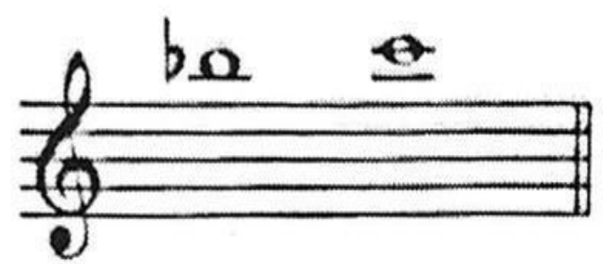
Key ingredient: *Rationales*

Rationales reveal the skills and steps needed to answer a question.

Evaluation Instance (MMMU, #744)

```
Select the type of the interval.

(A) perfect
(B) major
(C) minor
(D) diminished
```



Generated Rationale with localized skills

Step 1. Recognize the Image
- *Skill: **Perception, Visual Recognition, Symbol Identification, Treble clef recognition***
- *Conclusion: The image is of notes on the treble staff.*

Step 2. Identify the Notes
- *Skill: **Perception, Visual Recognition, Note Identification, Pitch recognition***
- *Conclusion: The notes are B and E.*

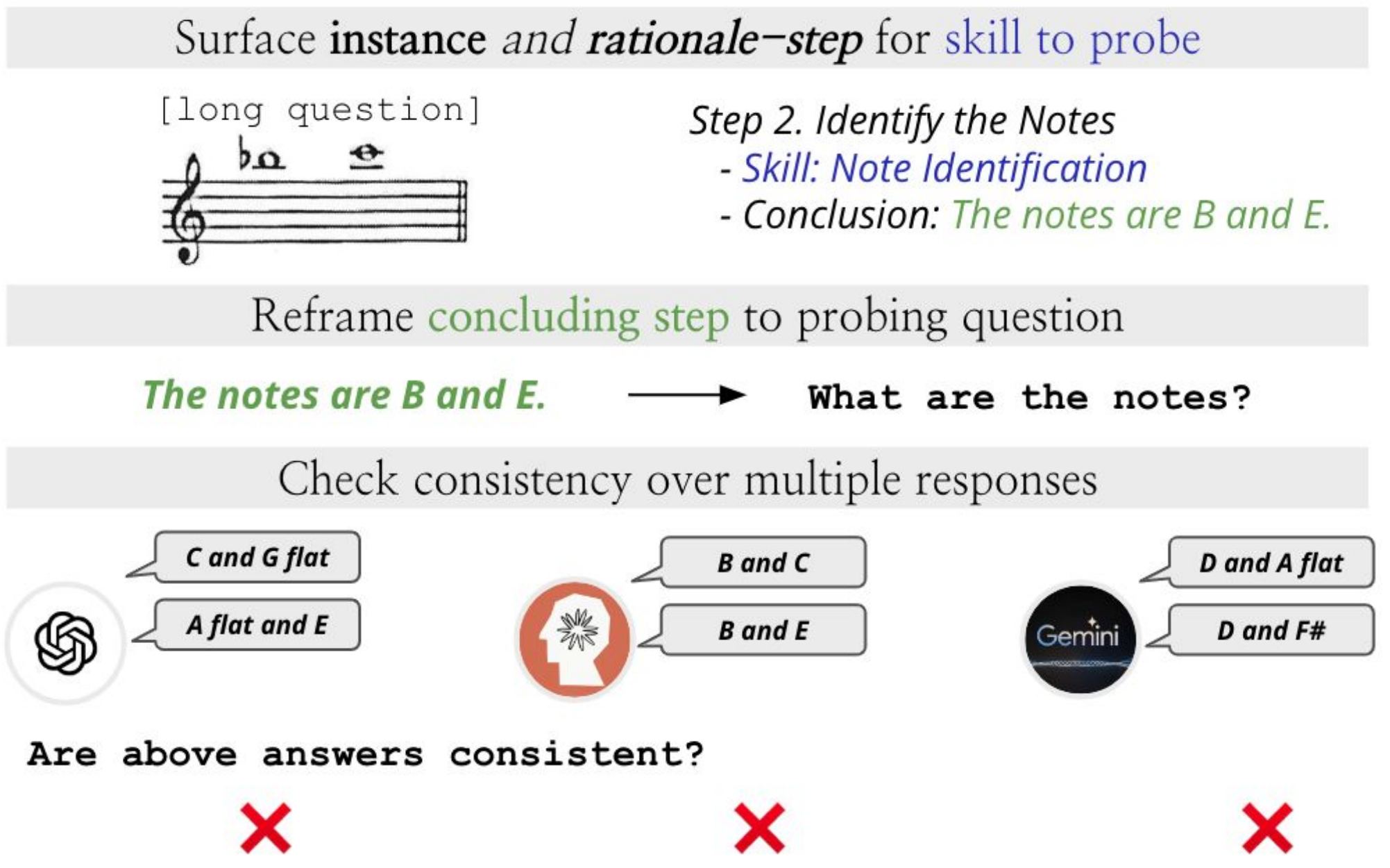
Step 3. Determine the Interval
- *Skill: **Knowledge, Music Theory, Interval Identification, Pitch relationships***

...

Testing along *"skill–slices"* (instances testing the same skill) unlocks richer insight from existing evaluation data.

Isolating Skills via Probing Questions

Reframing rationale steps where a skill is applied allows for generating questions that *test only one skill.*



Surface **instance** *and* ***rationale–step*** for skill to probe

```
[long question]
```

Step 2. Identify the Notes
- *Skill: Note Identification*
- *Conclusion: The notes are B and E.*

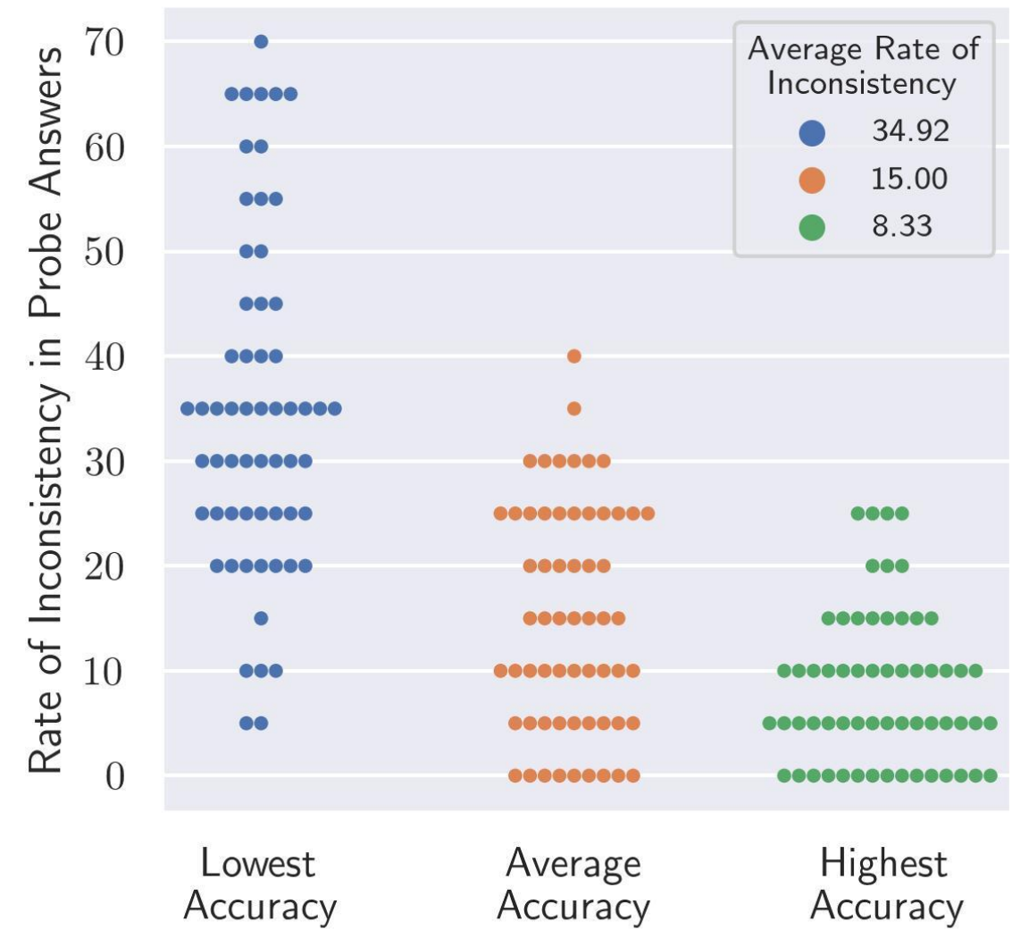
Reframe concluding step to probing question

The notes are B and E. ⟶ `What are the notes?`

Check consistency over multiple responses

`Are above answers consistent?`

✗ ✗ ✗

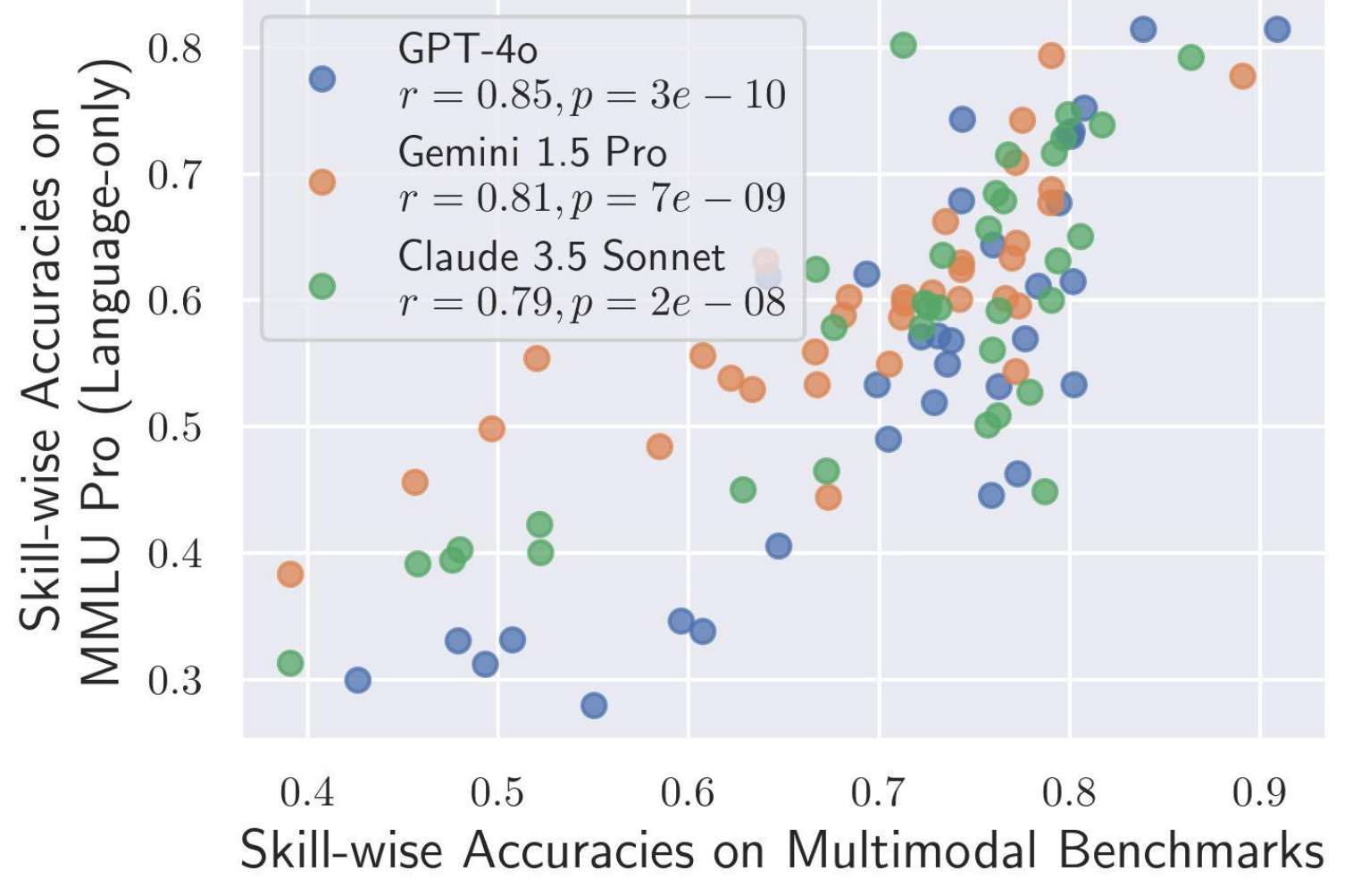


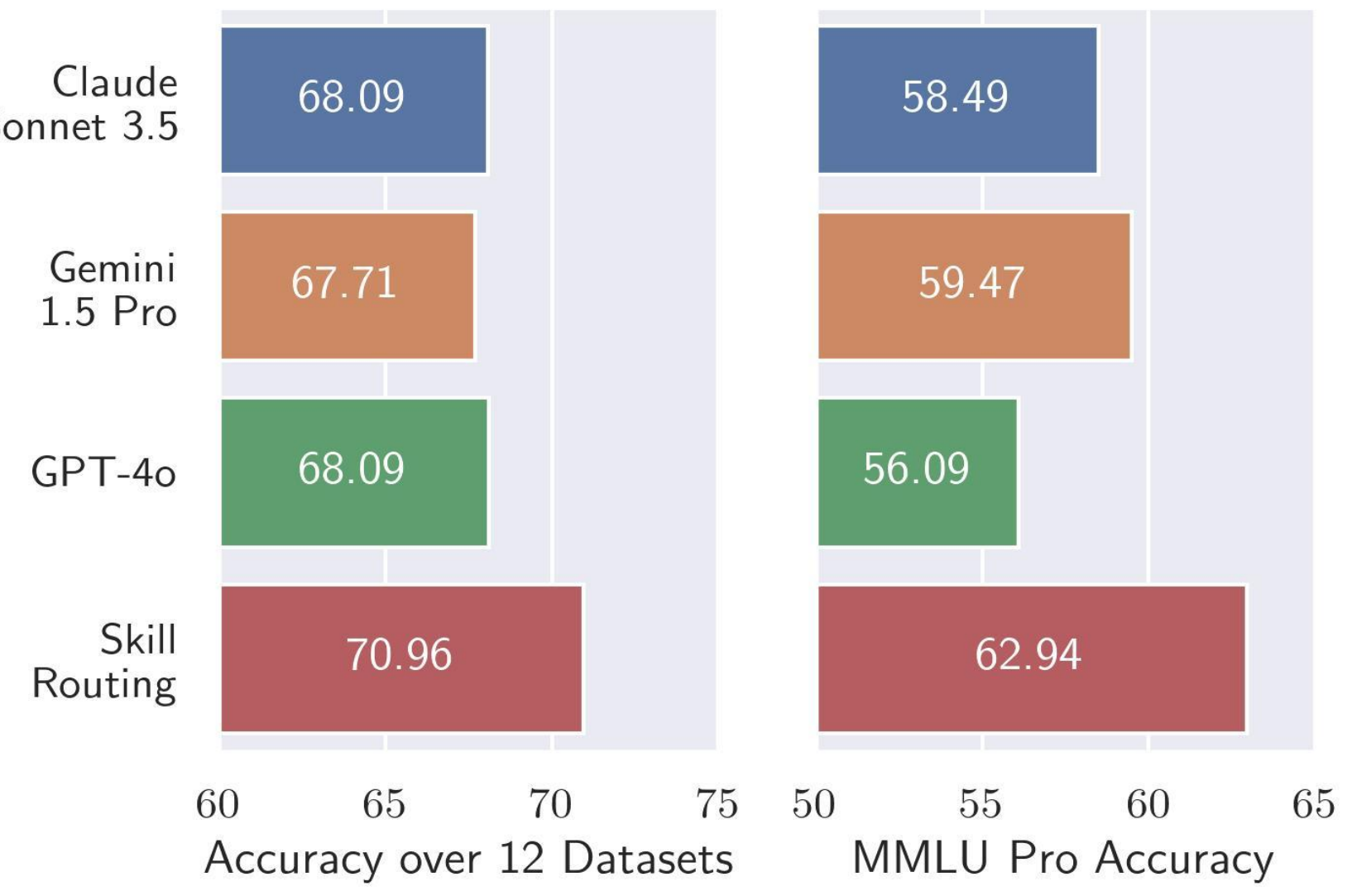
Consistency over multiple answers to the same probe question offers a 2nd corroborating and complementary (compared to *skill–slice* accuracies) measure of skill proficiency.

Skill–based Routing Improves Accuracy

Skill–slice accuracies generalize.

Highly correlated accuracies over slices drawn from distinct corpuses, *even when modality changes.*





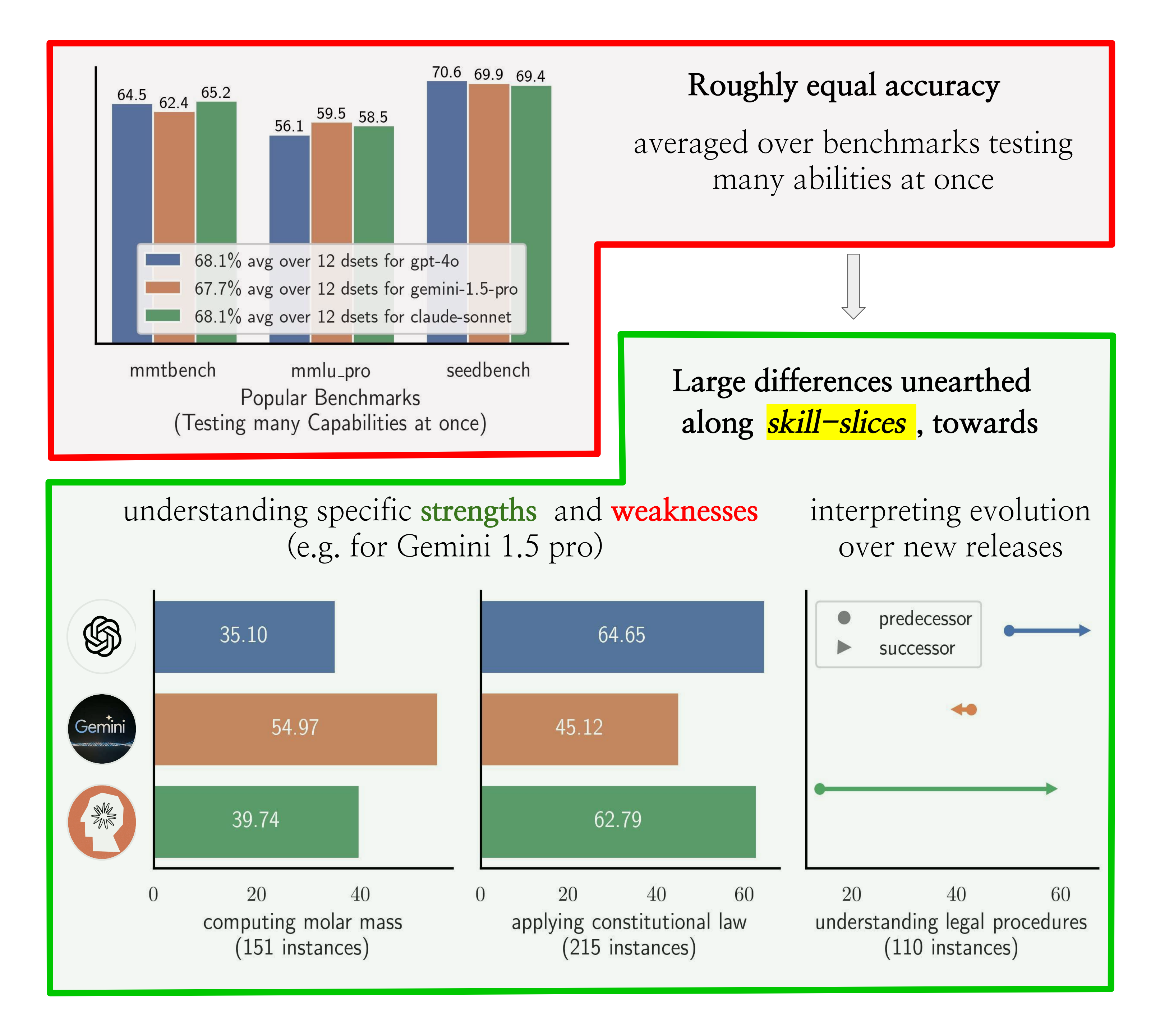
Routing each instance to the model with the highest slice accuracies for the relevant skills leads to accuracy gains. (+3 to 6.8% on MMLU–Pro)

Unearthing *Skill–Level Insights* for Understanding Tradeoffs of Foundation Models

Mazda Moayeri

Vidhisha Balachandran, Varun Chandrasekaran,
Safoora Yousefi, Thomas Fel, Soheil Feizi, Besmira Nushi, Neel Joshi, Vibhav Vineet

Work done with Microsoft Research AI Frontiers



Roughly equal accuracy

averaged over benchmarks testing many abilities at once

Large differences unearthed along *skill–slices*, towards

understanding specific strengths and weaknesses (e.g. for Gemini 1.5 pro)

interpreting evolution over new releases

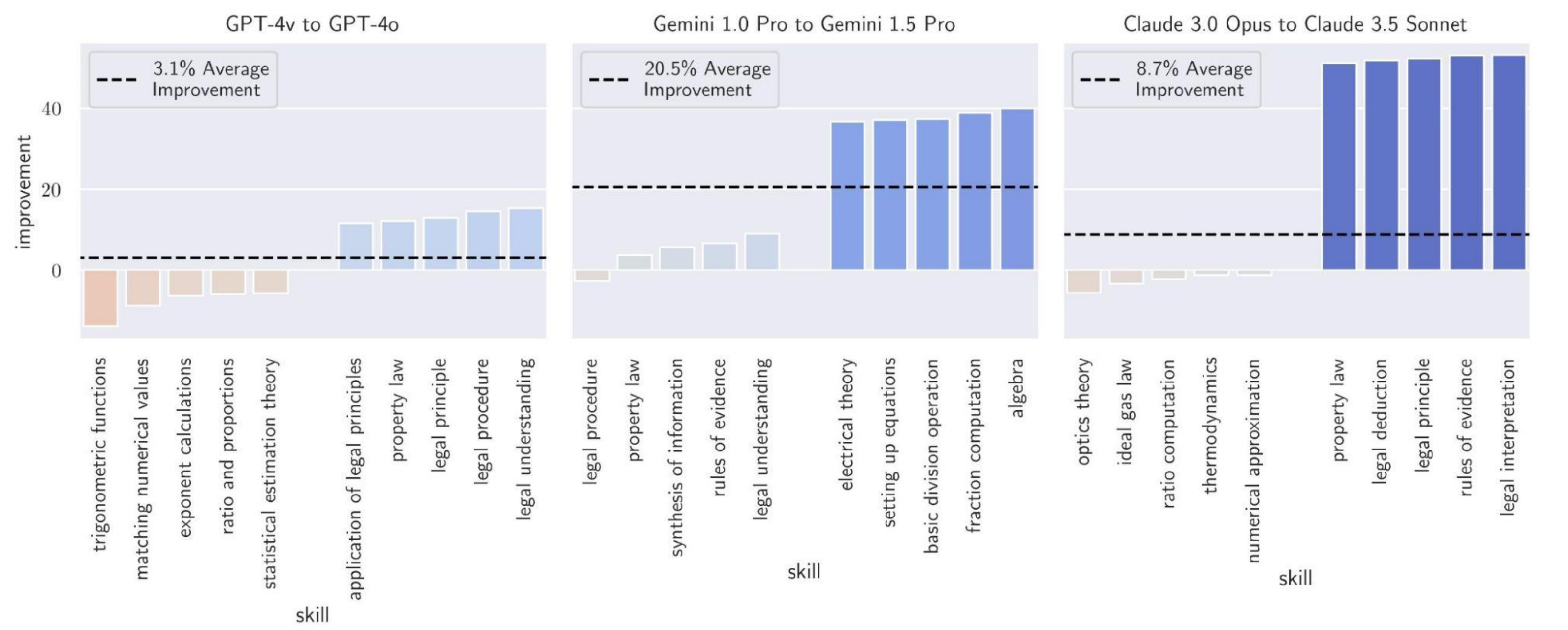
Foundation models, by design, encapsulate a wide breadth of skills. Thus, modern benchmarks test many skills, all at once – even in the same instance.

How many insights are we simply averaging away?

We present a method harnessing generated rationales to scalably recover the skill–level insights hiding within existing benchmark evaluations.

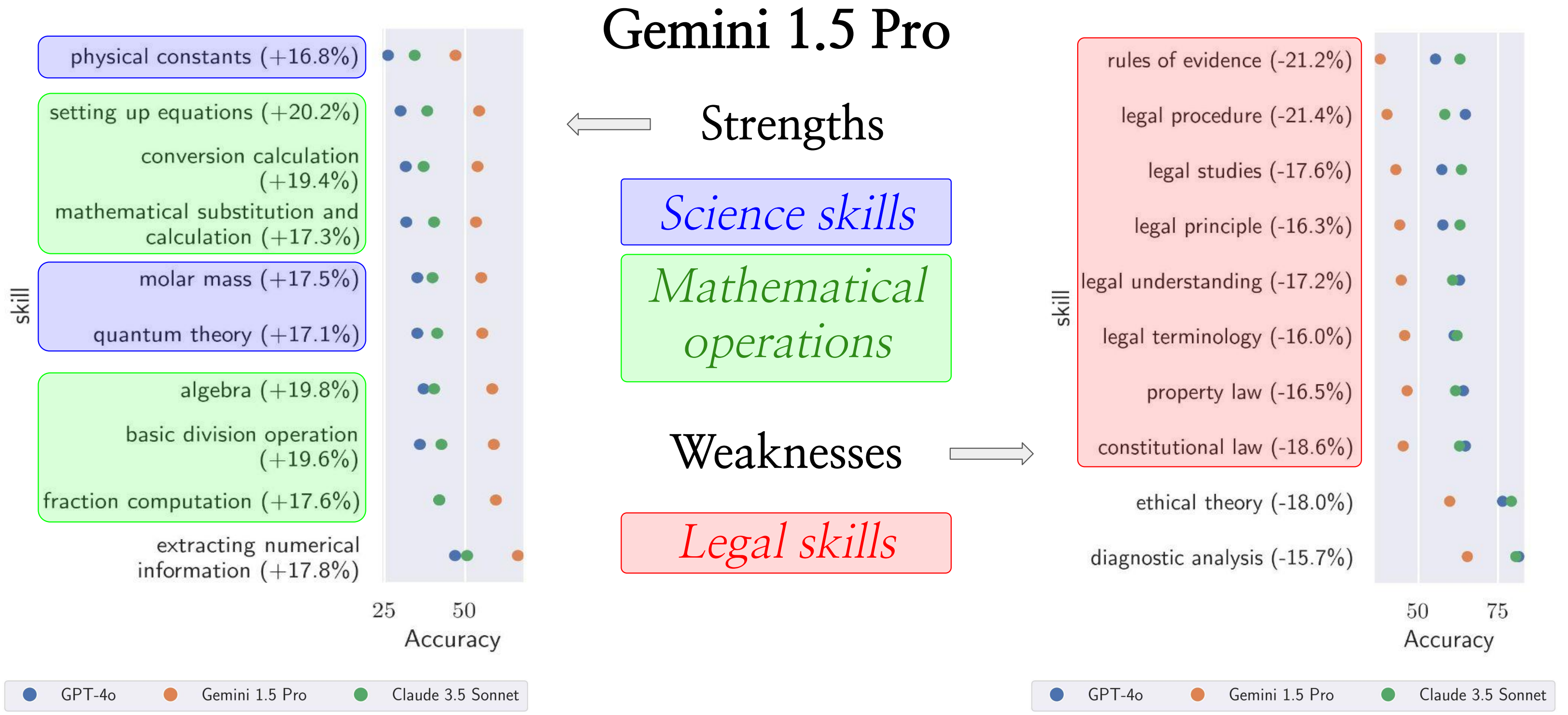
Evolution over Model Releases

Skill–slices uncover the most–improved skills from one release to the next.



Seems like legal skills were a recent priority for OpenAI and Anthropic. Below, we see legal skills are also the area where Gemini is furthest behind!

Strengths and Weaknesses of Top Models



Gemini 1.5 Pro

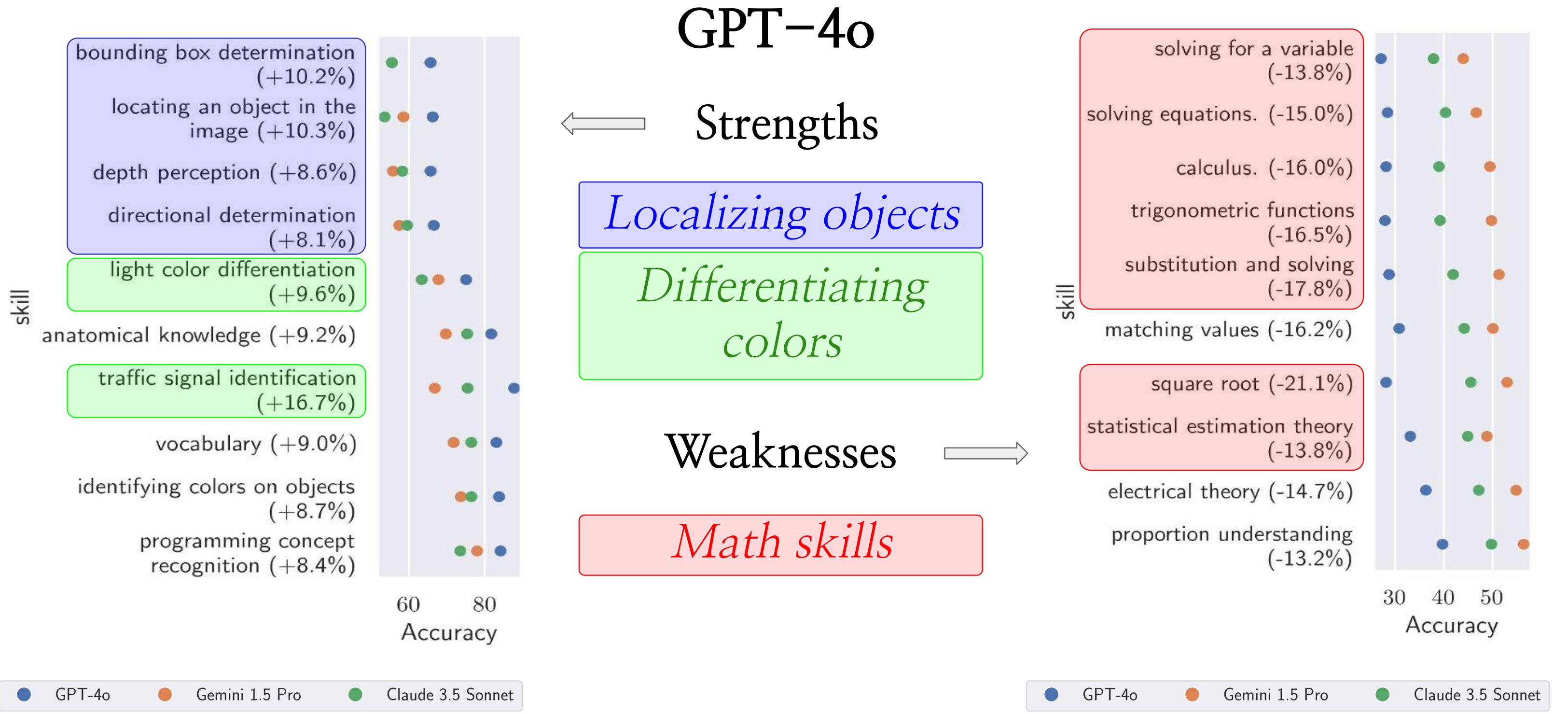
Strengths

Science skills

Mathematical operations

Weaknesses

Legal skills



GPT–4o

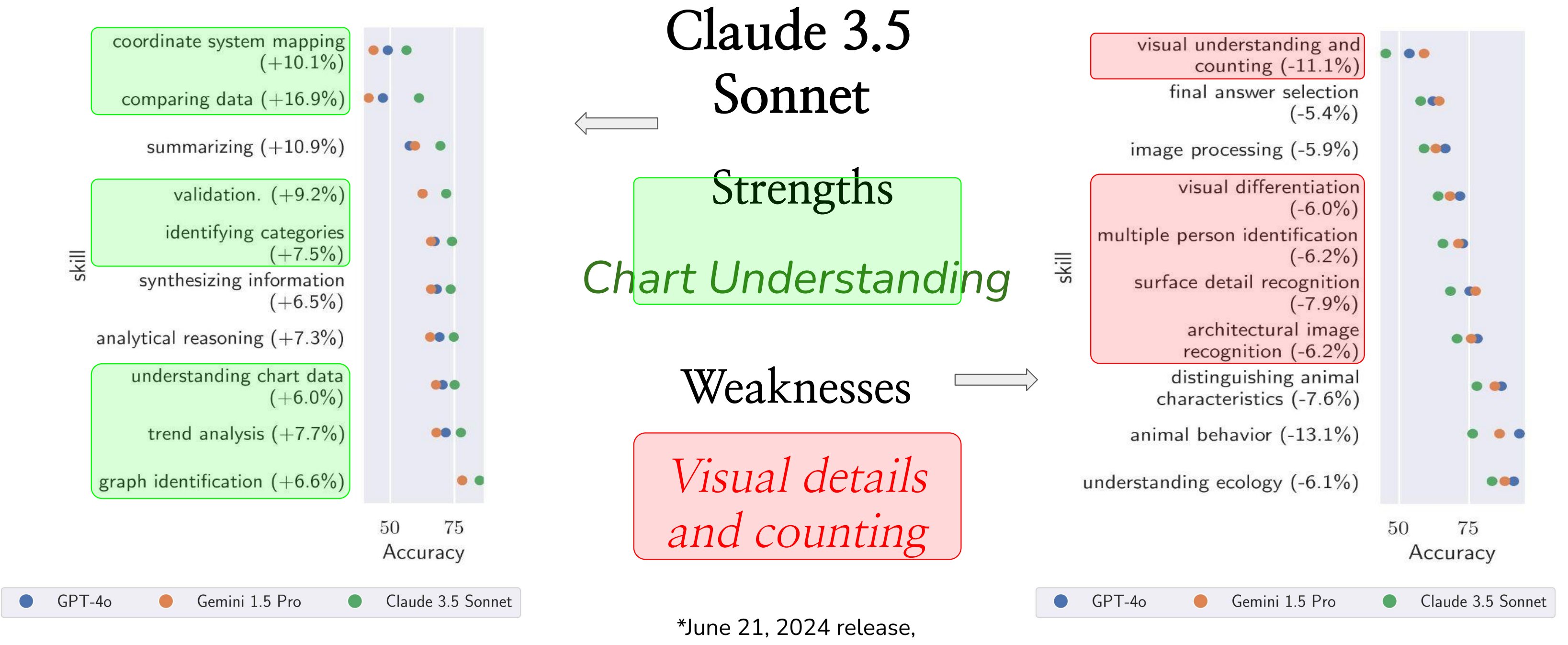
Strengths

Localizing objects

Differentiating colors

Weaknesses

Math skills



Claude 3.5 Sonnet

Strengths

Chart Understanding

Weaknesses

Visual details and counting

*June 21, 2024 release, not October 22 update