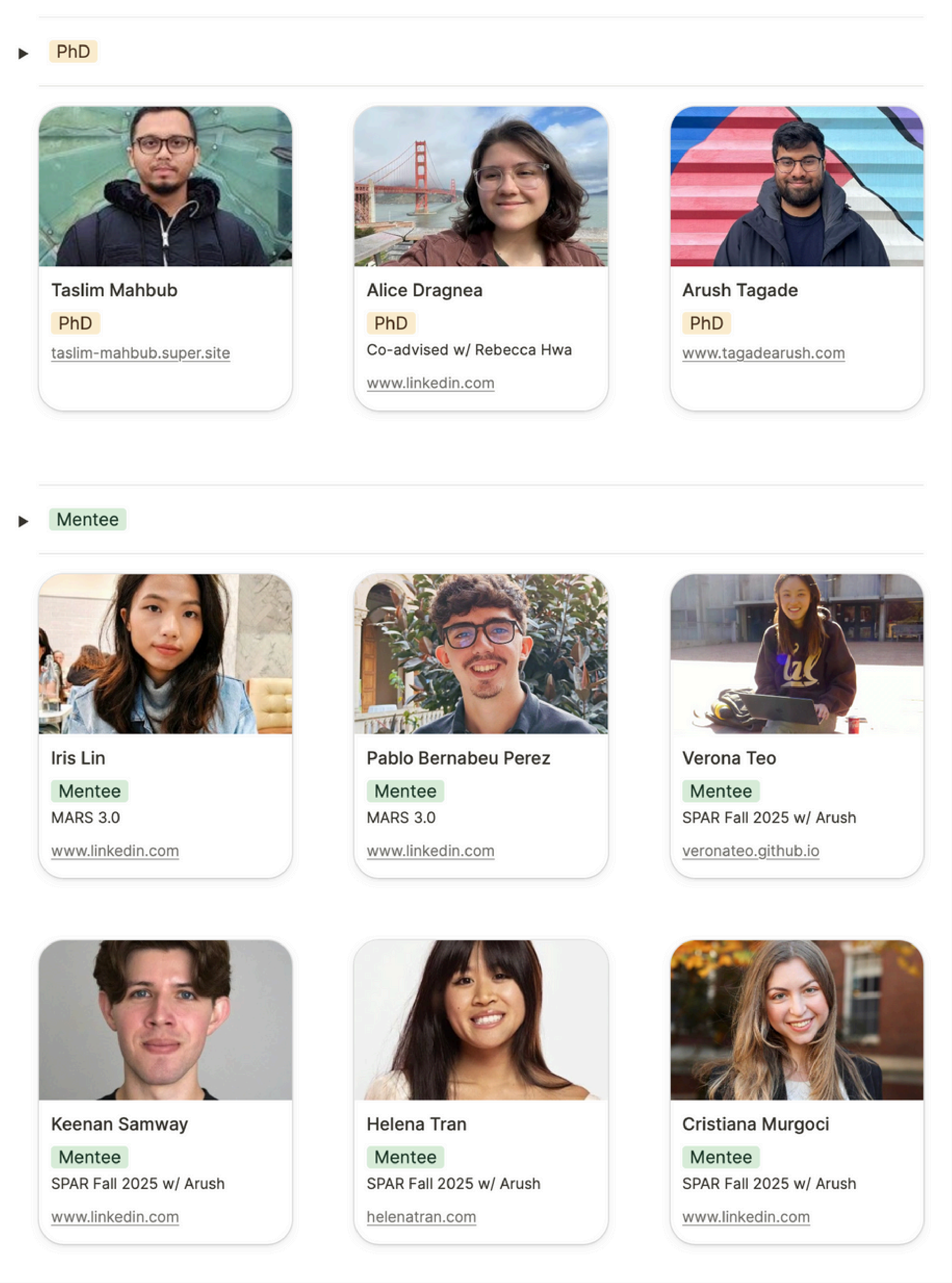
Self-Preference Undermines LM Evaluation

Shi Feng

George Washington University



▸ PhD

Taslim Mahbub
PhD
taslim-mahbub.super.site

Alice Dragnea
PhD
Co-advised w/ Rebecca Hwa
www.linkedin.com

Arush Tagade
PhD
www.tagadearush.com

▸ Mentee

Iris Lin
Mentee
MARS 3.0
www.linkedin.com

Pablo Bernabeu Perez
Mentee
MARS 3.0
www.linkedin.com

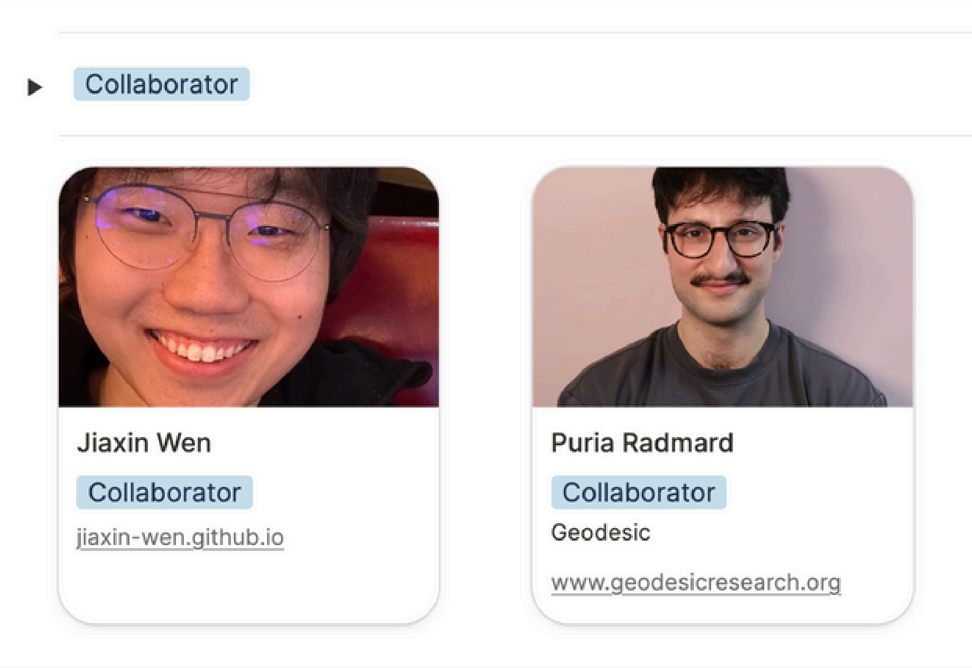
Verona Teo
Mentee
SPAR Fall 2025 w/ Arush
veronateo.github.io

Keenan Samway
Mentee
SPAR Fall 2025 w/ Arush
www.linkedin.com

Helena Tran
Mentee
SPAR Fall 2025 w/ Arush
helenatran.com

Cristiana Murgoci
Mentee
SPAR Fall 2025 w/ Arush
www.linkedin.com

Our research goal: to ensure human oversight over ***future*** AI systems.

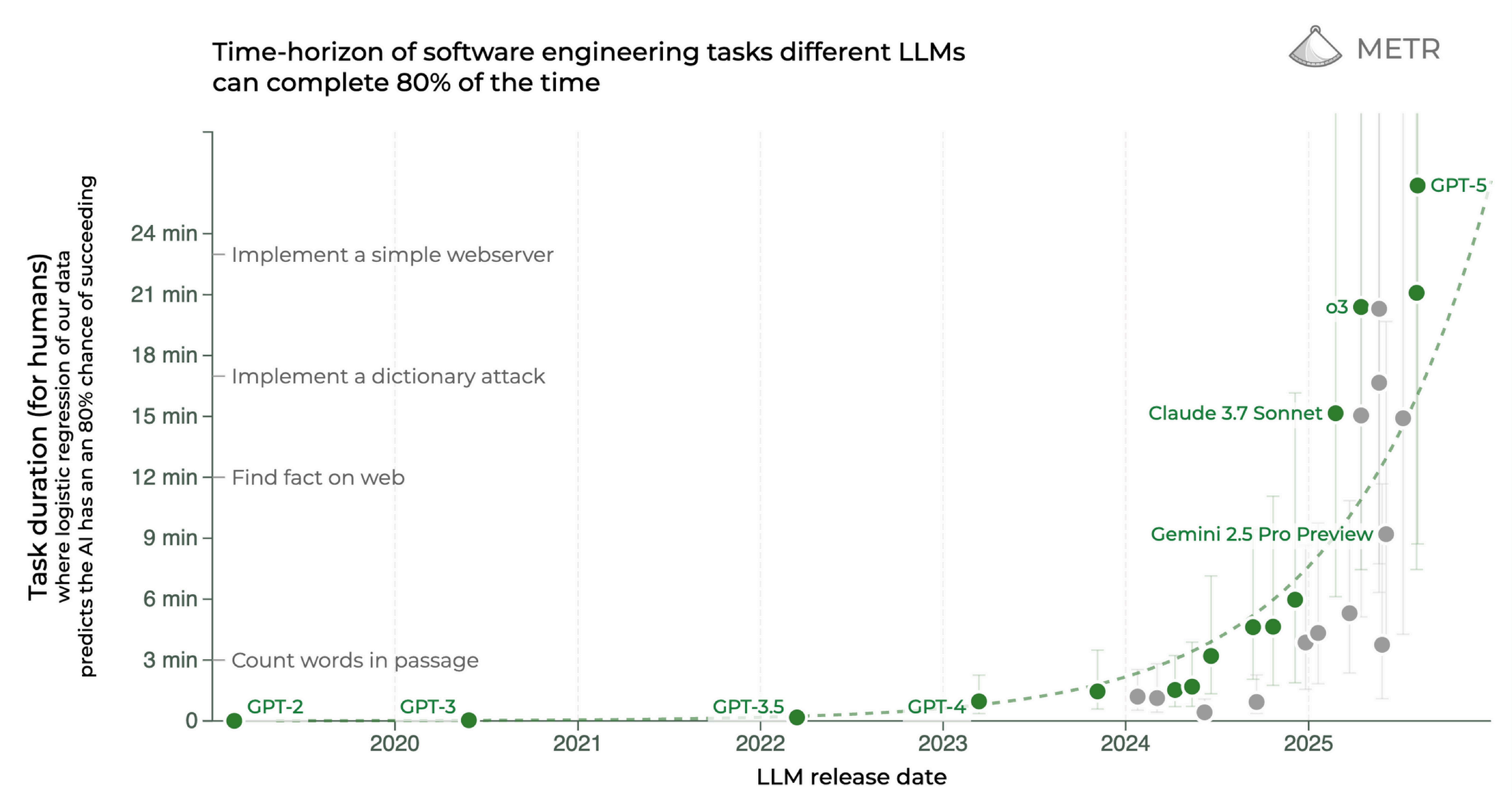


▸ Collaborator

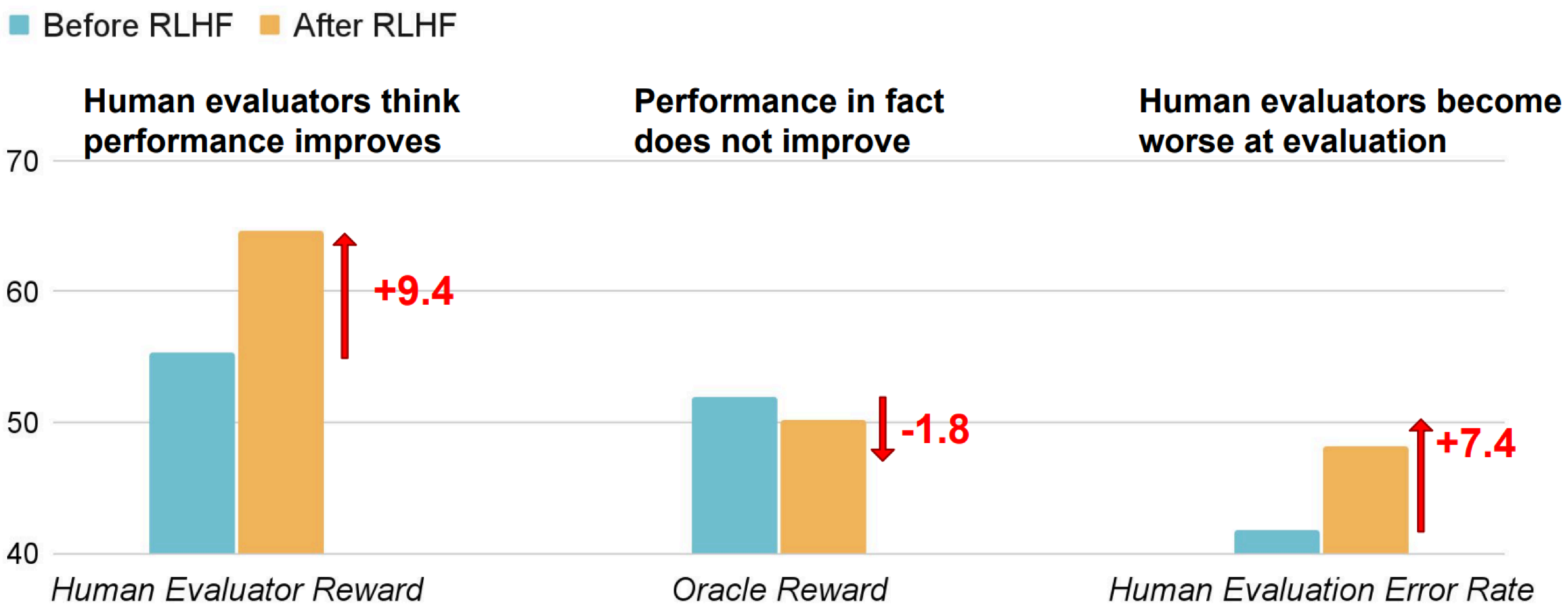
Jiaxin Wen
Collaborator
jiaxin-wen.github.io

Puria Radmard
Collaborator
Geodesic
www.geodesicresearch.org

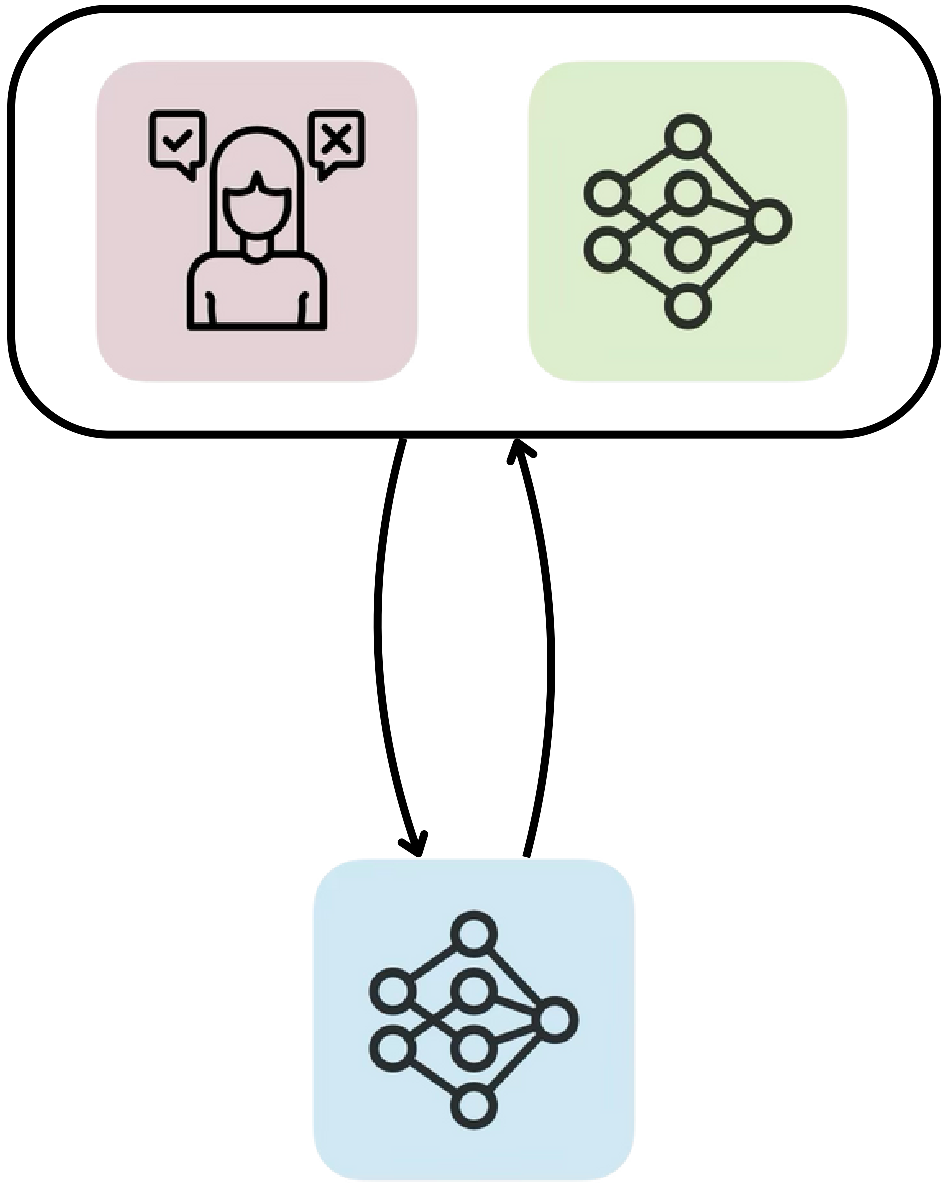
Increasing **Capability** of LMs



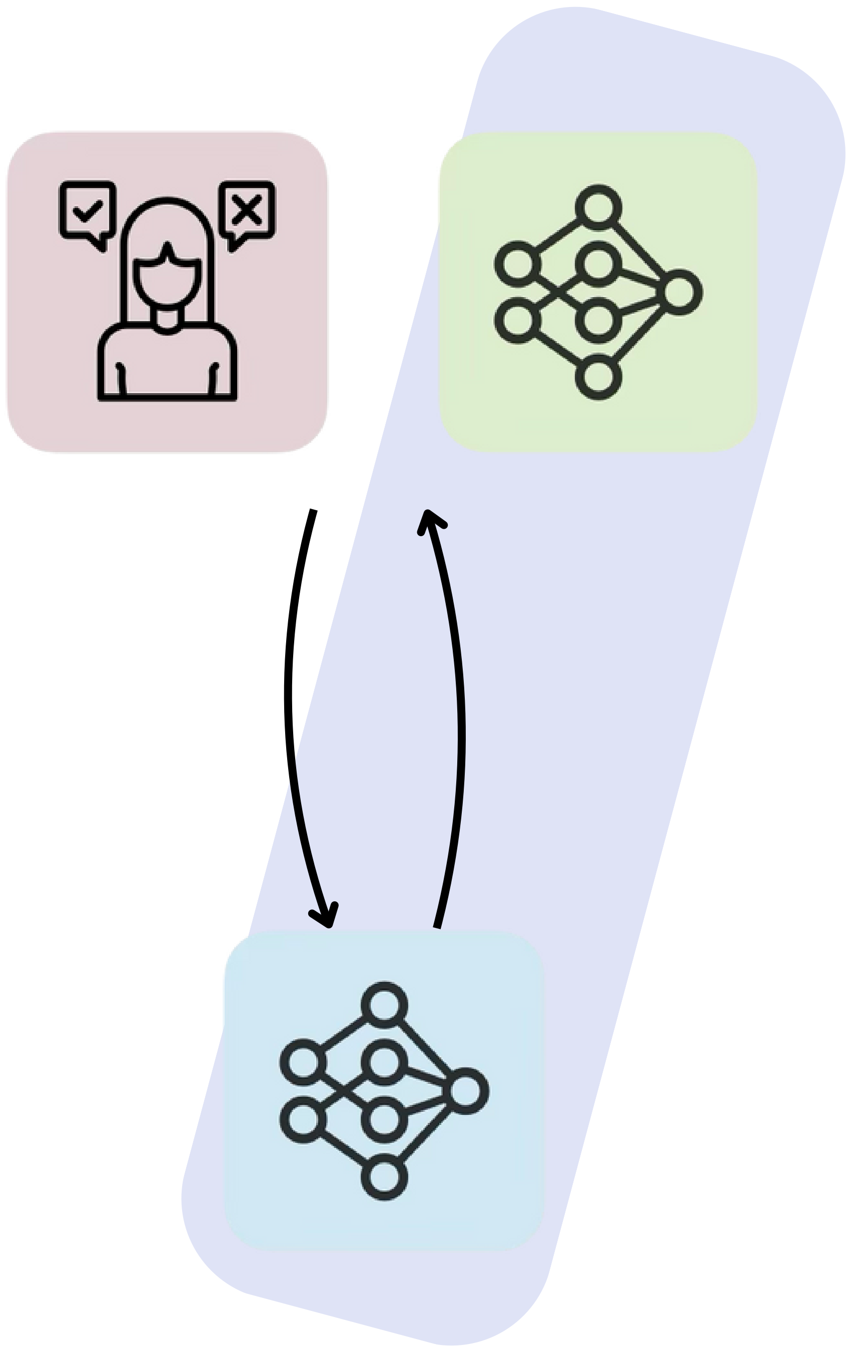
Increasing **Difficulty** of Human Supervision



Increasing **Need** for Human-AI Teaming



Safety Risks in Human-AI teaming



Self-preference

LM judges being biased towards their own outputs or those similar.

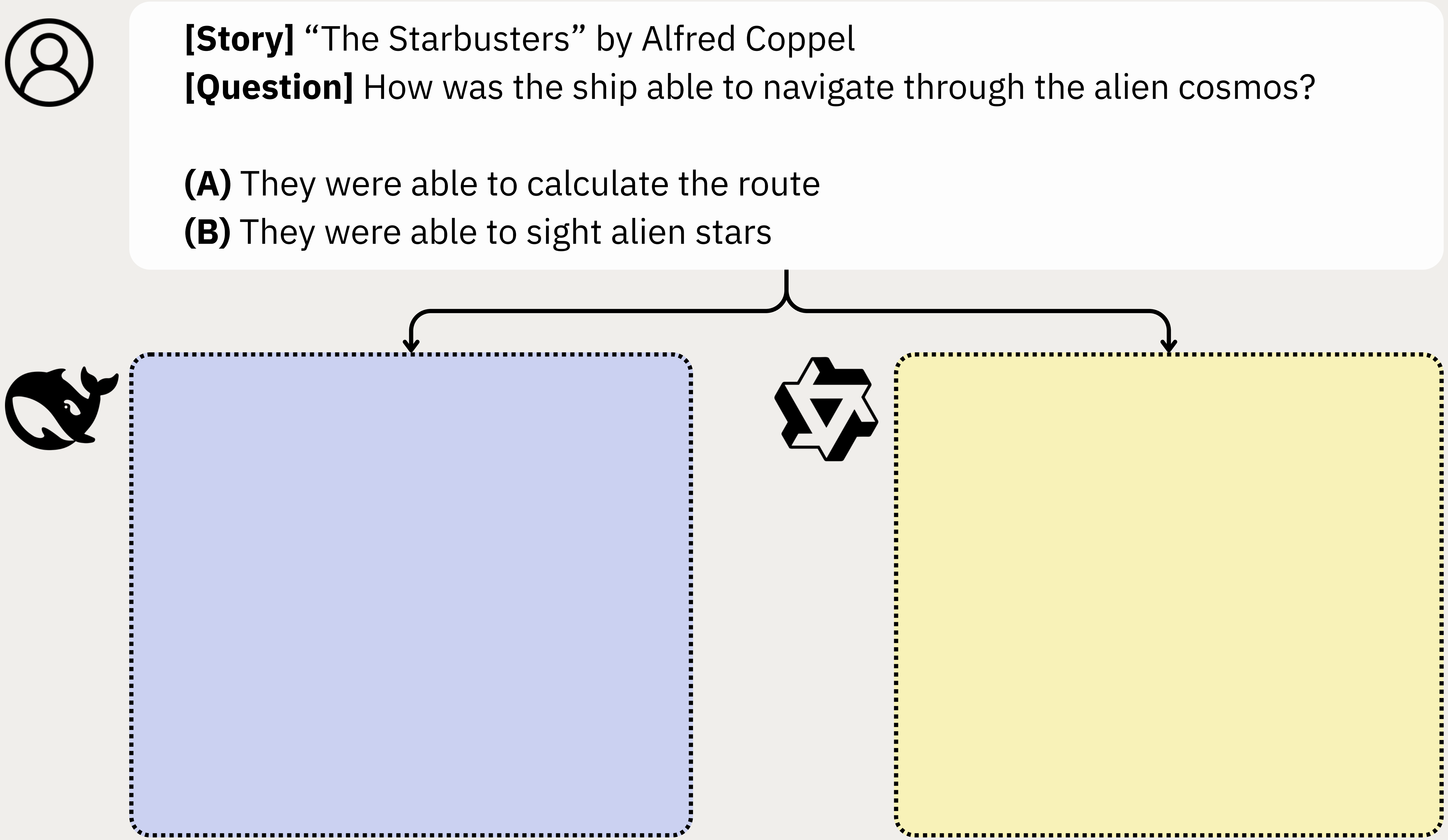
QuALITY question

[Story] “The Starbusters” by Alfred Coppel
[Question] How was the ship able to navigate through the alien cosmos?

(A) They were able to calculate the route
(B) They were able to sight alien stars

QuALITY question

Pair of Responses

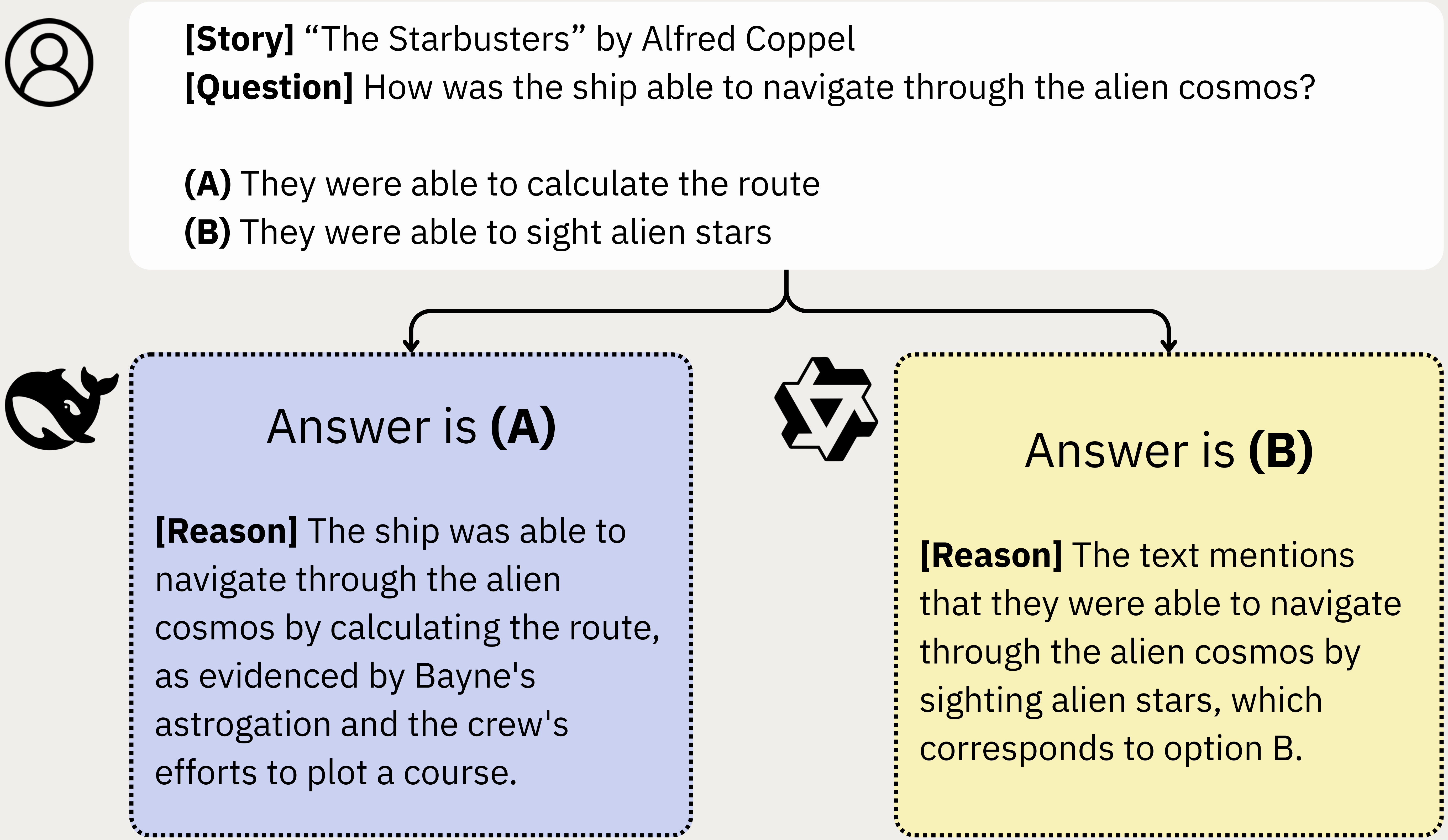


[Story] “The Starbusters” by Alfred Coppel

[Question] How was the ship able to navigate through the alien cosmos?

(A) They were able to calculate the route

(B) They were able to sight alien stars



QuALITY question

[Story] “The Starbusters” by Alfred Coppel

[Question] How was the ship able to navigate through the alien cosmos?

(A) They were able to calculate the route

(B) They were able to sight alien stars

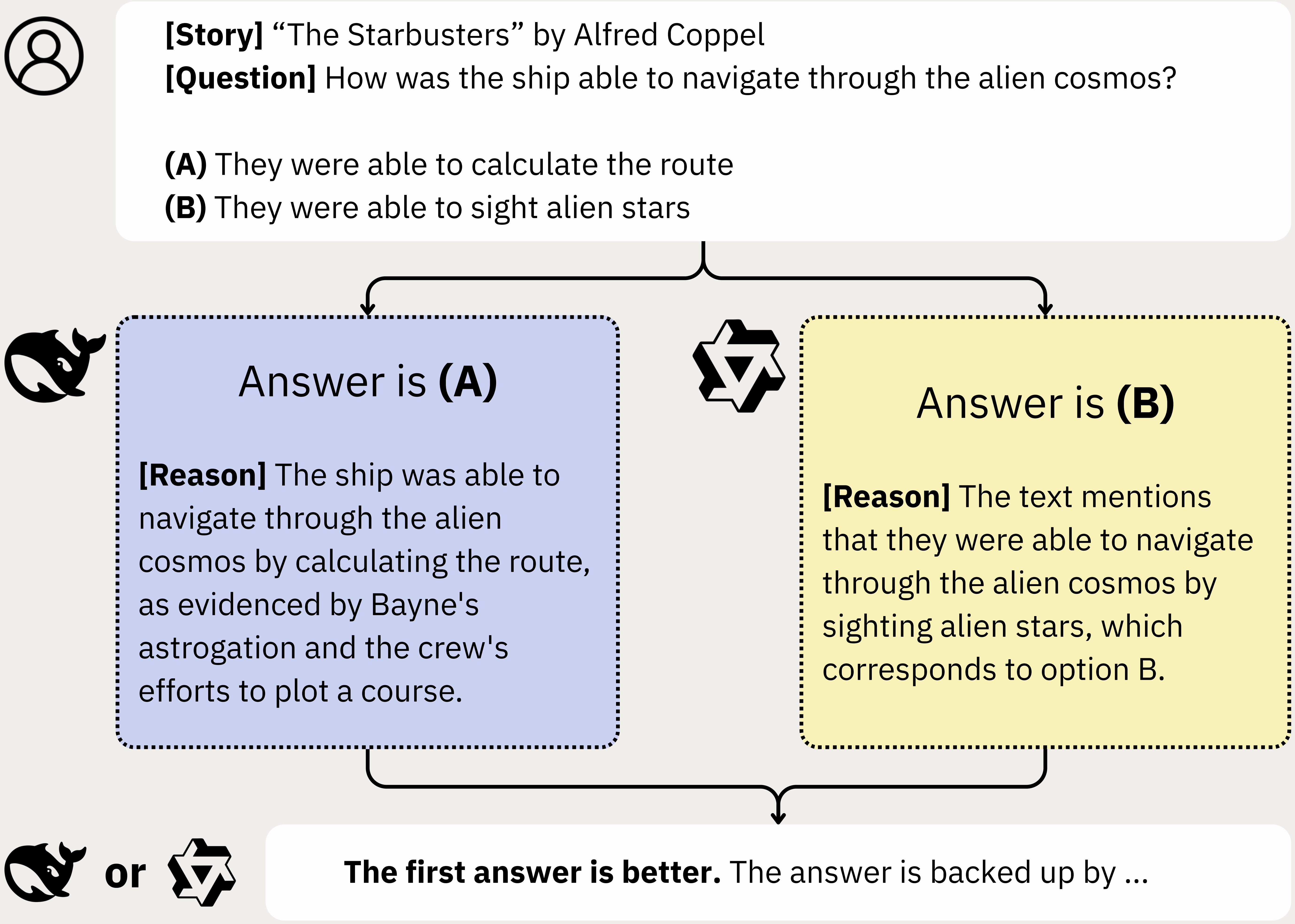
Pair of Responses

Answer is **(A)**

[Reason] The ship was able to navigate through the alien cosmos by calculating the route, as evidenced by Bayne's astrogation and the crew's efforts to plot a course.

Answer is **(B)**

[Reason] The text mentions that they were able to navigate through the alien cosmos by sighting alien stars, which corresponds to option B.



QuALITY question

[Story] "The Starbusters" by Alfred Coppel
[Question] How was the ship able to navigate through the alien cosmos?

(A) They were able to calculate the route
(B) They were able to sight alien stars

Pair of Responses

Answer is **(A)**

[Reason] The ship was able to navigate through the alien cosmos by calculating the route, as evidenced by Bayne's astrogation and the crew's efforts to plot a course.

Answer is **(B)**

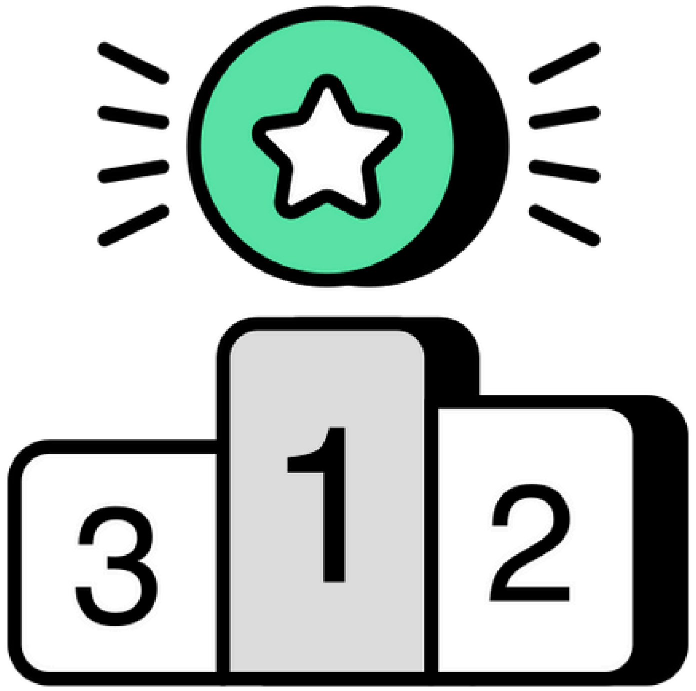
[Reason] The text mentions that they were able to navigate through the alien cosmos by sighting alien stars, which corresponds to option B.

Eval by LM

The first answer is better. The answer is backed up by ...

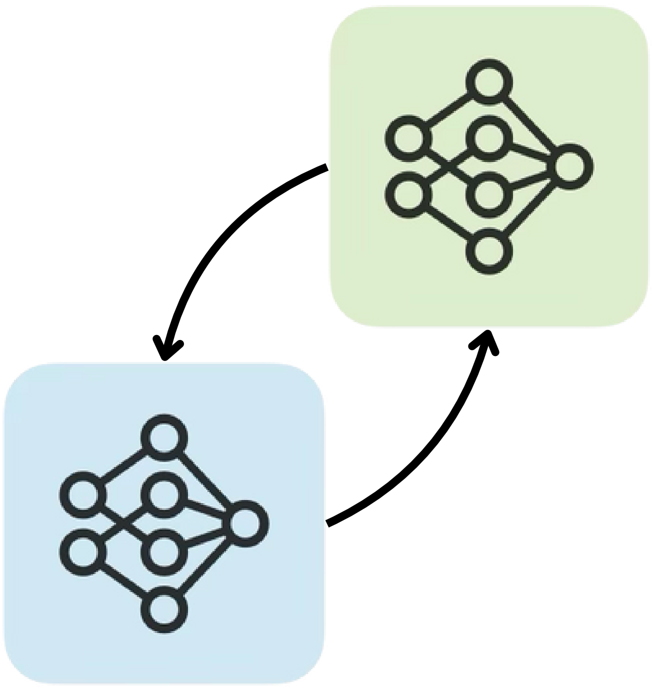
LM-as-a-judge is applicable everywhere

Benchmarks

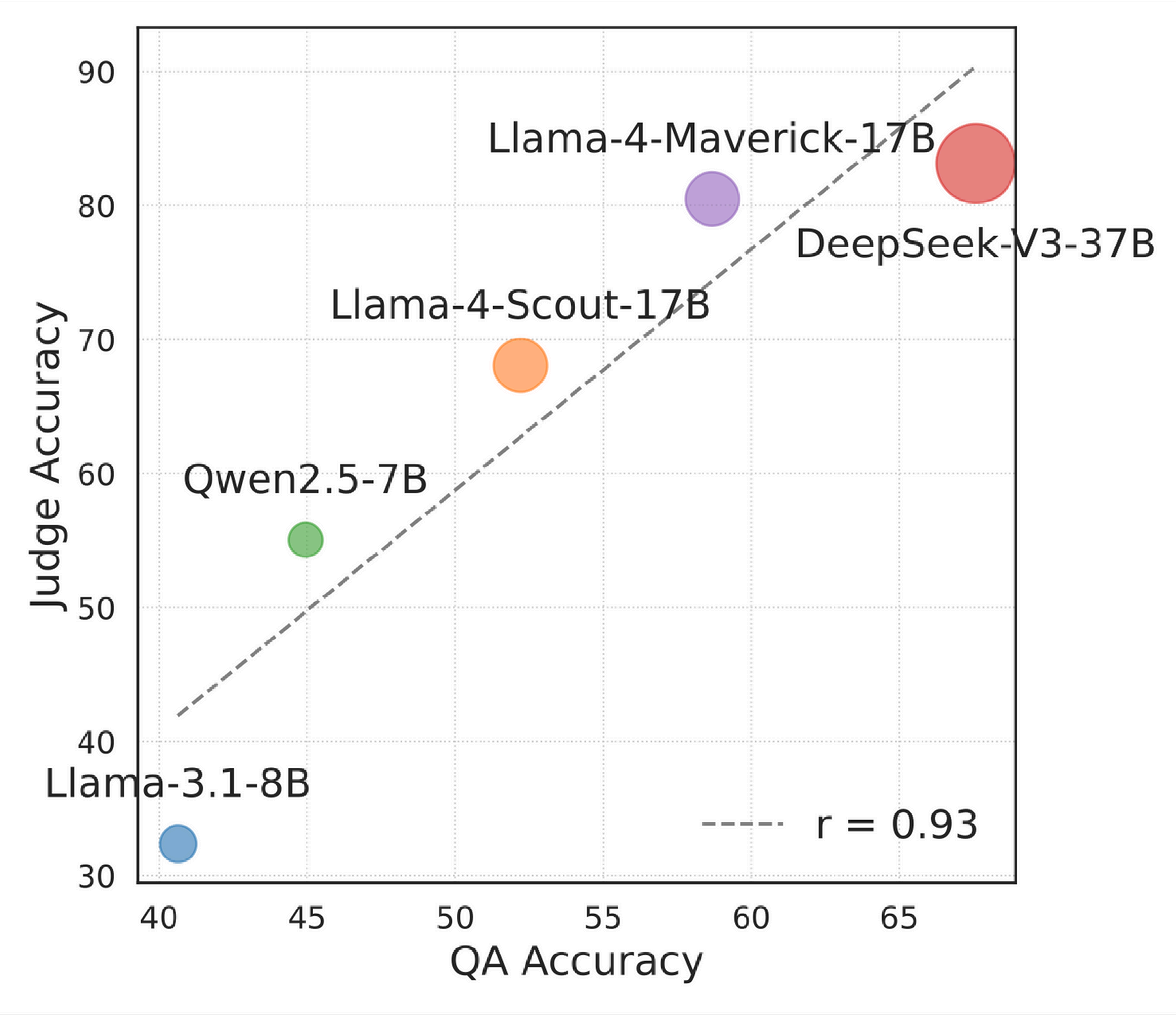


Safeguard

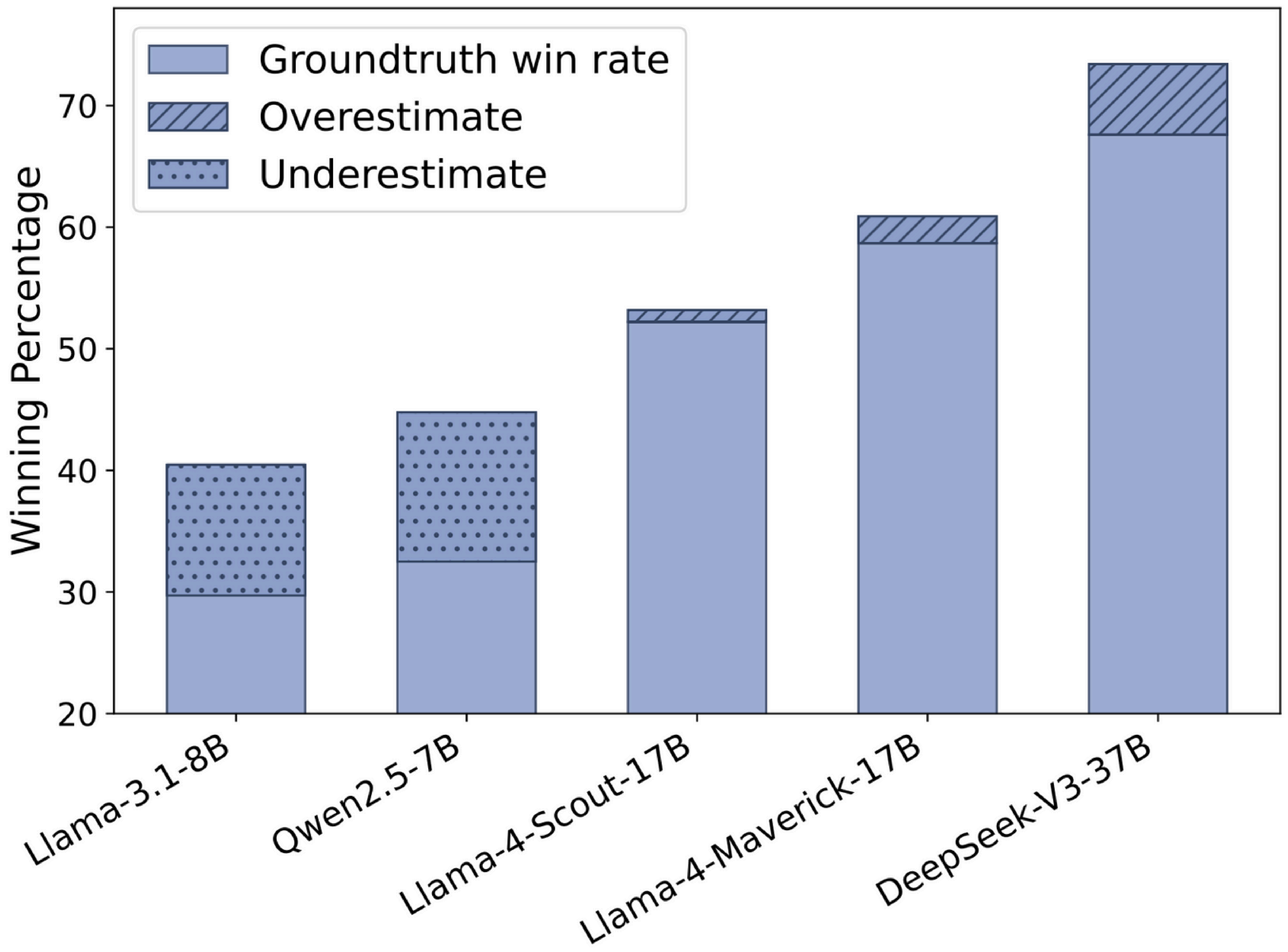
Training



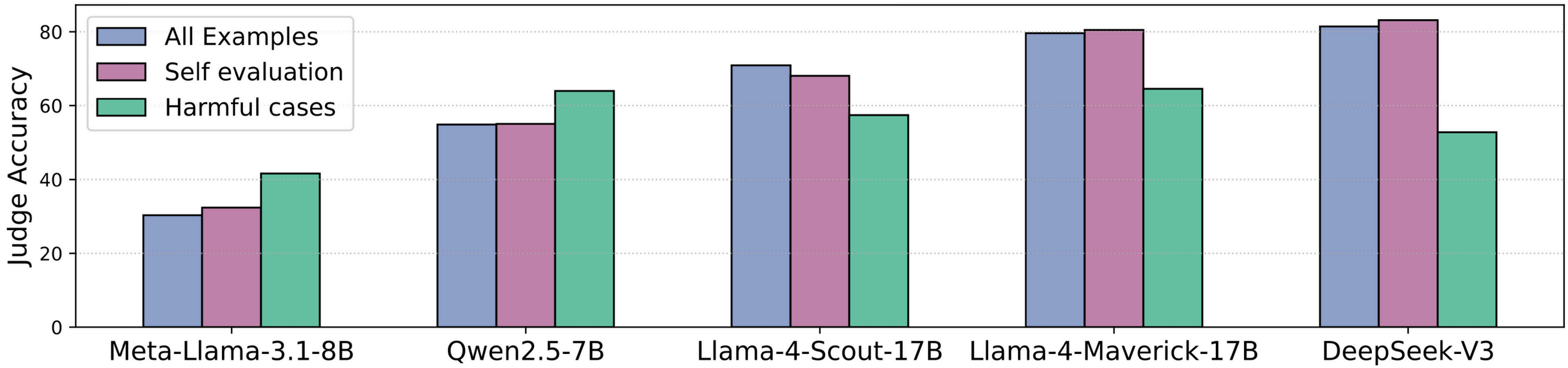
Stronger models are **better Judges**



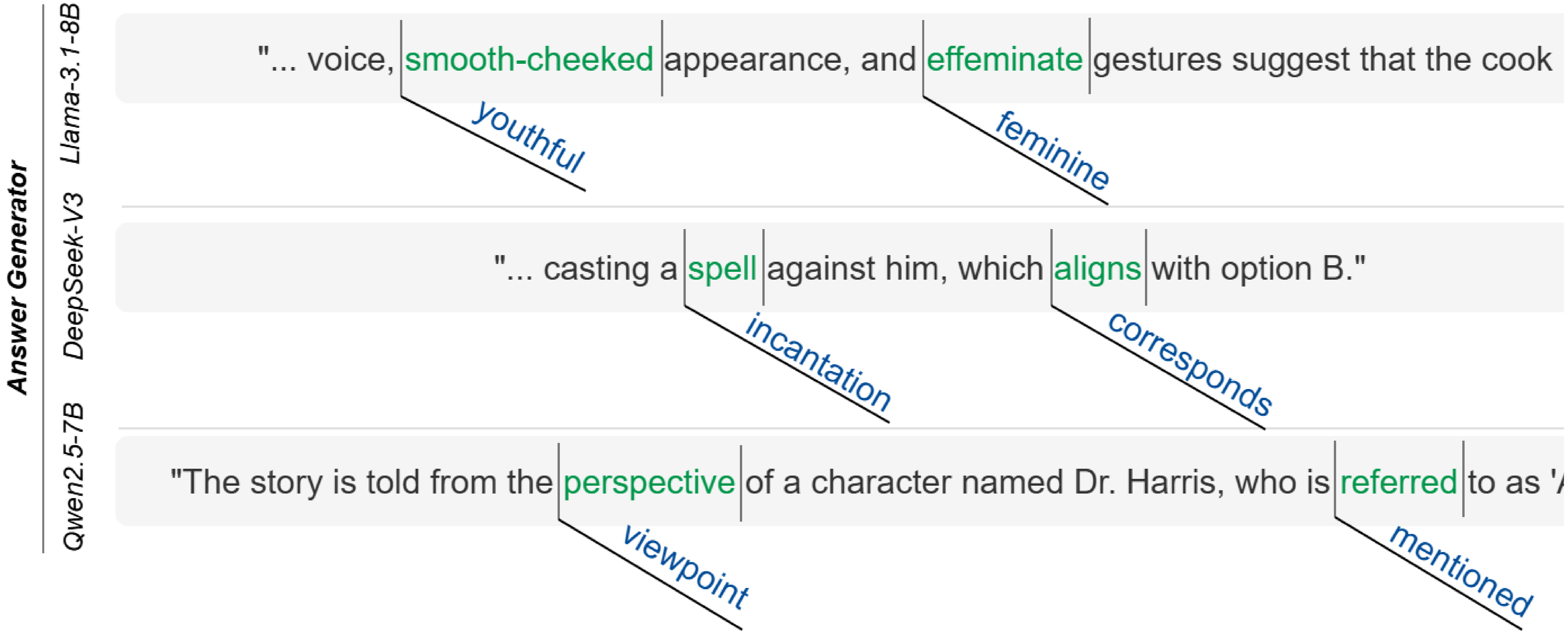
Stronger models **Overestimate** own accuracy



Larger accuracy drop when strong models are wrong



Mitigating self-preference by **perturbations**

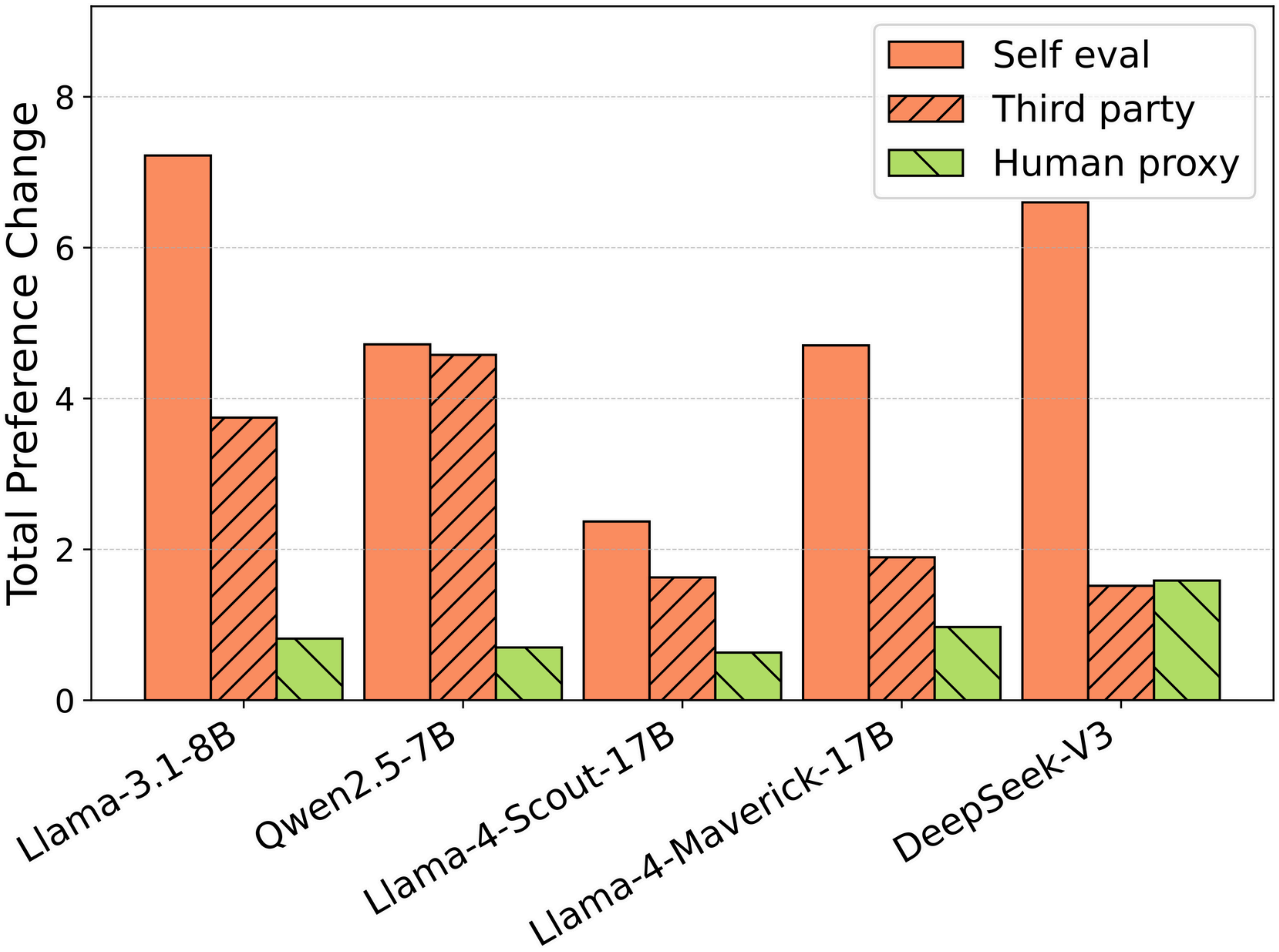


"... voice, smooth-cheeked appearance, and effeminate gestures suggest that the cook

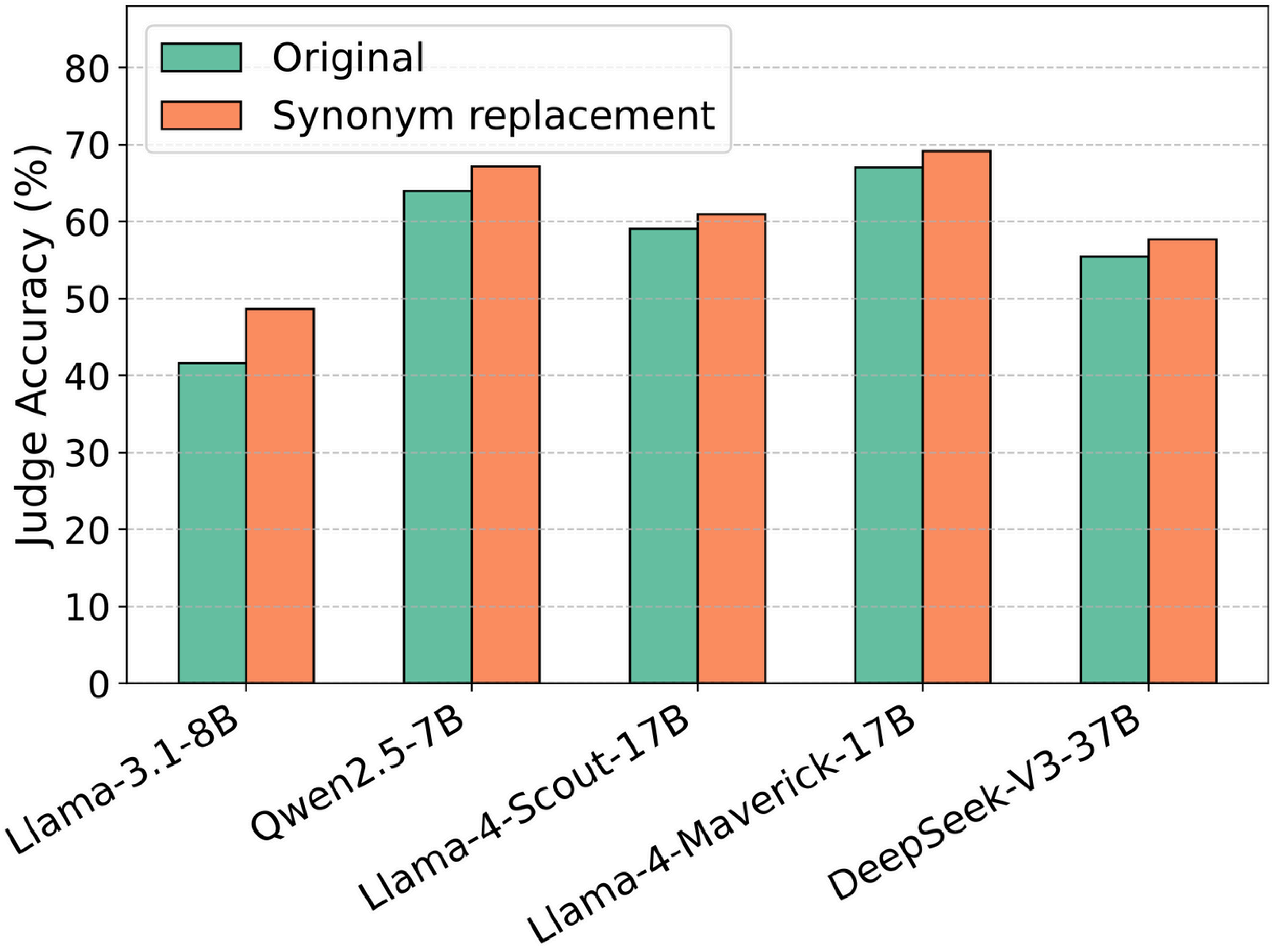
"... casting a spell against him, which aligns with option B."

"The story is told from the perspective of a character named Dr. Harris, who is referred to as '/

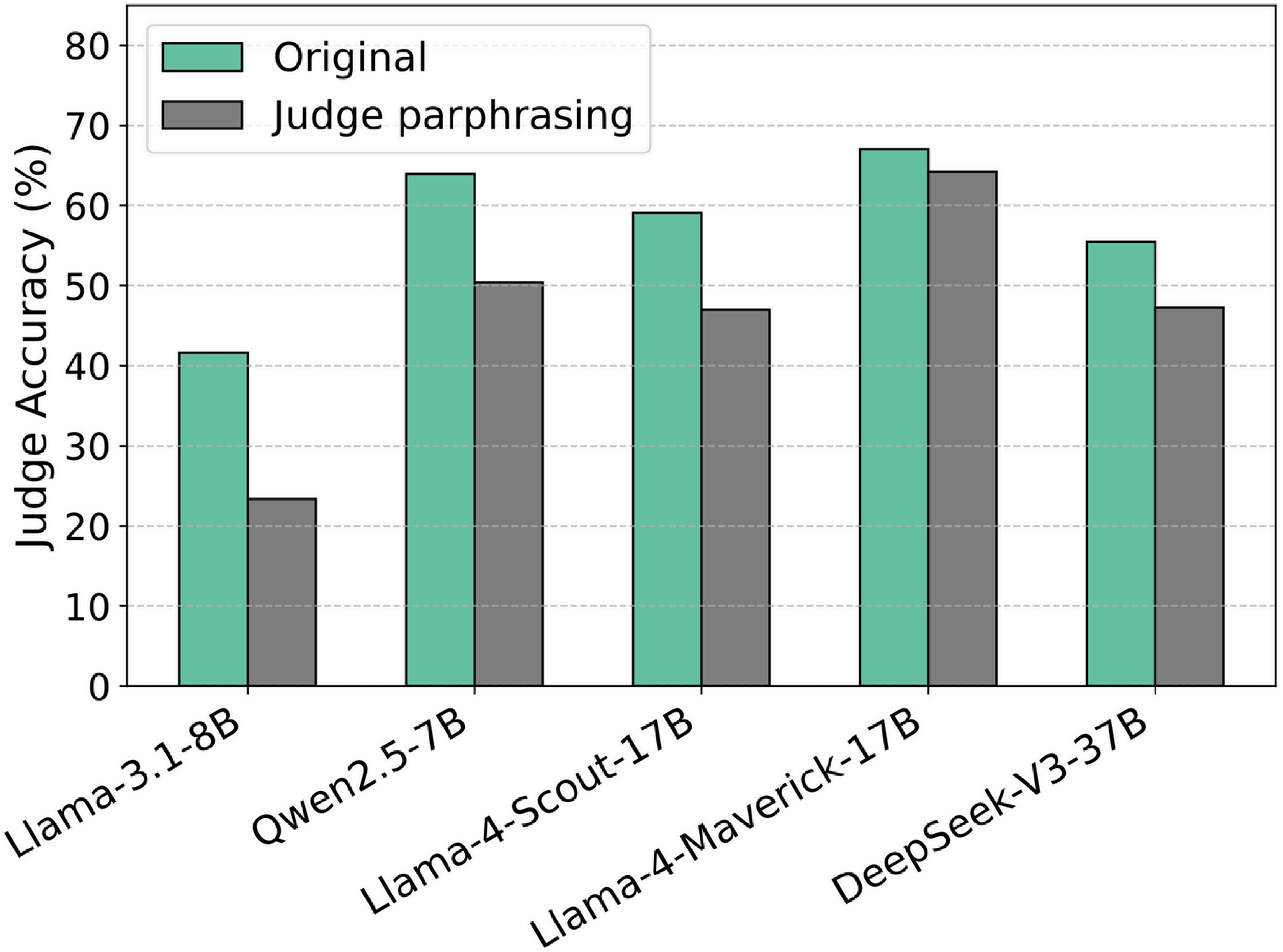
Judge decisions are **sensitive to perturbations**



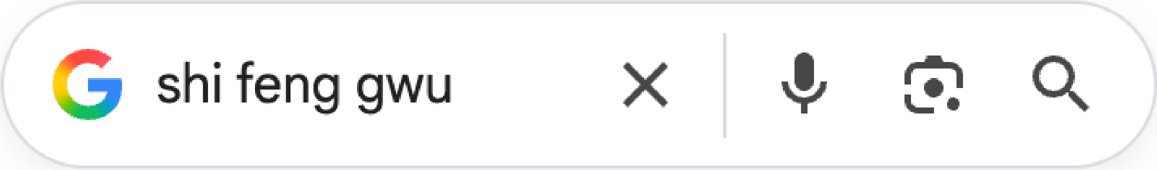
Judges are **more accurate** after perturbations



Not all perturbations help



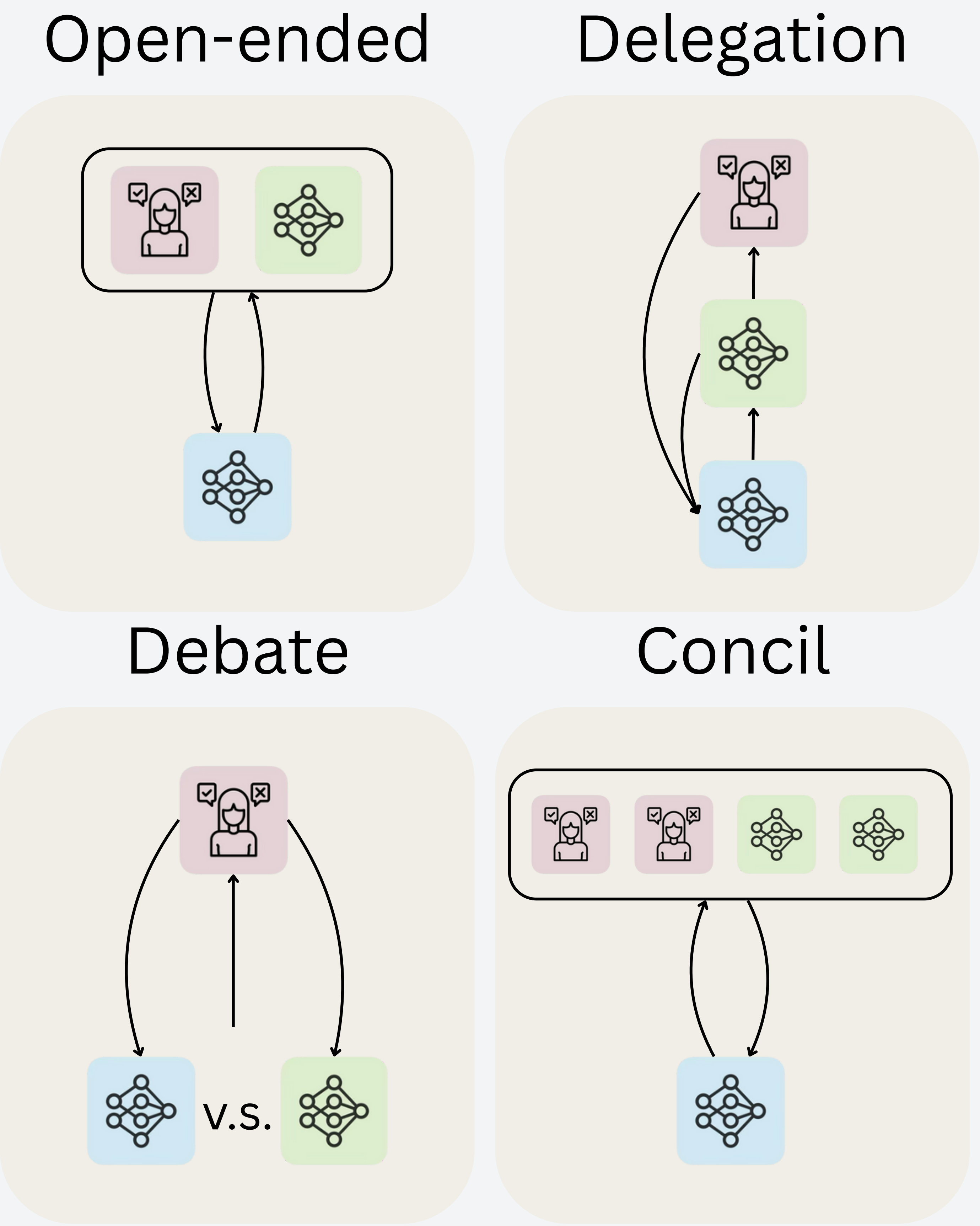
To learn more:



shi feng gwu

- **Self-preference** is an important factor in LM evaluation fairness and accuracy.
- **Authorship obfuscation** is a viable strategy to mitigate self-preference.
- Not all obfuscation methods are helpful. We need to better understand **how** LMs do self-recognition.

Human-AI teams are complex systems



- Self-preference can be amplified by the feedback loop of ML training
- How does SE create abstractions of these complex protocols?
- What are common ways to reduce amplification that we haven’t tried?
- What’s the standard method to evolve these protocols?