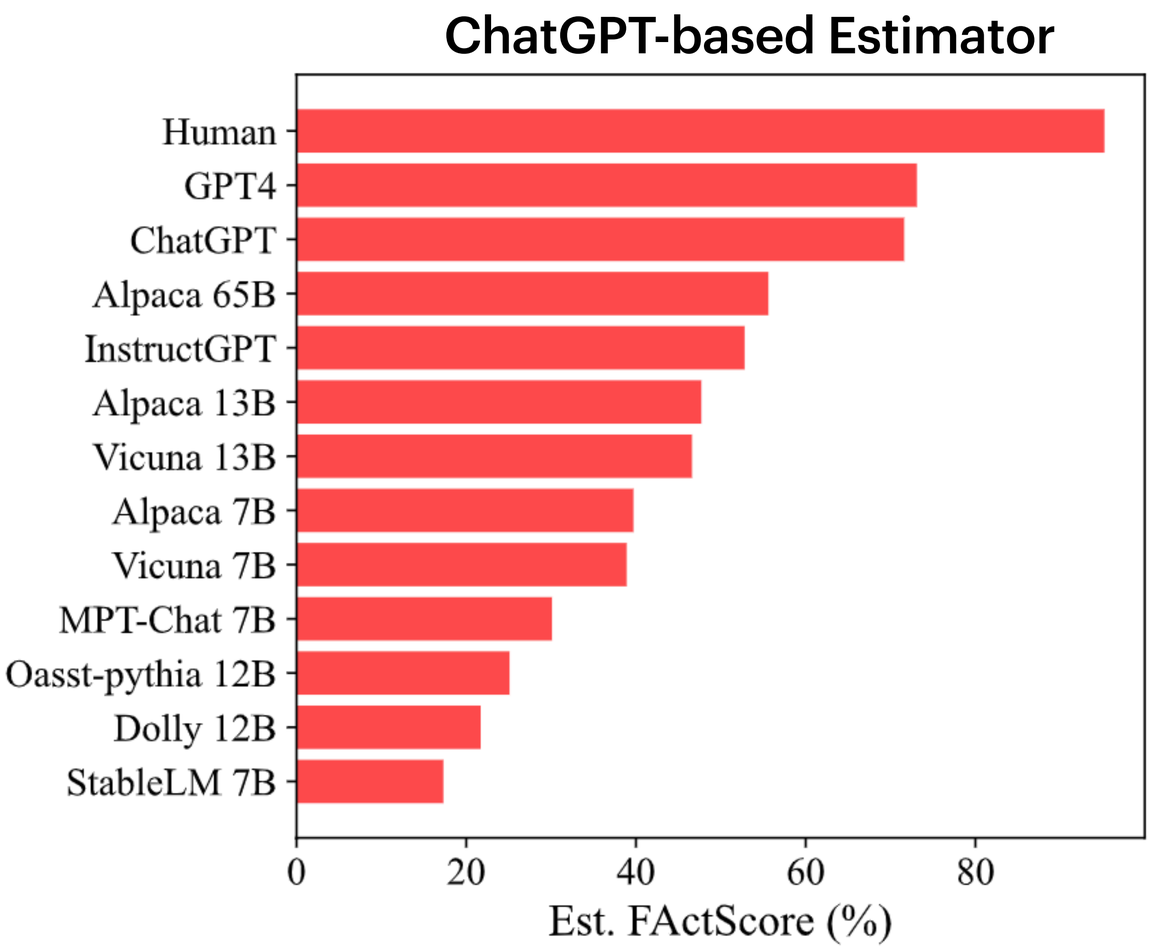
Diverse base LMs **Diverse data** **All very recent!**

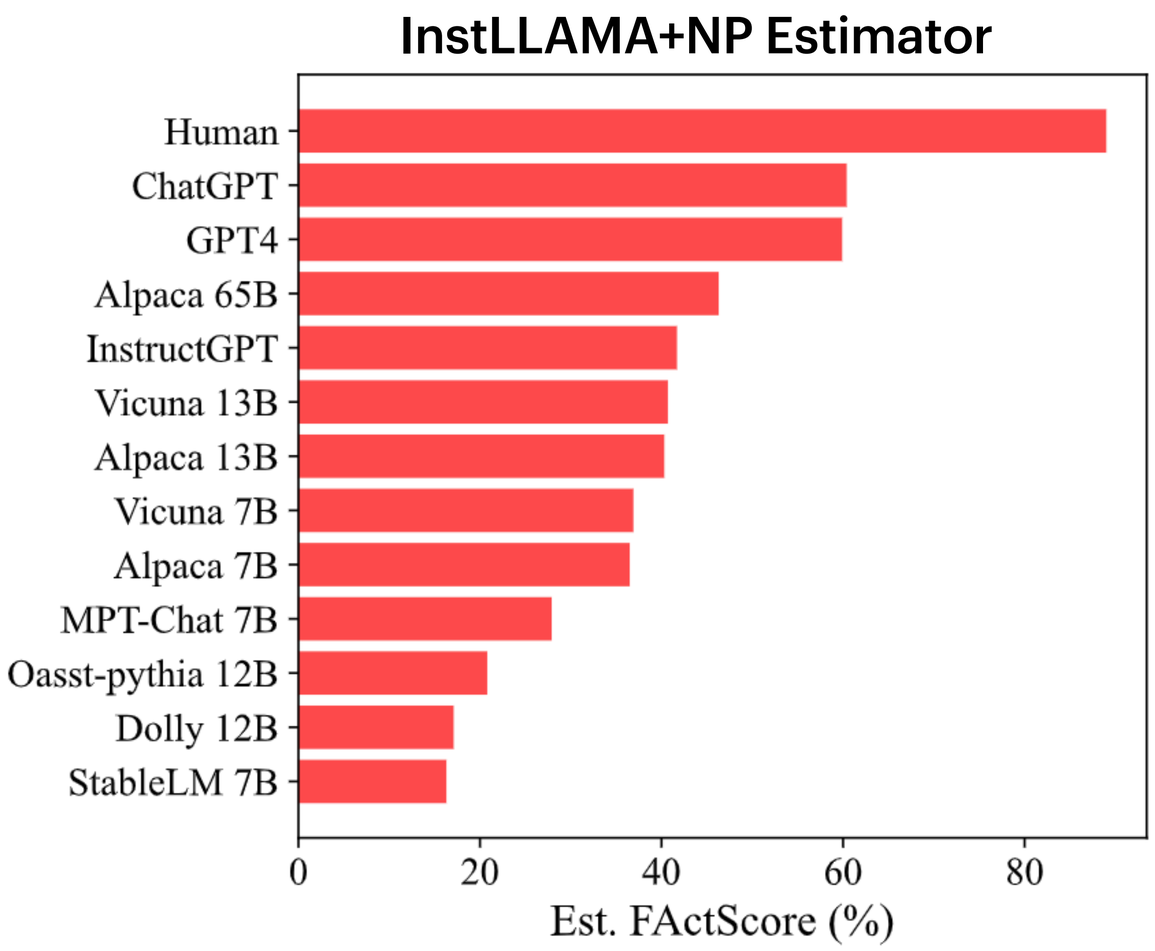
Results


ChatGPT-based Estimator
Human
GPT4
ChatGPT
Alpaca 65B
InstructGPT
Alpaca 13B
Vicuna 13B
Alpaca 7B
Vicuna 7B
MPT-Chat 7B
Oasst-pythia 12B
Dolly 12B
StableLM 7B
0
20
40
60
80
Est. FActScore (%)


InstLLAMA+NP Estimator
Human
ChatGPT
GPT4
Alpaca 65B
InstructGPT
Vicuna 13B
Alpaca 13B
Vicuna 7B
Alpaca 7B
MPT-Chat 7B
Oasst-pythia 12B
Dolly 12B
StableLM 7B
0
20
40
60
80
Est. FActScore (%)

Results

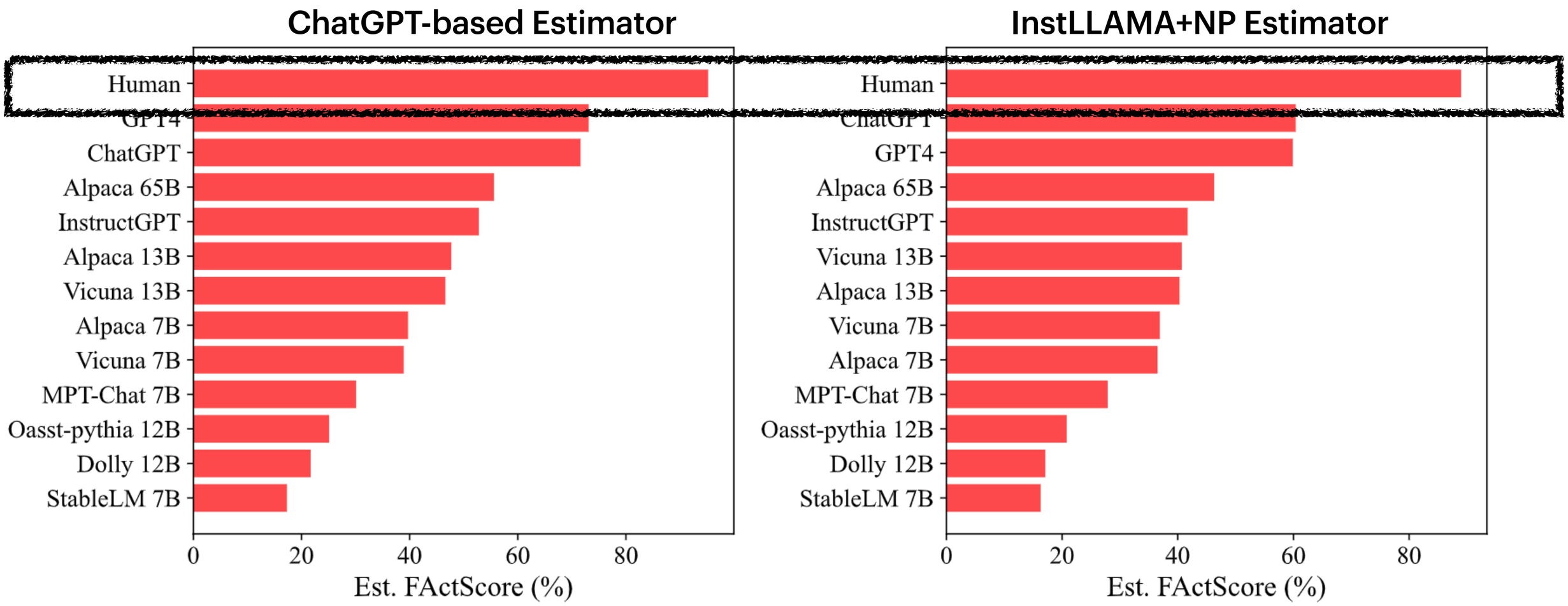




Pearson's r = 0.99

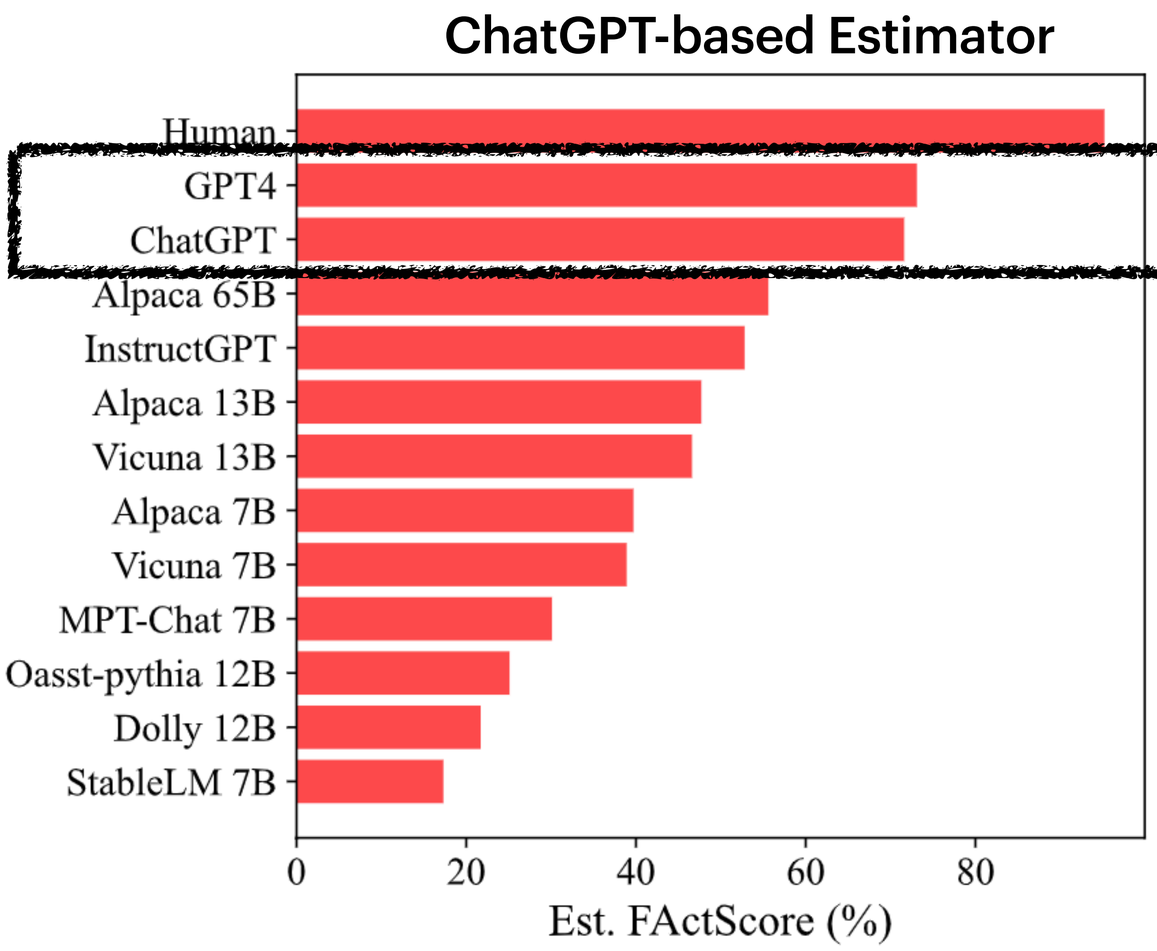
Results

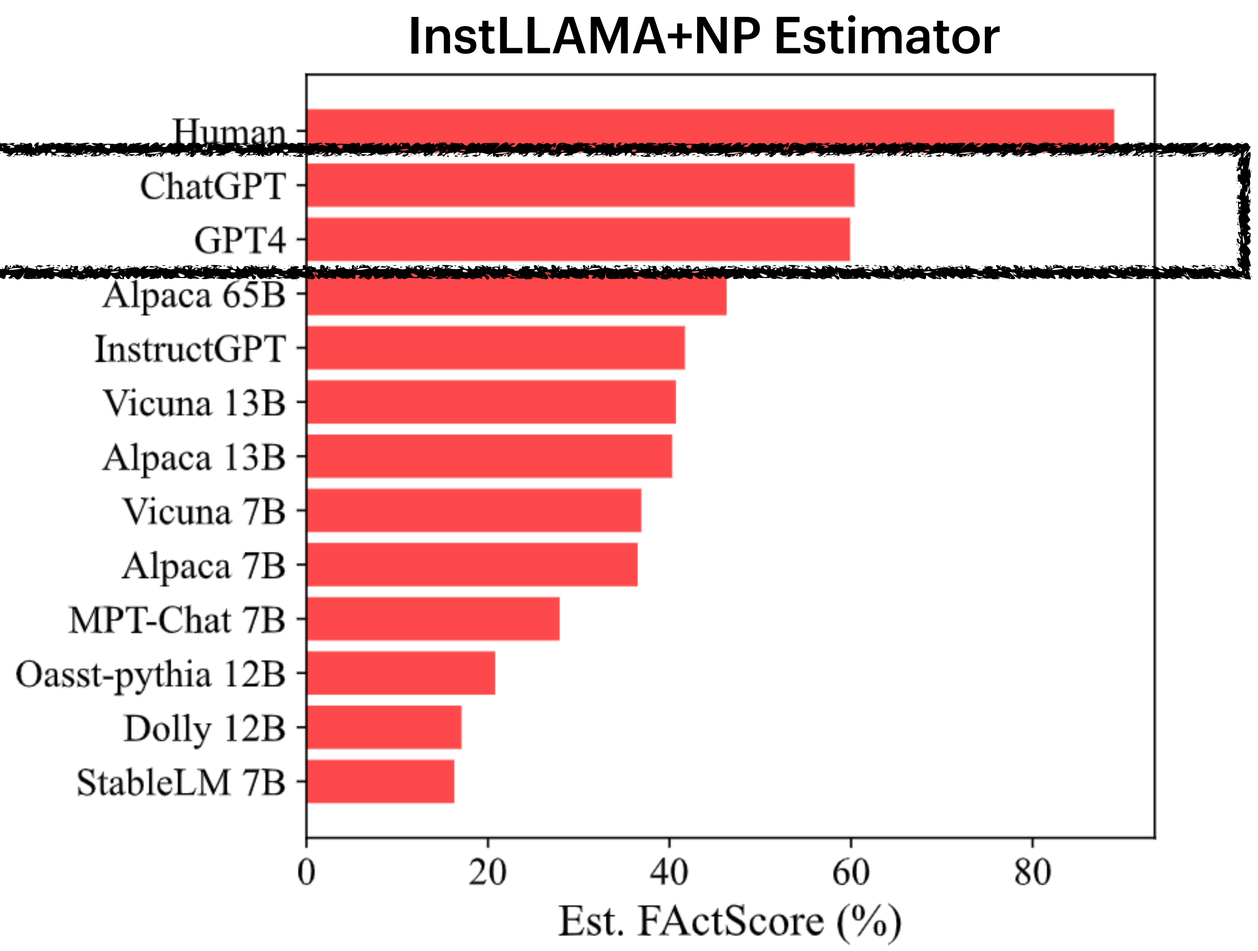
1. All models substantially lag behind human



Results

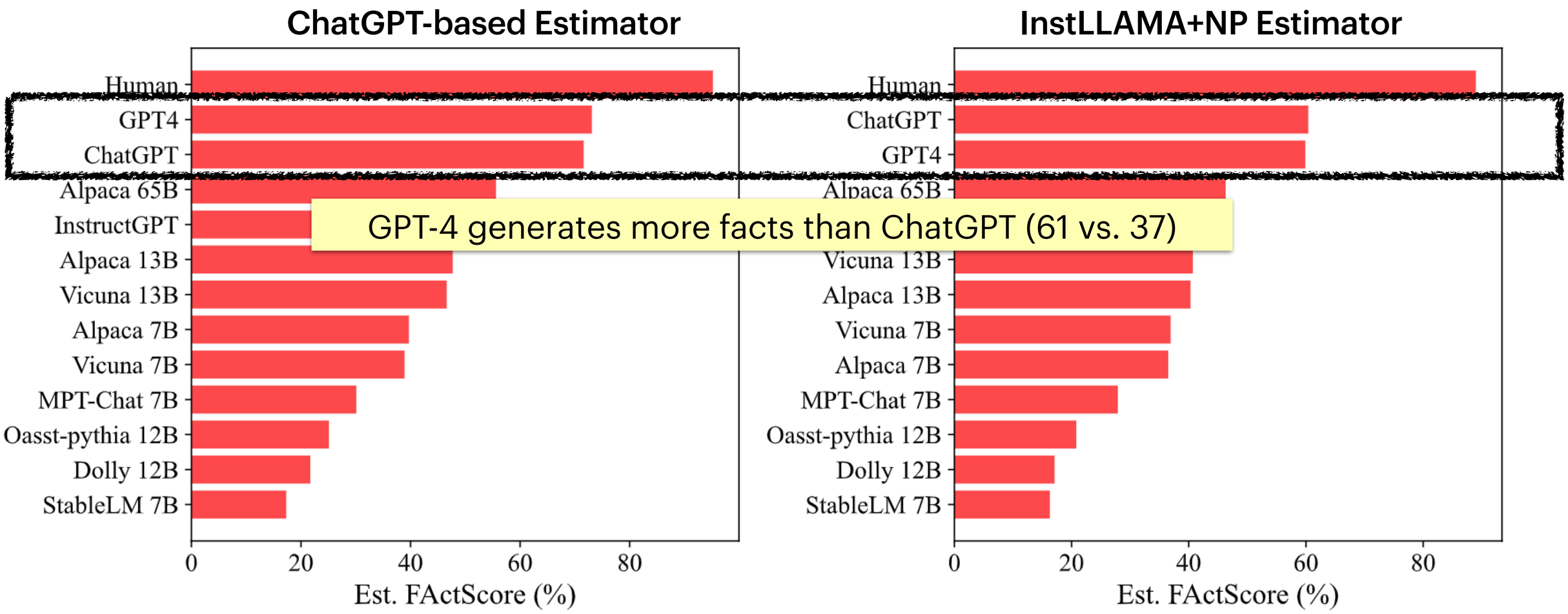
2. GPT-4 & ChatGPT are much better than public LMs





Results

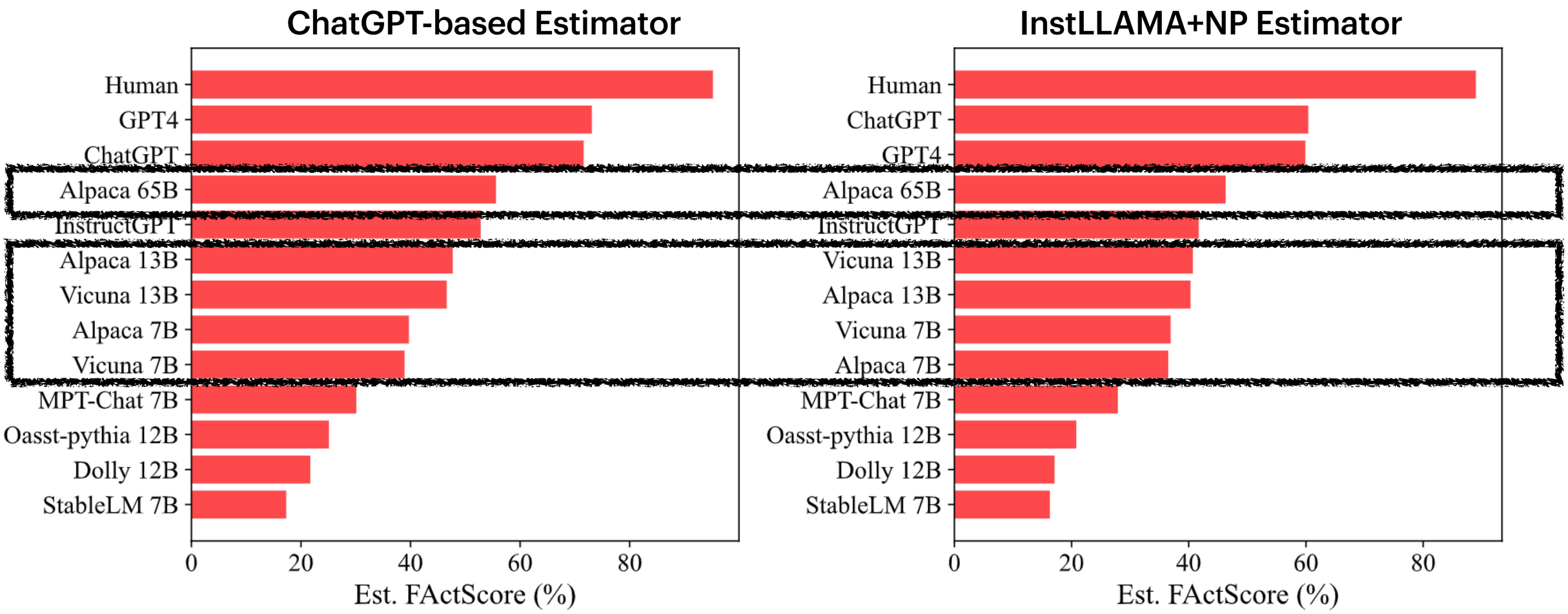
2. GPT-4 & ChatGPT are much better than public LMs



GPT-4 generates more facts than ChatGPT (61 vs. 37)

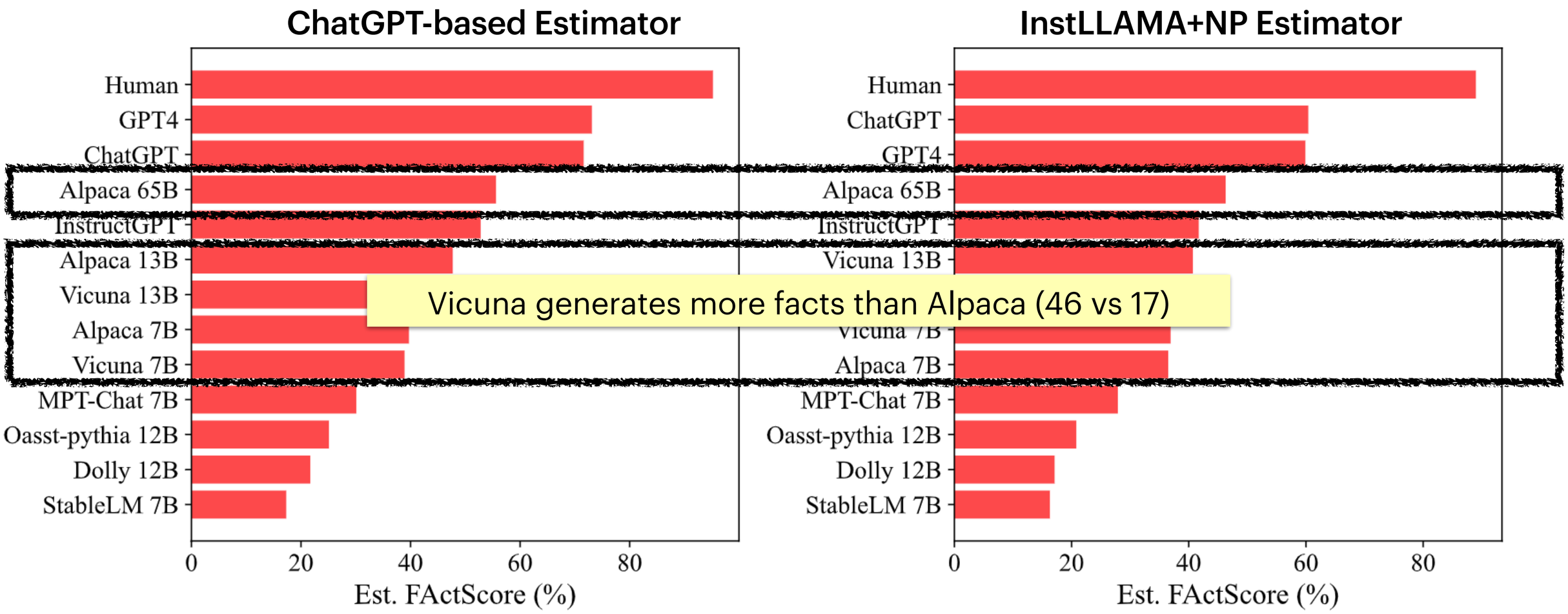
Results

3. Alpaca & Vicuna come the next



Results

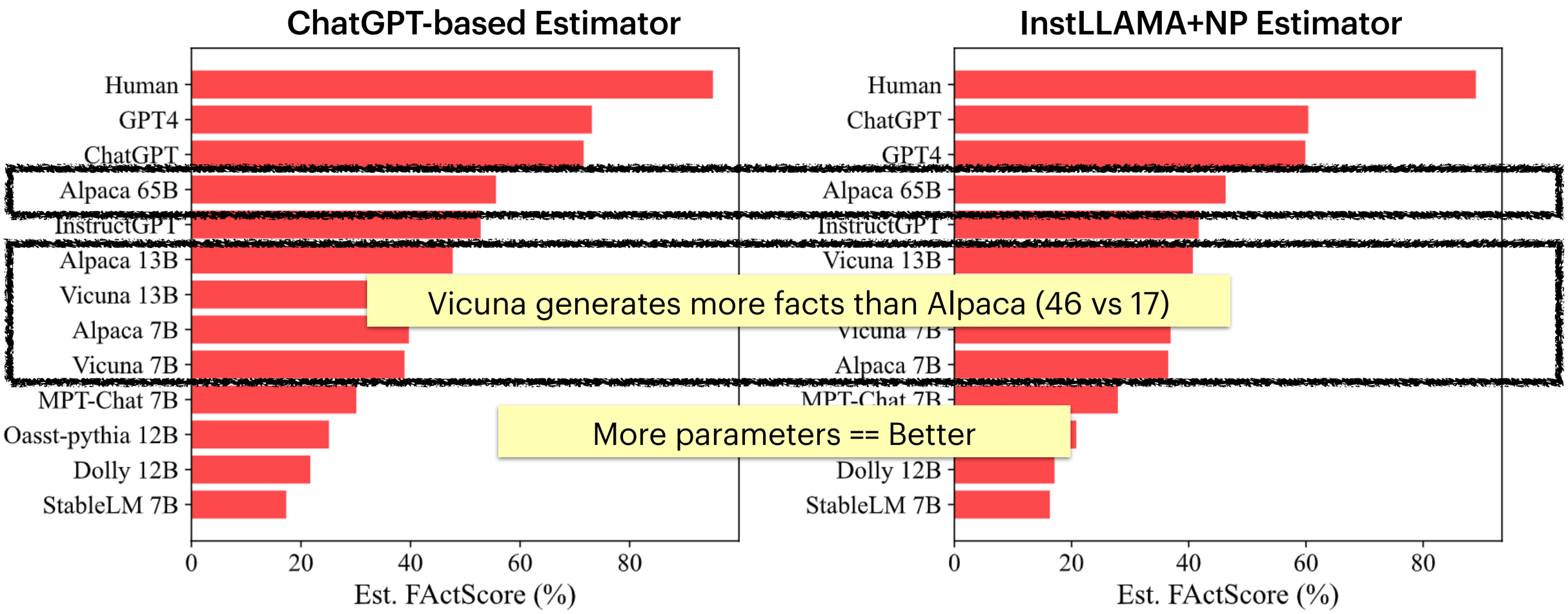
3. Alpaca & Vicuna come the next



Vicuna generates more facts than Alpaca (46 vs 17)

Results

3. Alpaca & Vicuna come the next

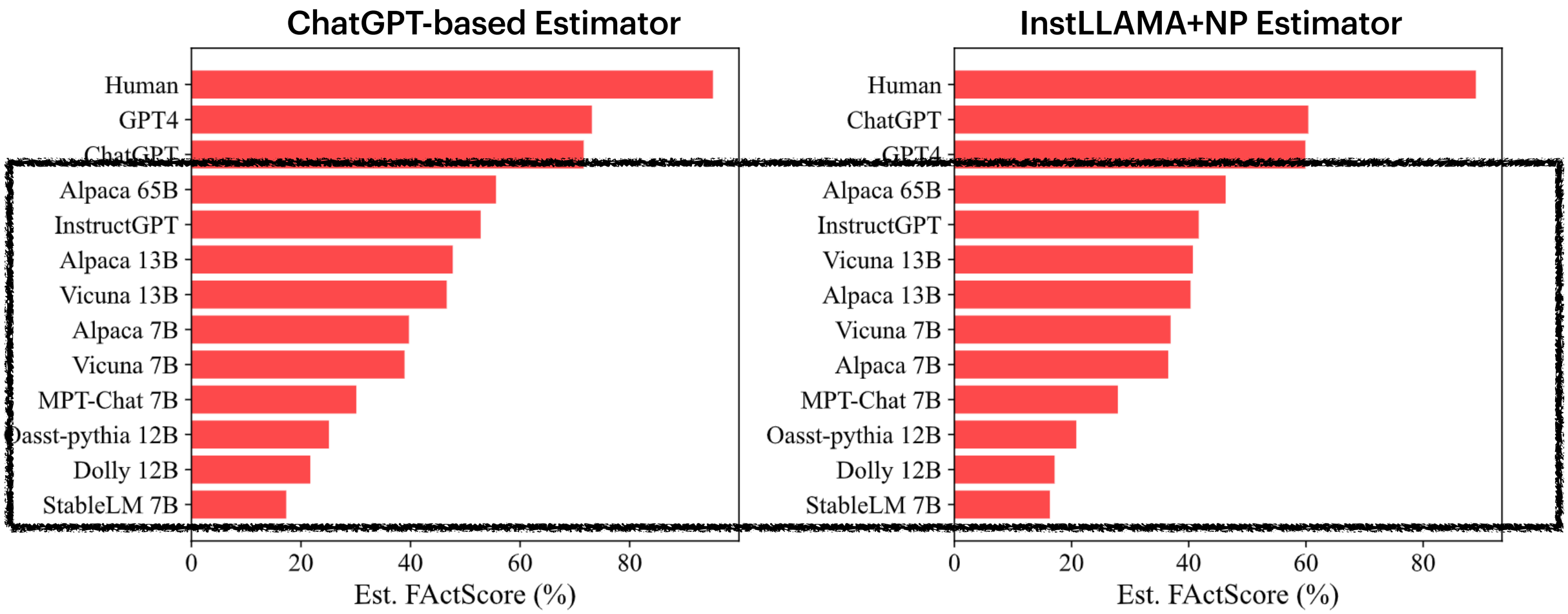


Vicuna generates more facts than Alpaca (46 vs 17)

More parameters == Better

Results

4. Large variance between public LMs



So far, FActScore only considers *Precision*

So far, FActScore only considers *Precision*

Bridget Moynahan is an American actress.

Bridget Moynahan is an American actress and model. She is best known for her roles in I, Robot and Blue Bloods. She studied acting at the Caymichael Patten Studio, and made her TV debut in 1999. (...) She married businessman Andrew Frankel in 2015.

So far, FActScore only considers *Precision*

Bridget Moynahan is an American actress.

2/2

Bridget Moynahan is an American actress and model. She is best known for her roles in I, Robot and Blue Bloods. She studied acting at the Caymichael Patten Studio, and made her TV debut in 1999. (...) She married businessman Andrew Frankel in 2015.

So far, FActScore only considers *Precision*

Bridget Moynahan is an American actress.

2/2

Bridget Moynahan is an American actress and model. She is best known for her roles in I, Robot and Blue Bloods. She studied acting at the Caymichael Patten Studio, and made her TV debut in 1999. (...) She married businessman Andrew Frankel in 2015.

19/20

So far, FActScore only considers *Precision*

Bridget Moynahan is an American actress. → 2/2

Bridget Moynahan is an American actress and model. She is best known for her roles in I, Robot and Blue Bloods. She studied acting at the Caymichael Patten Studio, and made her TV debut in 1999. (...) She married businessman Andrew Frankel in 2015. → 19/20

A single number that considers both FActScore and # of facts?

So far, FActScore only considers *Precision*

Bridget Moynahan is an American actress. → 2/2

Bridget Moynahan is an American actress and model. She is best known for her roles in I, Robot and Blue Bloods. She studied acting at the Caymichael Patten Studio, and made her TV debut in 1999. (...) She married businessman Andrew Frankel in 2015. → 19/20

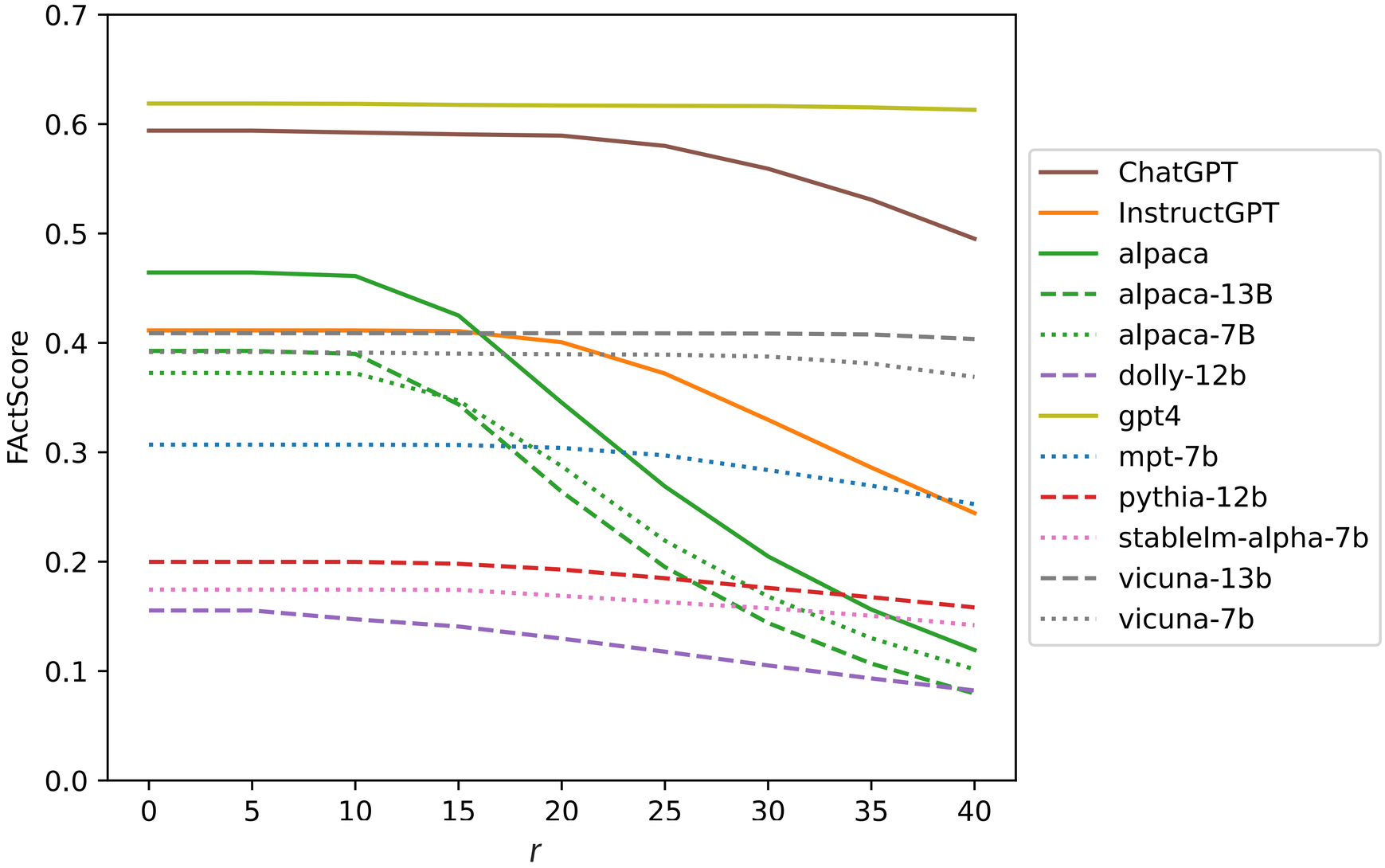
A single number that considers both FActScore and # of facts?

# of facts in model generation

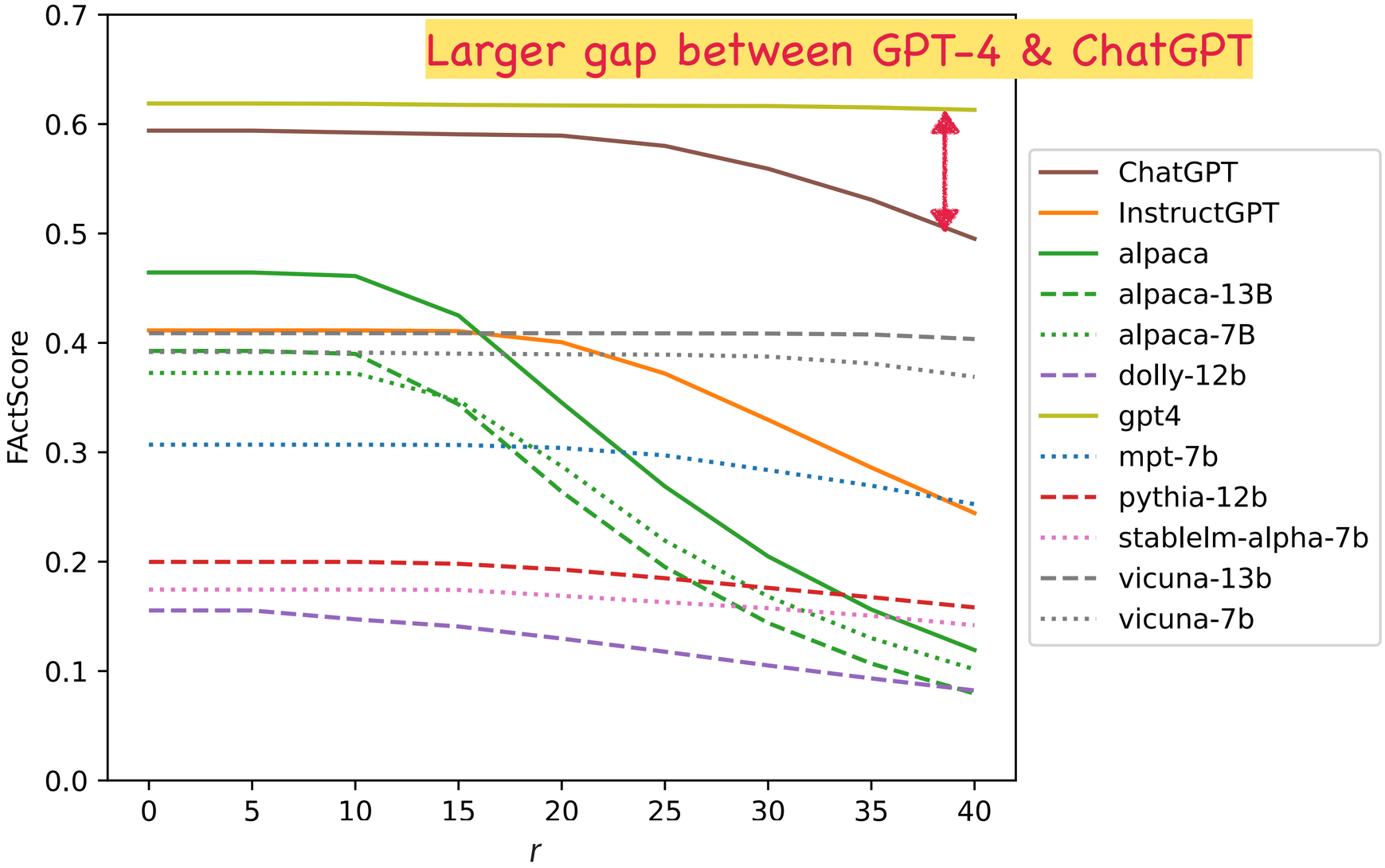
$$\mathrm{bp}(c, r) = 1 \text{ if } c < r \text{ else } \mathrm{e}^{1-\frac{r}{c}}$$

reference # of facts (hyperparam)

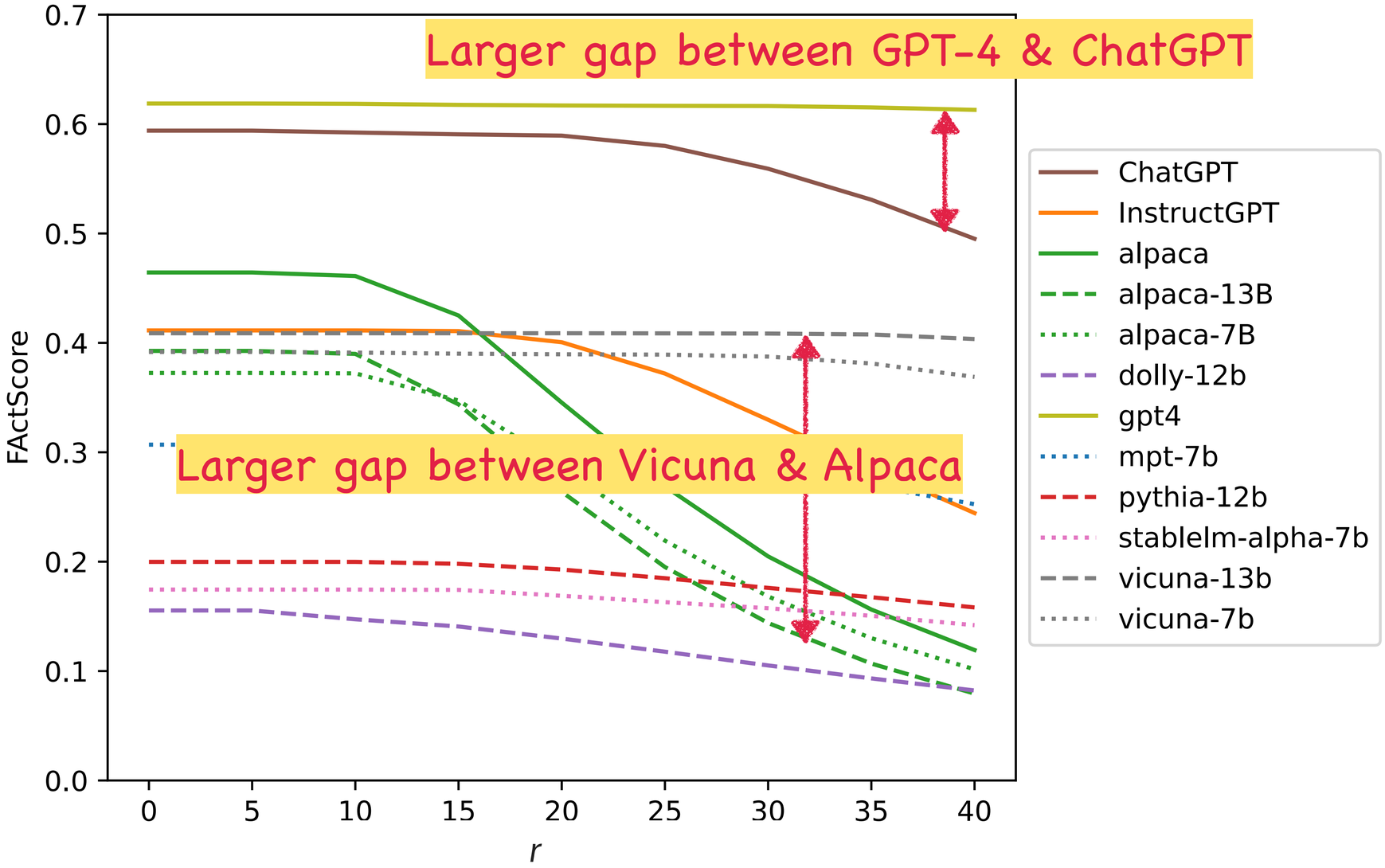
Adjusted FActScore



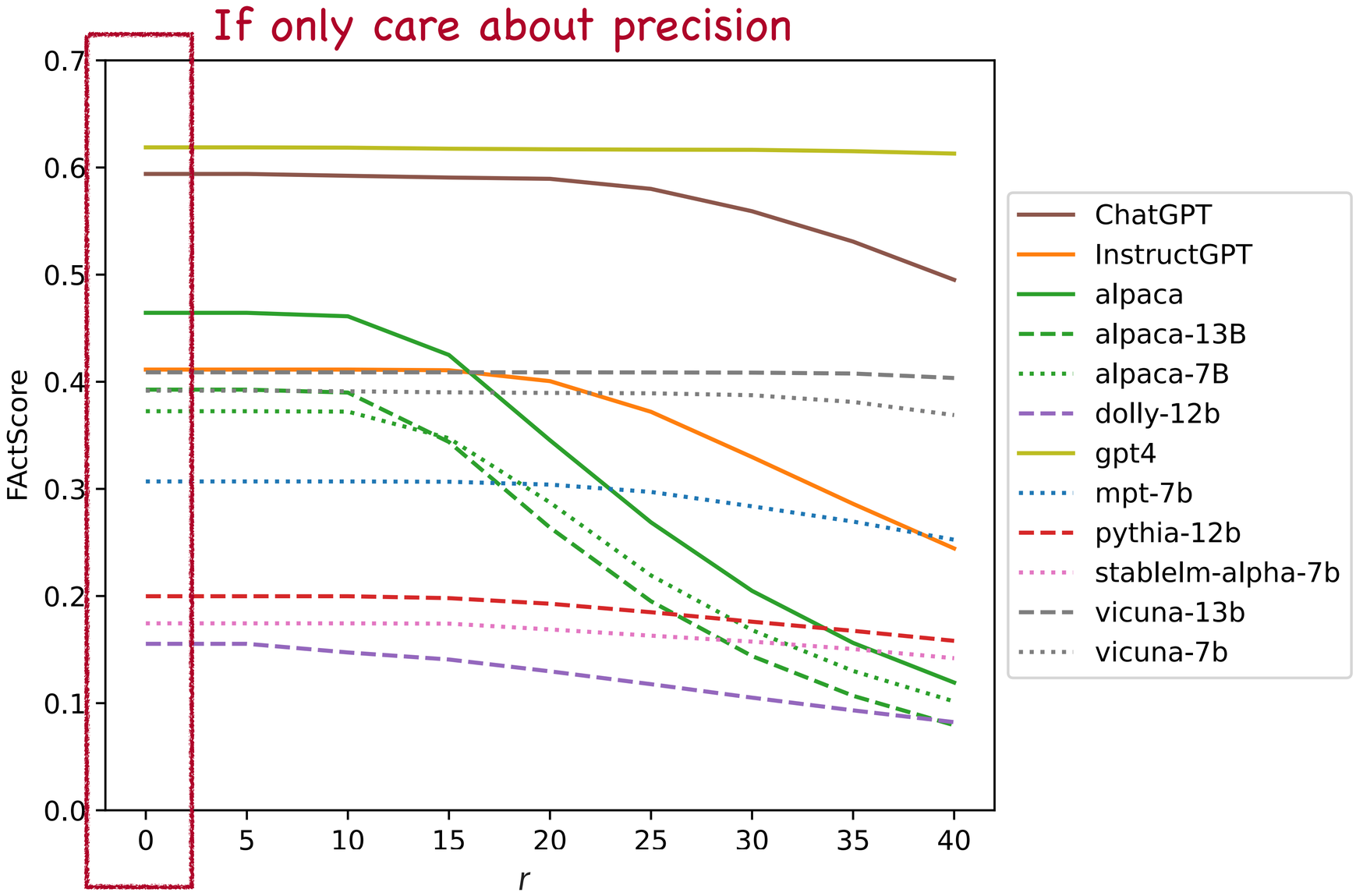
Adjusted FActScore



Adjusted FActScore

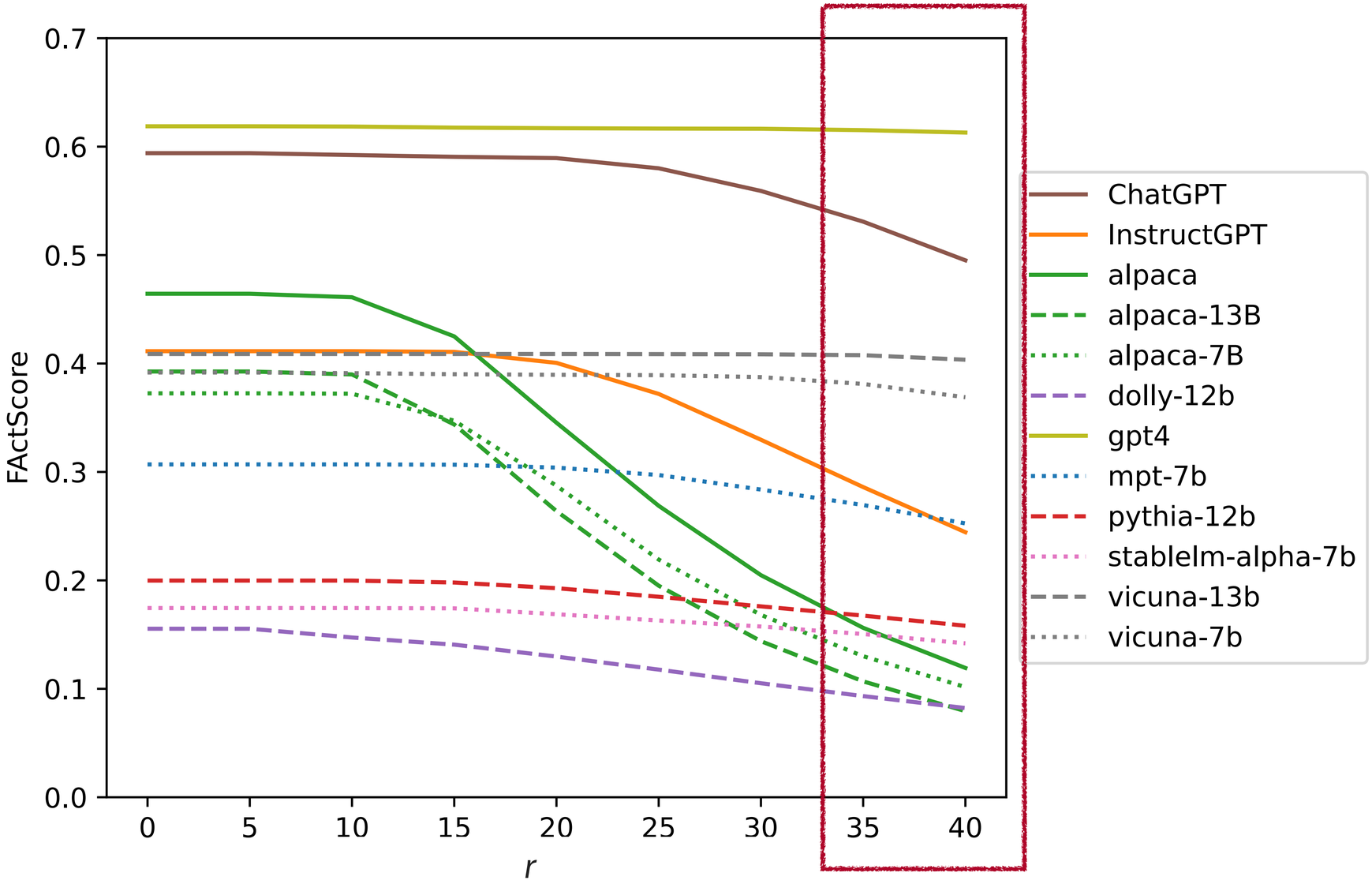


Adjusted FActScore

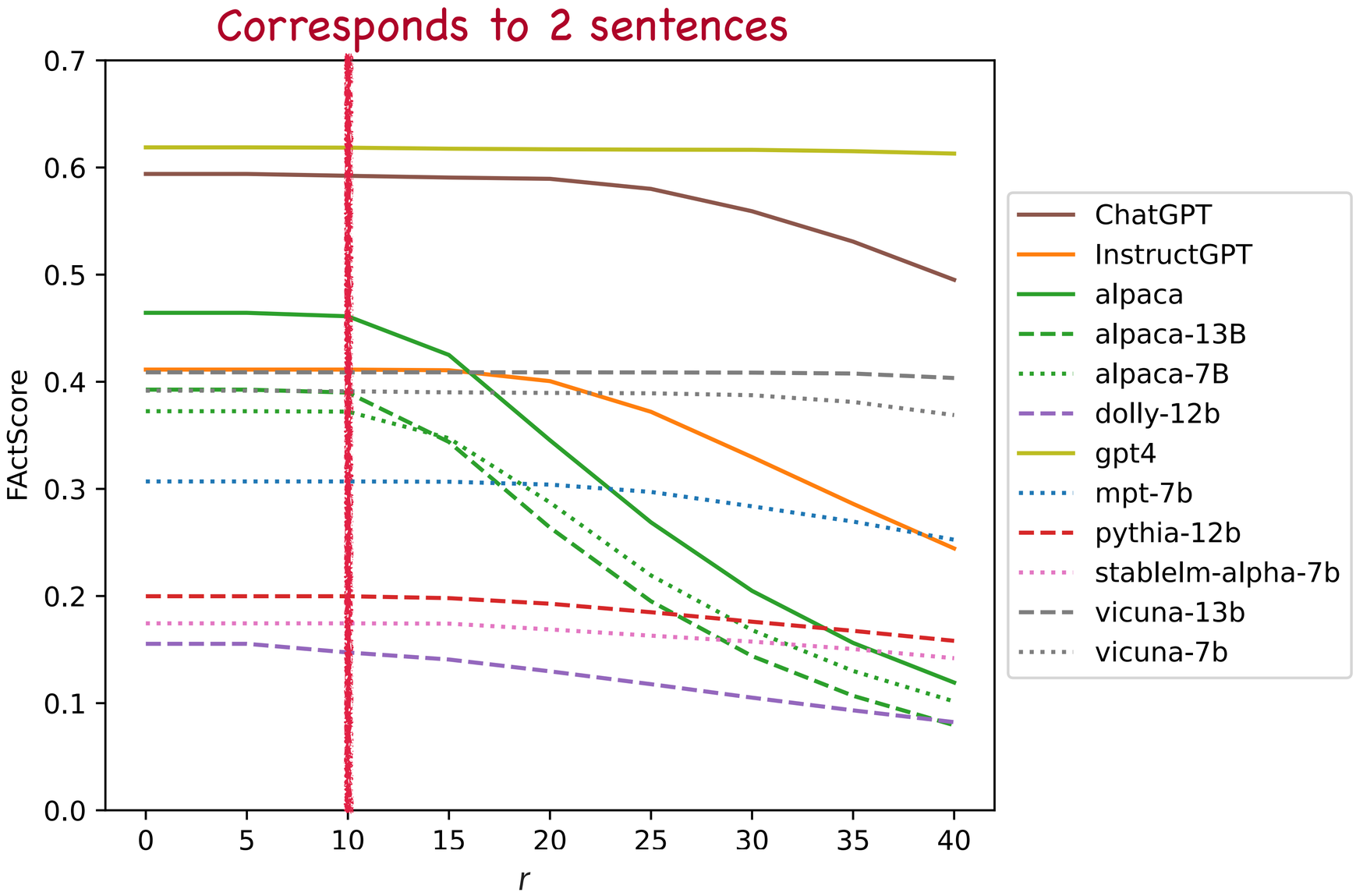


Adjusted FActScore

If long generation is "hard" constraint

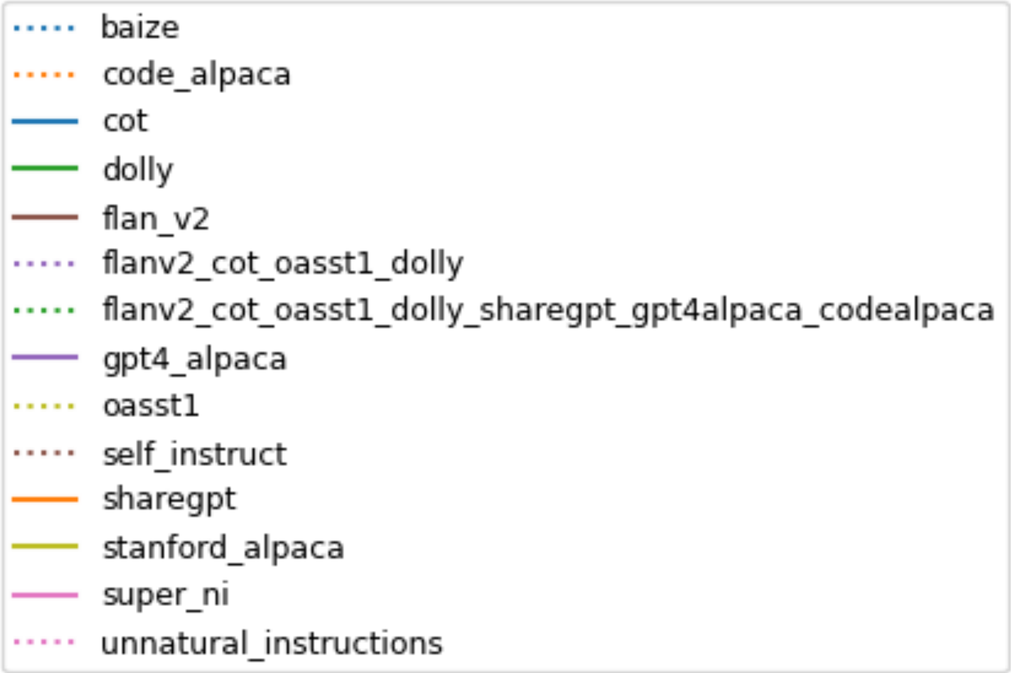


Adjusted FActScore



Adjusted FActScore

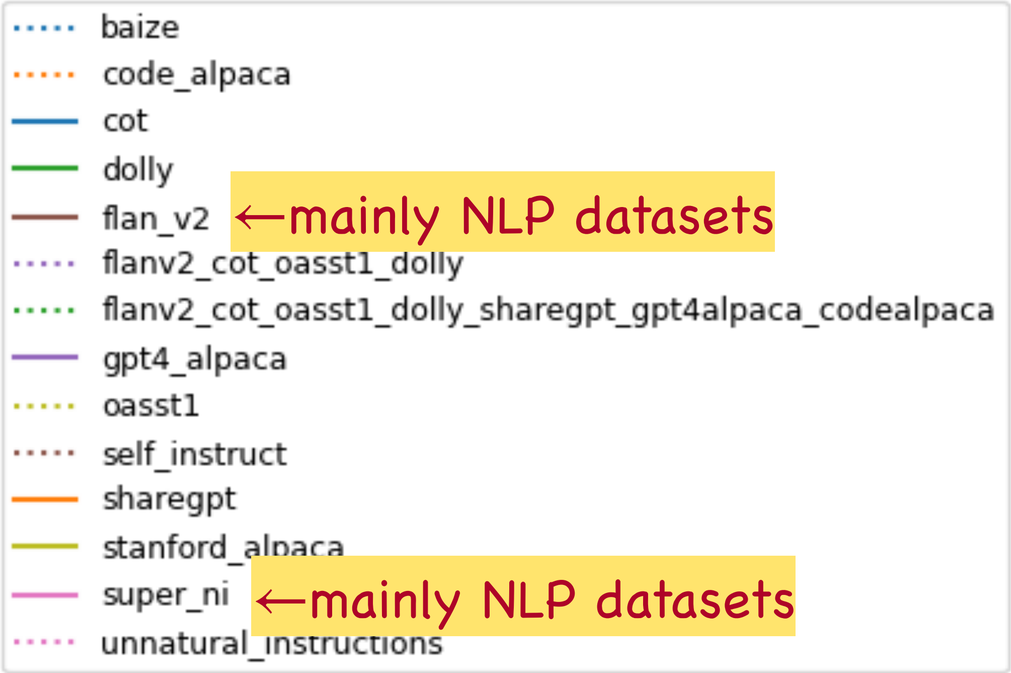
When the training data is more diverse?



Wang et al. 2023. “How Far Can Camels Go? Exploring the State of Instruction Tuning on Open Resources”

Adjusted FActScore

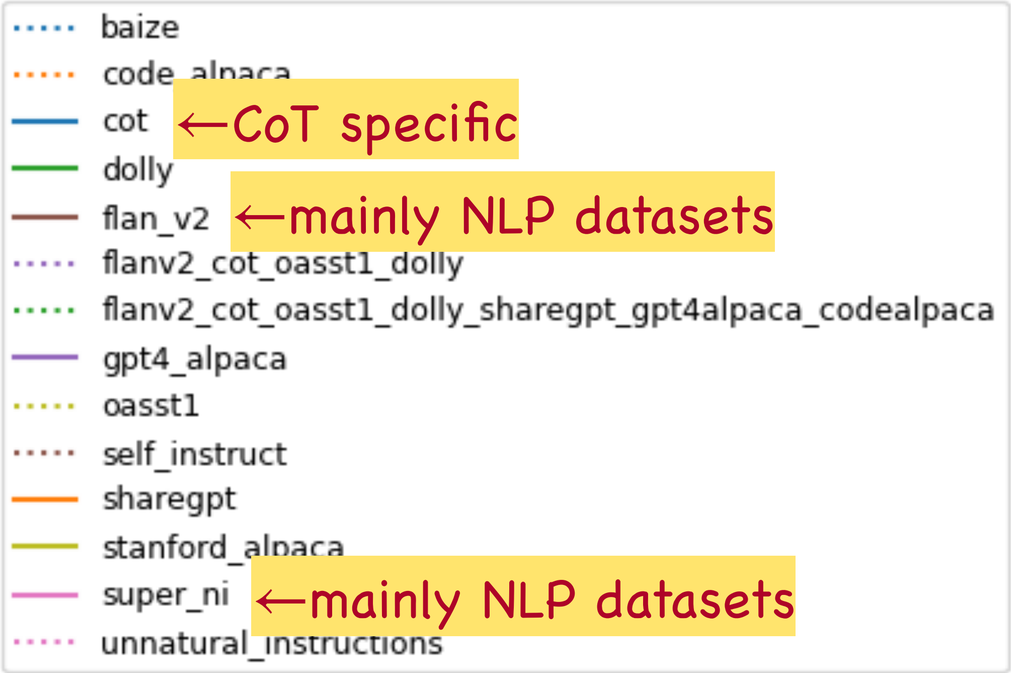
When the training data is more diverse?



Wang et al. 2023. "How Far Can Camels Go? Exploring the State of Instruction Tuning on Open Resources"

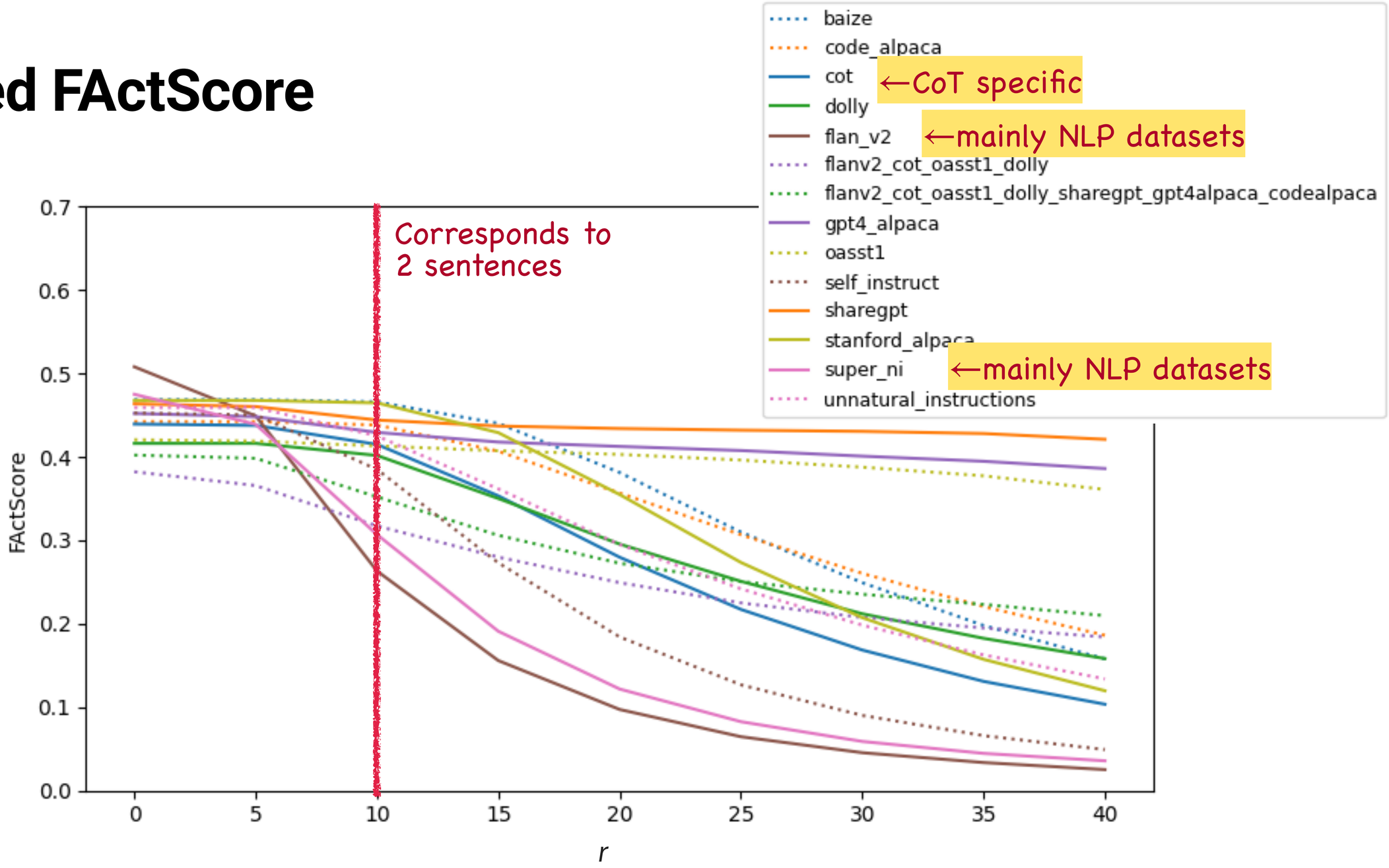
Adjusted FActScore

When the training data is more diverse?



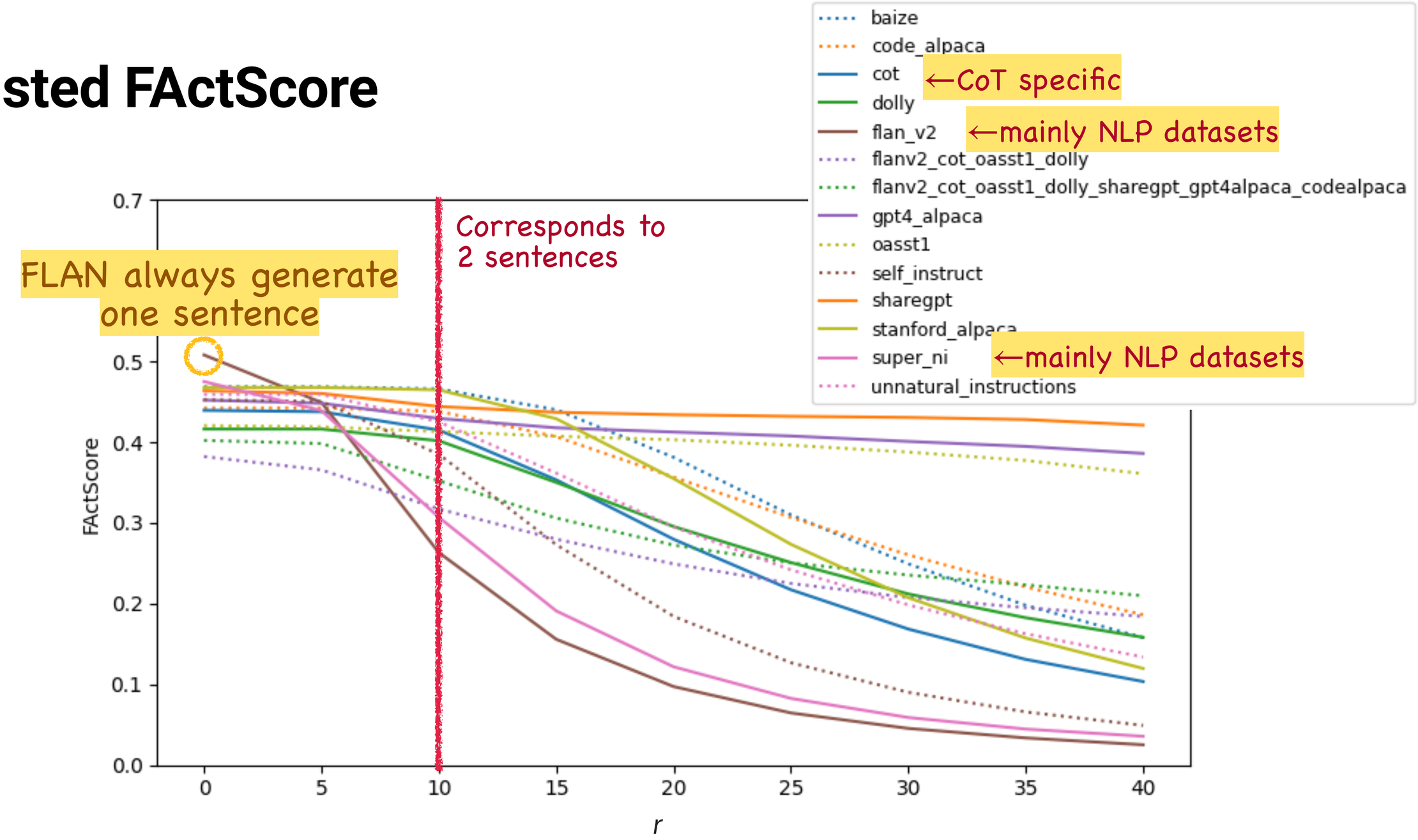
Wang et al. 2023. "How Far Can Camels Go? Exploring the State of Instruction Tuning on Open Resources"

Adjusted FActScore



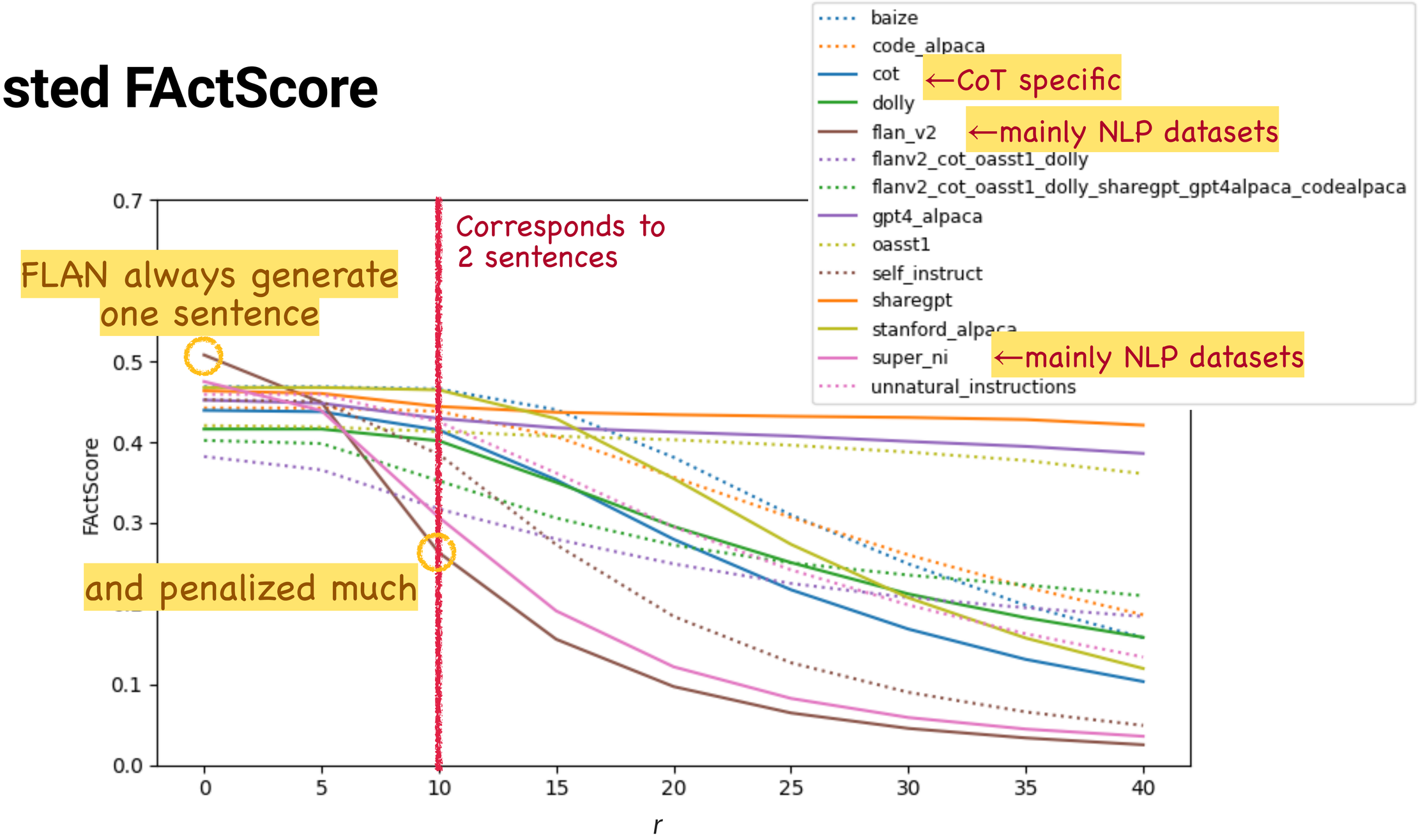
Wang et al. 2023. "How Far Can Camels Go? Exploring the State of Instruction Tuning on Open Resources"

Adjusted FActScore



Wang et al. 2023. "How Far Can Camels Go? Exploring the State of Instruction Tuning on Open Resources"

Adjusted FActScore



Wang et al. 2023. "How Far Can Camels Go? Exploring the State of Instruction Tuning on Open Resources"

FActScore in PIP!

```
pip install factscore
```

```python
from factscore.factscorer import FactScorer

fs = FactScorer(openai_key="...")

scoreA = fs.get_score(topics, modelA_generations) # 0.667
scoreB = fs.get_score(topics, modelB_generations) # 0.010
```

https://github.com/shmsw25/FActScore

We appreciate feedback!

Summary

Summary

- **FActScore**: a new evaluation of an LM that breaks a generation into atomic facts and validates each against a given knowledge source.

Summary

- **FActScore**: a new evaluation of an LM that breaks a generation into atomic facts and validates each against a given knowledge source.
- Human evaluation: the state-of-the-art LMs with & without search have low FActScores.

Summary

- **FActScore**: a new evaluation of an LM that breaks a generation into atomic facts and validates each against a given knowledge source.
- Human evaluation: the state-of-the-art LMs with & without search have low FActScores.
- We can approximate FActScore with < 2% error rate, allowing evaluation of a large set of new LMs without manual human efforts.

Summary

- **FActScore**: a new evaluation of an LM that breaks a generation into atomic facts and validates each against a given knowledge source.
- Human evaluation: the state-of-the-art LMs with & without search have low FActScores.
- We can approximate FActScore with < 2% error rate, allowing evaluation of a large set of new LMs without manual human efforts.
- Future work: next slide!

Future directions?

Future directions?

Extension beyond biographies

Evaluate recall

Incorporate to RLHF

1. Beyond biographies

Assumption: Every atomic fact is verifiable, specific, and undebatable

1. Beyond biographies

Assumption: Every atomic fact is verifiable, specific, and undebatable

Input: Tell me about Seattle.

Seattle is a vibrant city located in the Pacific Northwest region of the United States. Known for its stunning natural surroundings, thriving tech industry, and rich cultural scene, Seattle has a distinct character that attracts residents and visitors alike.

1. Beyond biographies

Assumption: Every atomic fact is verifiable, specific, and undebatable

Input: Tell me about Seattle.

Seattle is a vibrant city located in the Pacific Northwest region of the United States. Known for its stunning natural surroundings, thriving tech industry, and rich cultural scene, Seattle has a distinct character that attracts residents and visitors alike.

1. Beyond biographies

Assumption: Every atomic fact is verifiable, specific, and undebatable

Seattle has thriving tech industry.

Seattle

Article Talk

From Wikipedia, the free encyclopedia

Logging was Seattle's first major industry, but by the late 19th century, the city had become a commercial and shipbuilding center as a gateway to Alaska during the Klondike Gold Rush. Growth after World War II was partially due to the local Boeing company, which established Seattle as a center for aircraft manufacturing. The Seattle area developed into a technology center from the 1980s onwards with companies like Microsoft becoming established in the region; Microsoft founder Bill Gates is a Seattleite by birth. Internet retailer Amazon was founded in Seattle in 1994, and major airline Alaska Airlines is based in SeaTac, Washington, serving Seattle's international airport, Seattle–Tacoma International Airport. The stream of new software, biotechnology, and Internet companies led to an economic revival, which increased the city's population by almost 50,000 between 1990 and 2000.

1. Beyond biographies

Assumption: Every atomic fact is verifiable, specific, and undebatable

Seattle has thriving tech industry.

Seattle

Article Talk

From Wikipedia, the free encyclopedia

Logging was Seattle's first major industry, but by the late 19th century, the city had become a commercial and shipbuilding center as a gateway to Alaska during the Klondike Gold Rush. Growth after World War II was partially due to the local Boeing company, which established Seattle as a center for aircraft manufacturing. The Seattle area developed into a technology center from the 1980s onwards with companies like Microsoft becoming established in the region; Microsoft founder Bill Gates is a Seattleite by birth. Internet retailer Amazon was founded in Seattle in 1994, and major airline Alaska Airlines is based in SeaTac, Washington, serving Seattle's international airport, Seattle–

Bios are good because they are relatively free from these issues, but still extension beyond bios is important!

2. Evaluating recall

Evaluation & models that consider **recall**

Tell me a bio of Mary I of England.

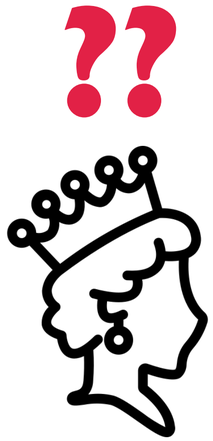
Mary I (February 18, 1516 — November 17, 1558) was Queen of England and Ireland from July 1553 until her death. Mary was the daughter of King Henry VIII and his first wife, Catherine of Aragon. After her father annulled his marriage to Catherine, Mary's status was in doubt, and she was excluded from the line of succession. Her half-brother, Edward VI, succeeded their father in 1547, and Mary had to live with the humiliation of being declared illegitimate.

2. Evaluating recall

Evaluation & models that consider **recall**

Tell me a bio of Mary I of England.

Mary I (February 18, 1516 — November 17, 1558) was Queen of England and Ireland from July 1553 until her death. Mary was the daughter of King Henry VIII and his first wife, Catherine of Aragon. After her father annulled his marriage to Catherine, Mary's status was in doubt, and she was excluded from the line of succession. Her half-brother, Edward VI, succeeded their father in 1547, and Mary had to live with the humiliation of being declared illegitimate.



3. Incorporation to RLHF

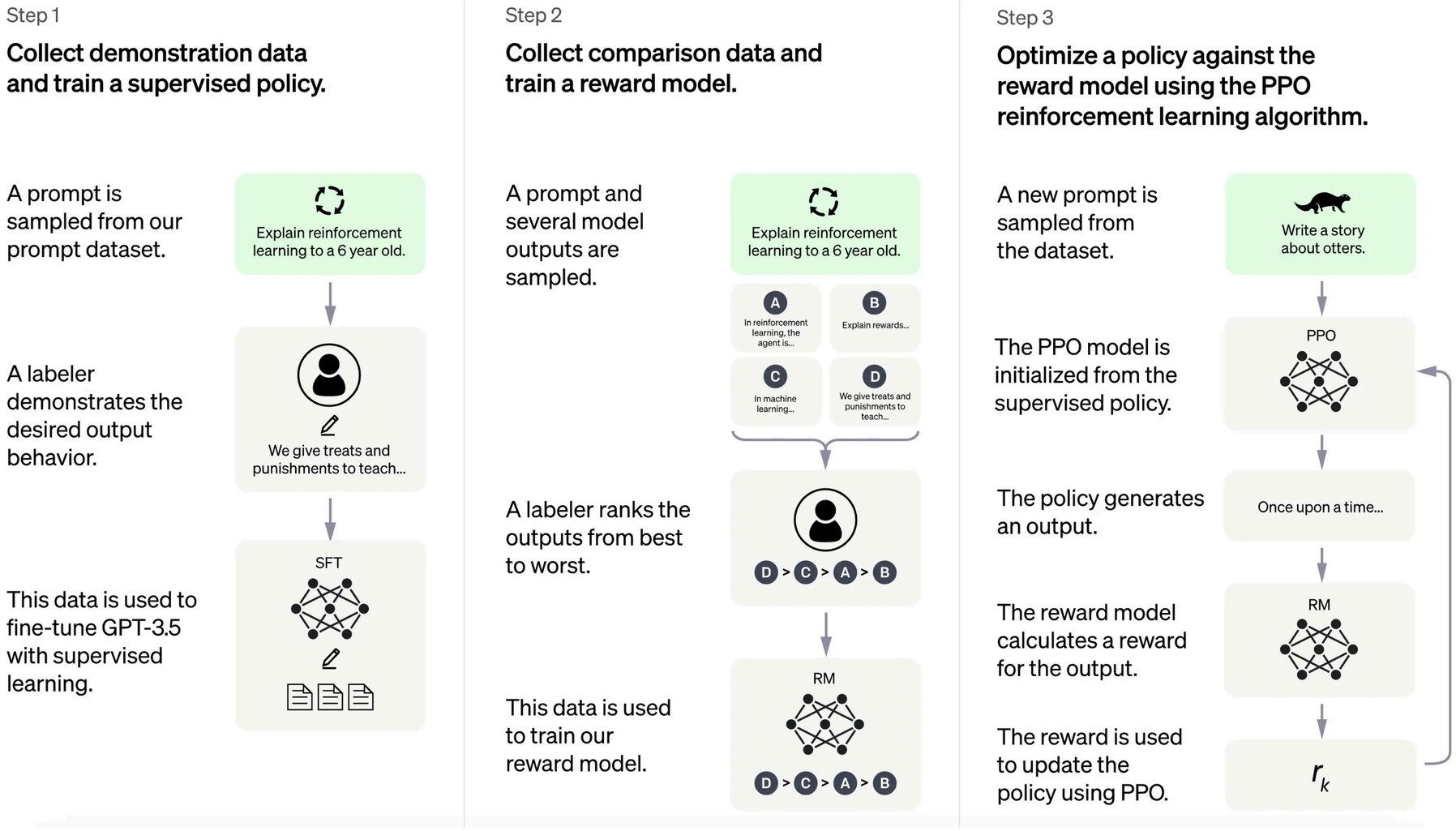


Figure from OpenAI

3. Incorporation to RLHF

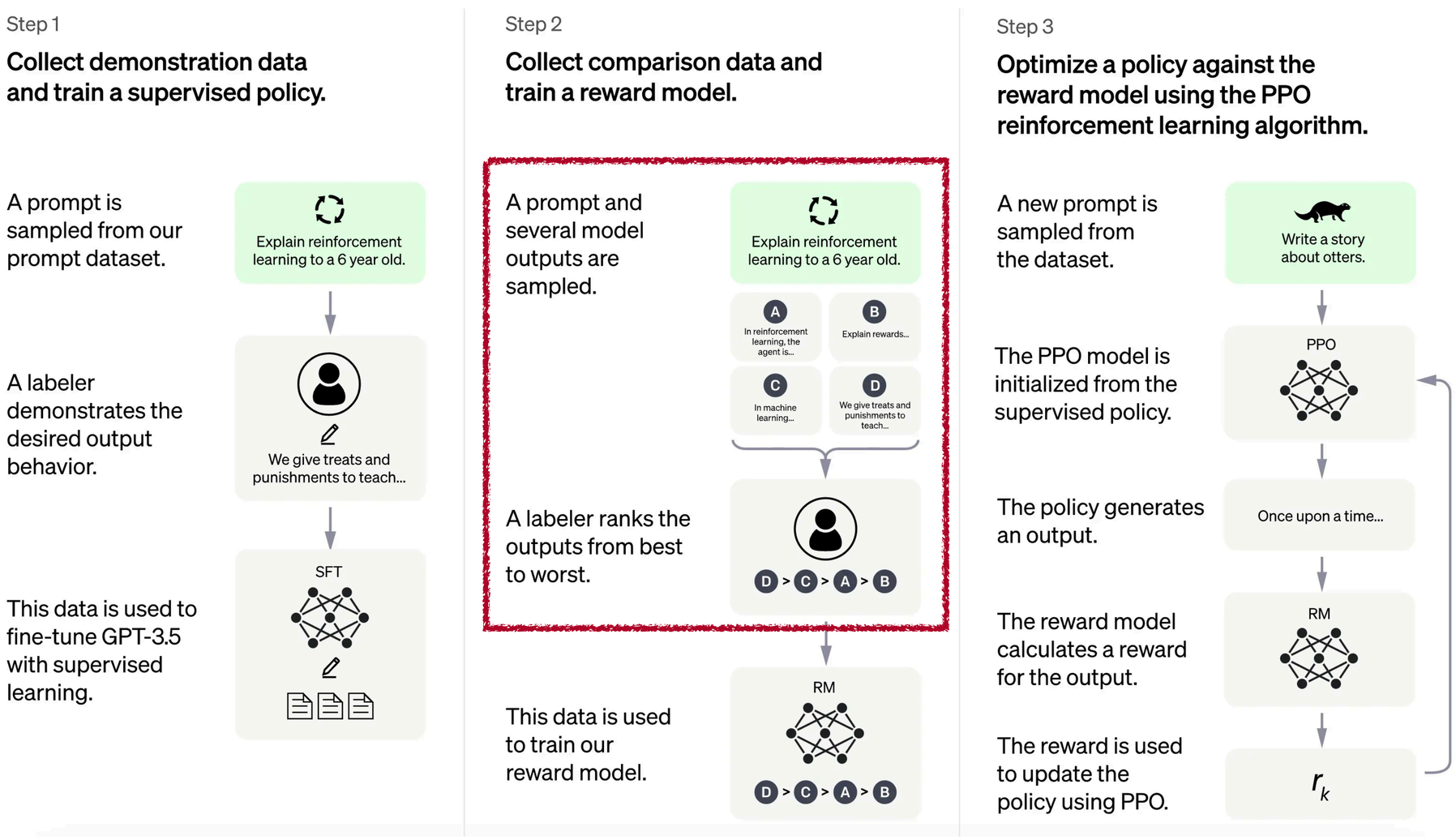
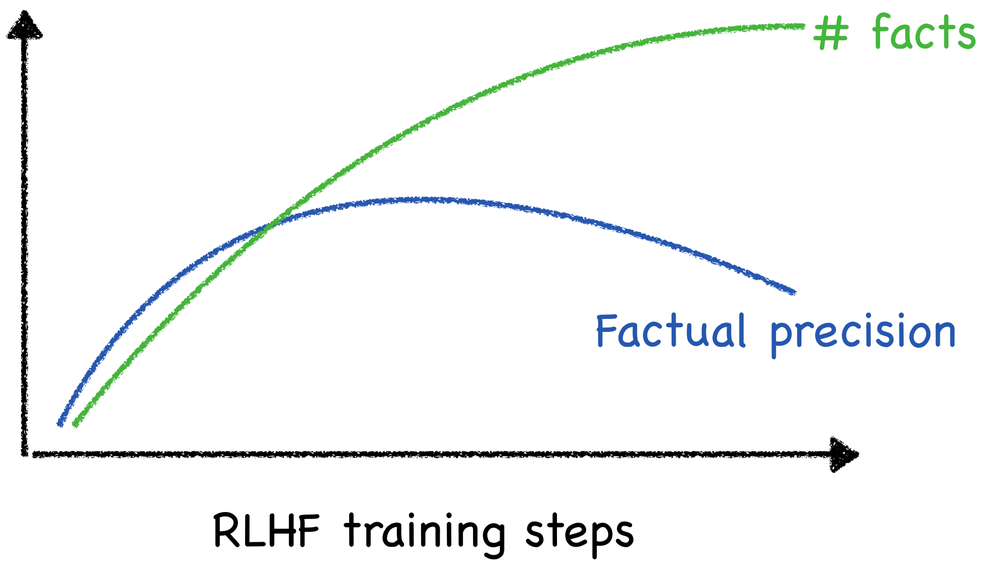


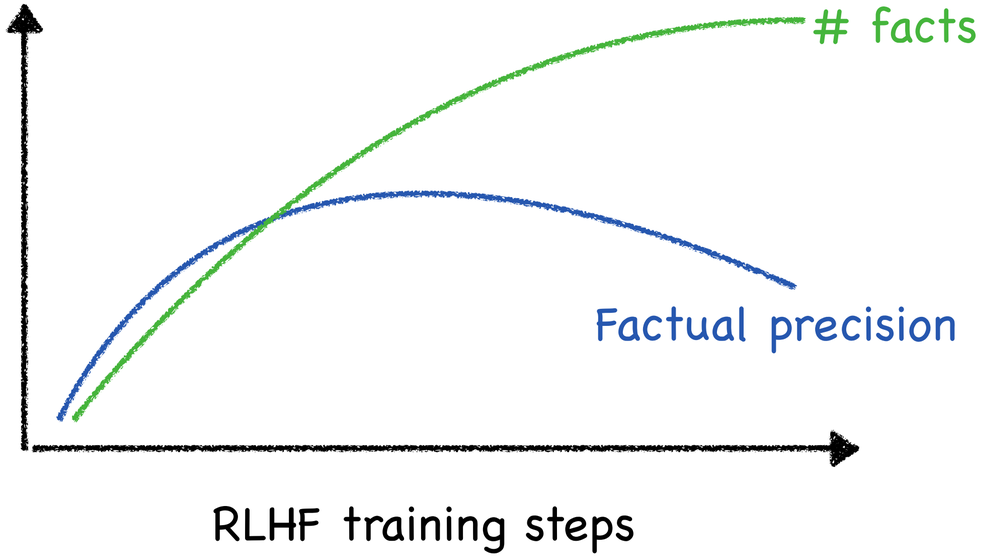
Figure from OpenAI

3. Incorporation to RLHF

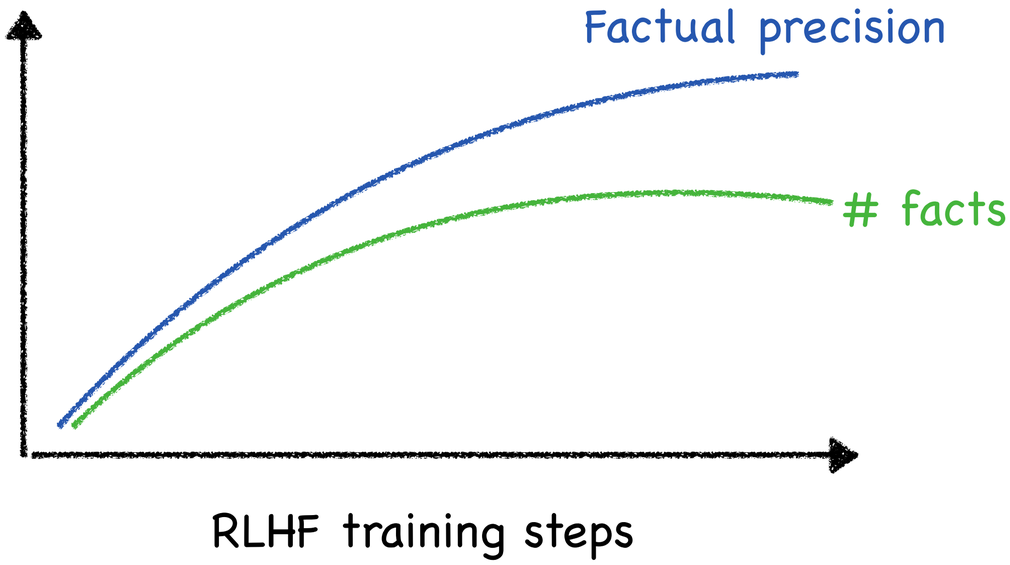


Failure mode

3. Incorporation to RLHF



Failure mode



Ideal

Thank you for listening

https://arxiv.org/abs/2305.14251

https://github.com/shmsw25/FActScore

@sewon__min & @kalpeshk2011

Anonymous feedback: tinyurl.com/sewon-min-talk