From Distributional to Overton Pluralism

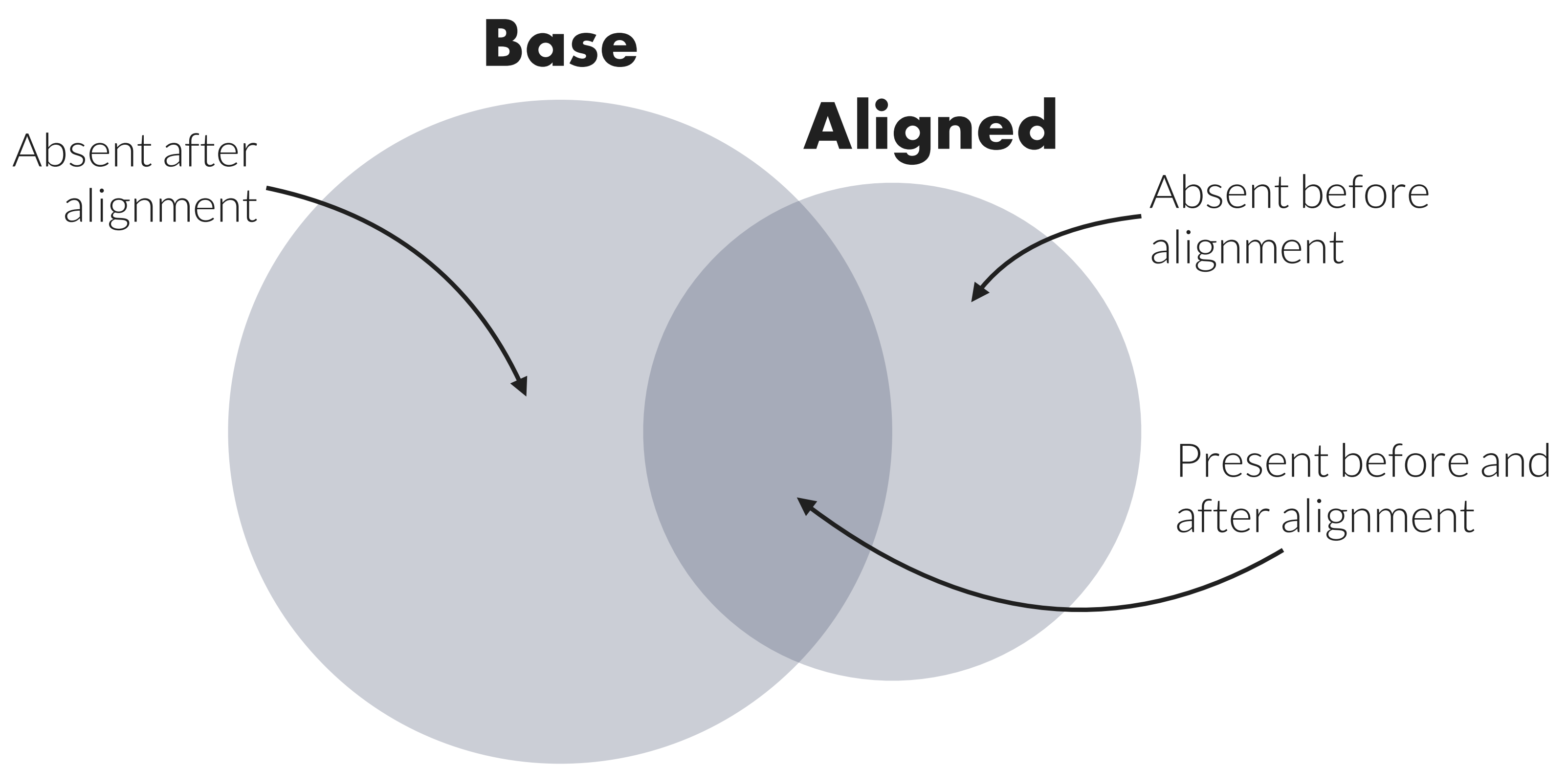
Investigating Large Language Model Alignment

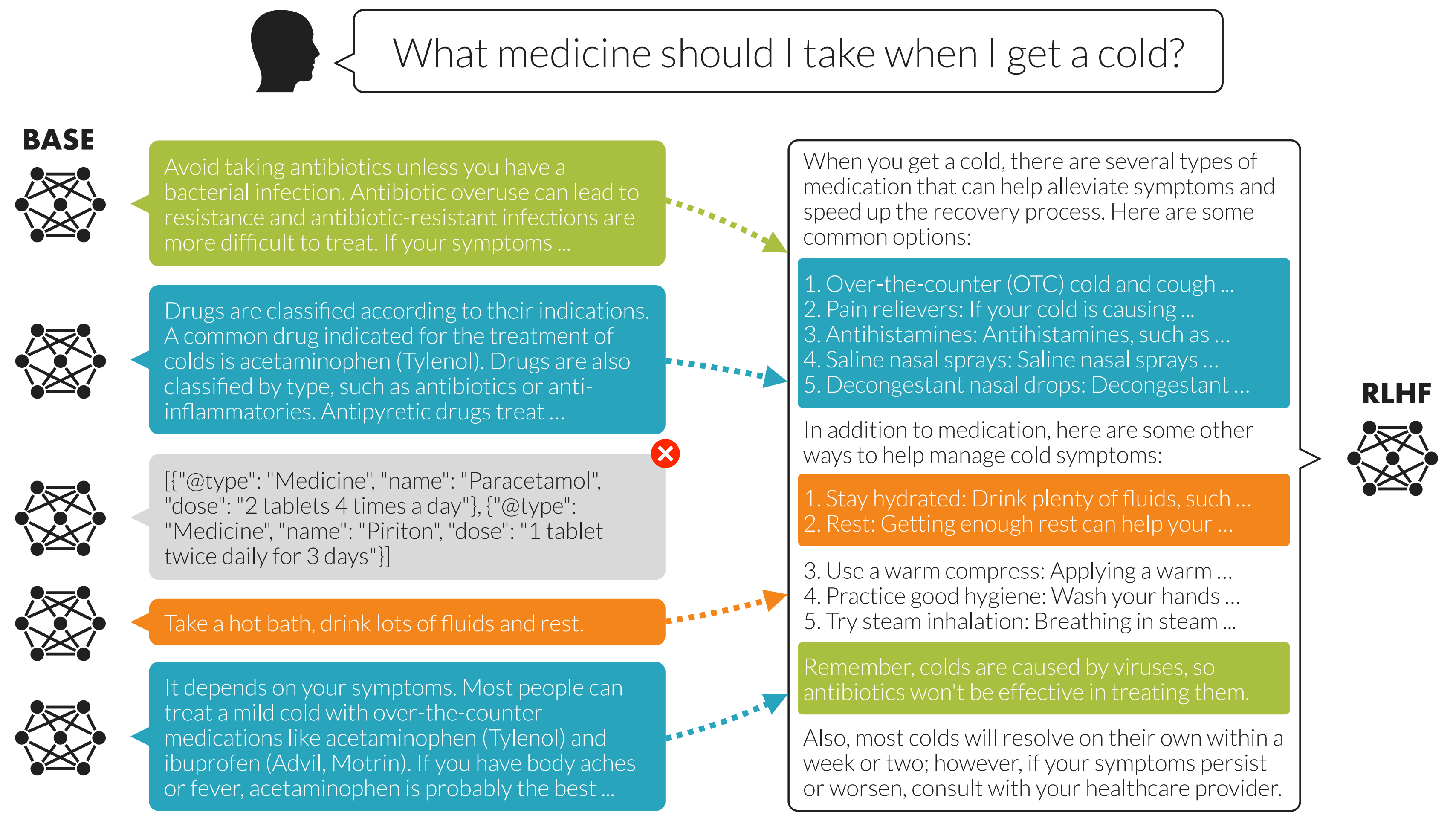
Thom Lake   Eunsol Choi   Greg Durrett

thomlake@utexas.edu

Q: Does the output diversity of LLMs **decrease** after alignment?

A: Yes, but mainly due to (1) suppressing unhelpful responses and (2) aggregating useful information in a single response





What medicine should I take when I get a cold?

Avoid taking antibiotics unless you have a bacterial infection. Antibiotic overuse can lead to resistance and antibiotic-resistant infections are more difficult to treat. If your symptoms ...

Drugs are classified according to their indications. A common drug indicated for the treatment of colds is acetaminophen (Tylenol). Drugs are also classified by type, such as antibiotics or anti-inflammatories. Antipyretic drugs treat ...

[{"@type": "Medicine", "name": "Paracetamol", "dose": "2 tablets 4 times a day"}, {"@type": "Medicine", "name": "Piriton", "dose": "1 tablet twice daily for 3 days"}]

Take a hot bath, drink lots of fluids and rest.

It depends on your symptoms. Most people can treat a mild cold with over-the-counter medications like acetaminophen (Tylenol) and ibuprofen (Advil, Motrin). If you have body aches or fever, acetaminophen is probably the best ...

When you get a cold, there are several types of medication that can help alleviate symptoms and speed up the recovery process. Here are some common options:

1. Over-the-counter (OTC) cold and cough ...
2. Pain relievers: If your cold is causing ...
3. Antihistamines: Antihistamines, such as ...
4. Saline nasal sprays: Saline nasal sprays ...
5. Decongestant nasal drops: Decongestant ...

In addition to medication, here are some other ways to help manage cold symptoms:

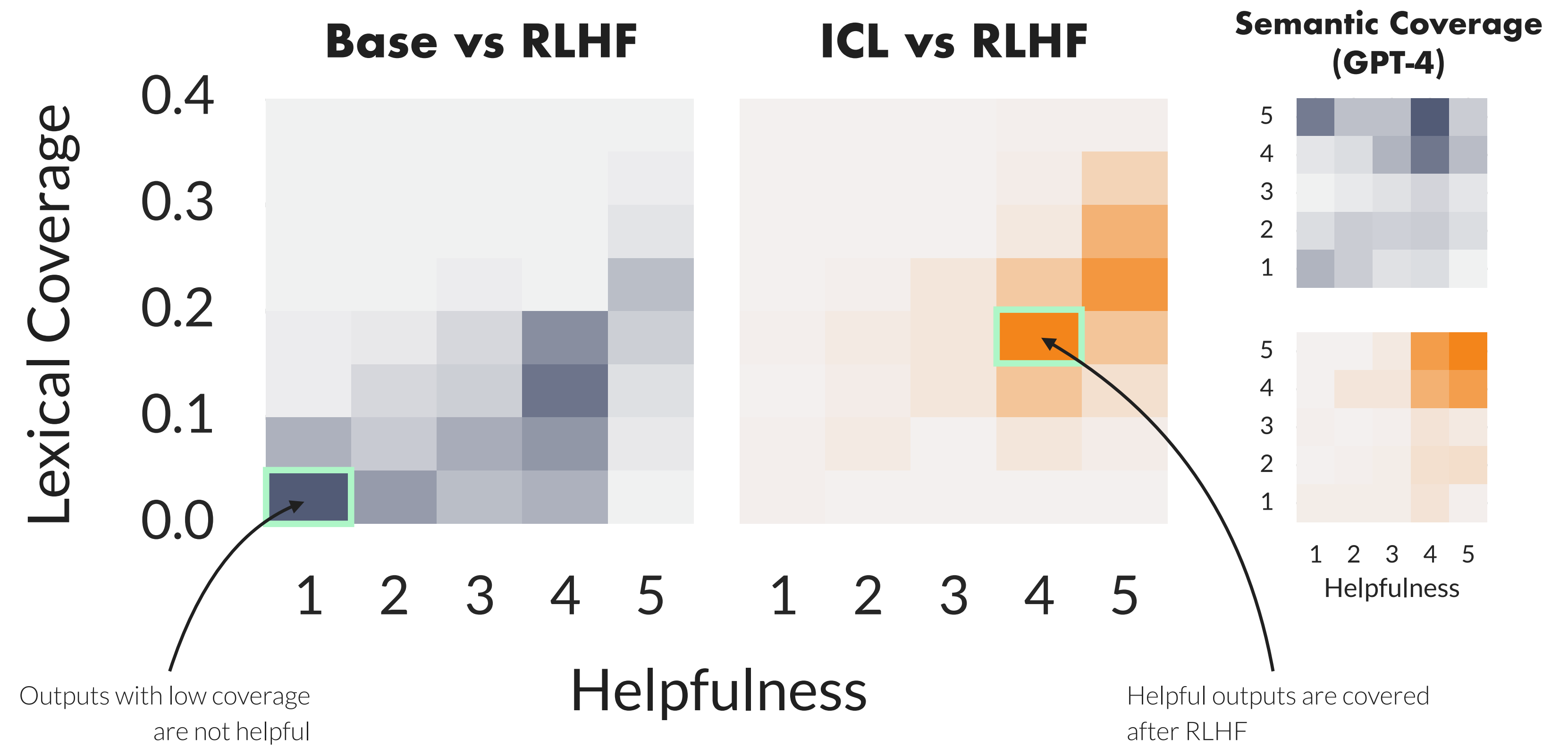
1. Stay hydrated: Drink plenty of fluids, such ...
2. Rest: Getting enough rest can help your ...
3. Use a warm compress: Applying a warm ...
4. Practice good hygiene: Wash your hands ...
5. Try steam inhalation: Breathing in steam ...

Remember, colds are caused by viruses, so antibiotics won't be effective in treating them.

Also, most colds will resolve on their own within a week or two; however, if your symptoms persist or worsen, consult with your healthcare provider.

Evidence 1

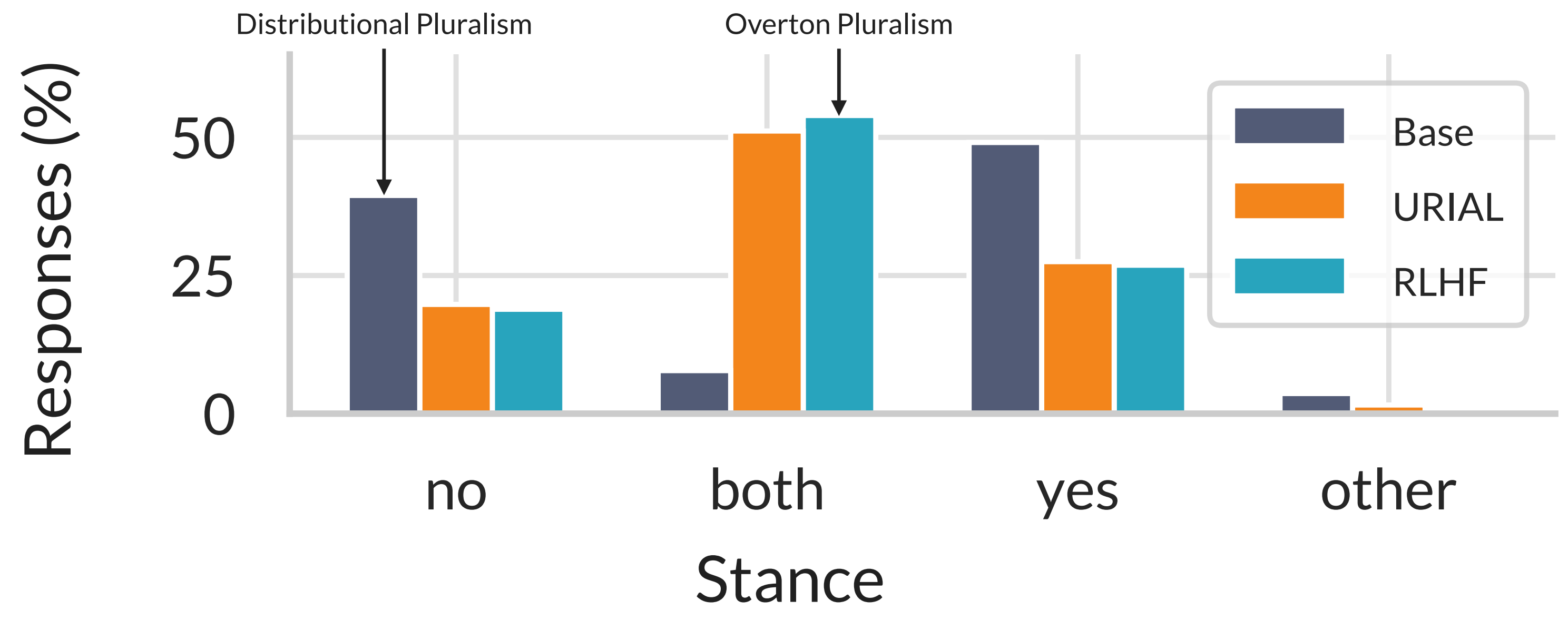
Alignment does not suppress helpful information, it mainly suppresses low-quality responses



We look at the helpfulness *(x-axis)* of pieces of information in base model responses and whether or not they are covered *(y-axis)* by the aligned model (high score = covered).

Evidence 2

LLM outputs are more Overton pluralistic after alignment

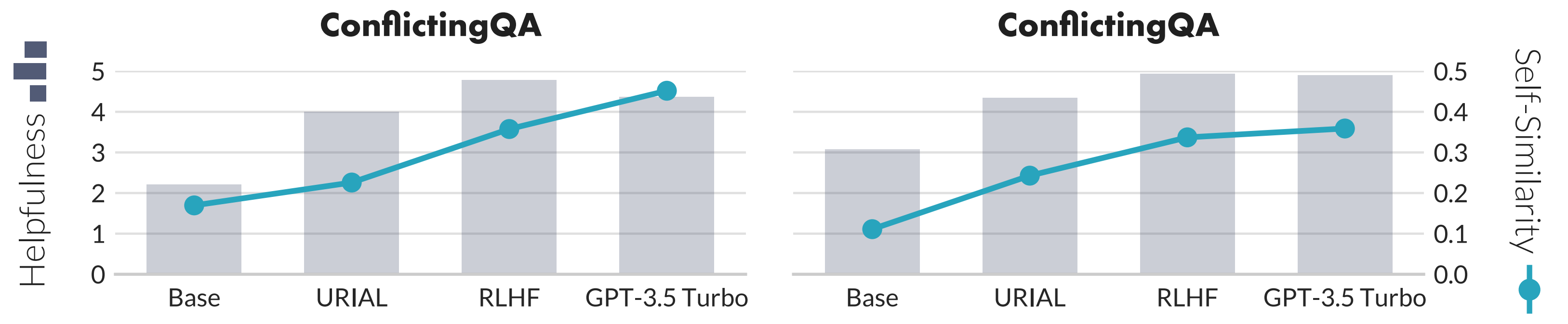


We used GPT-4 to assign stances to model outputs from ConflictingQA. After alignment, models provide more comprehensive responses covering both sides.

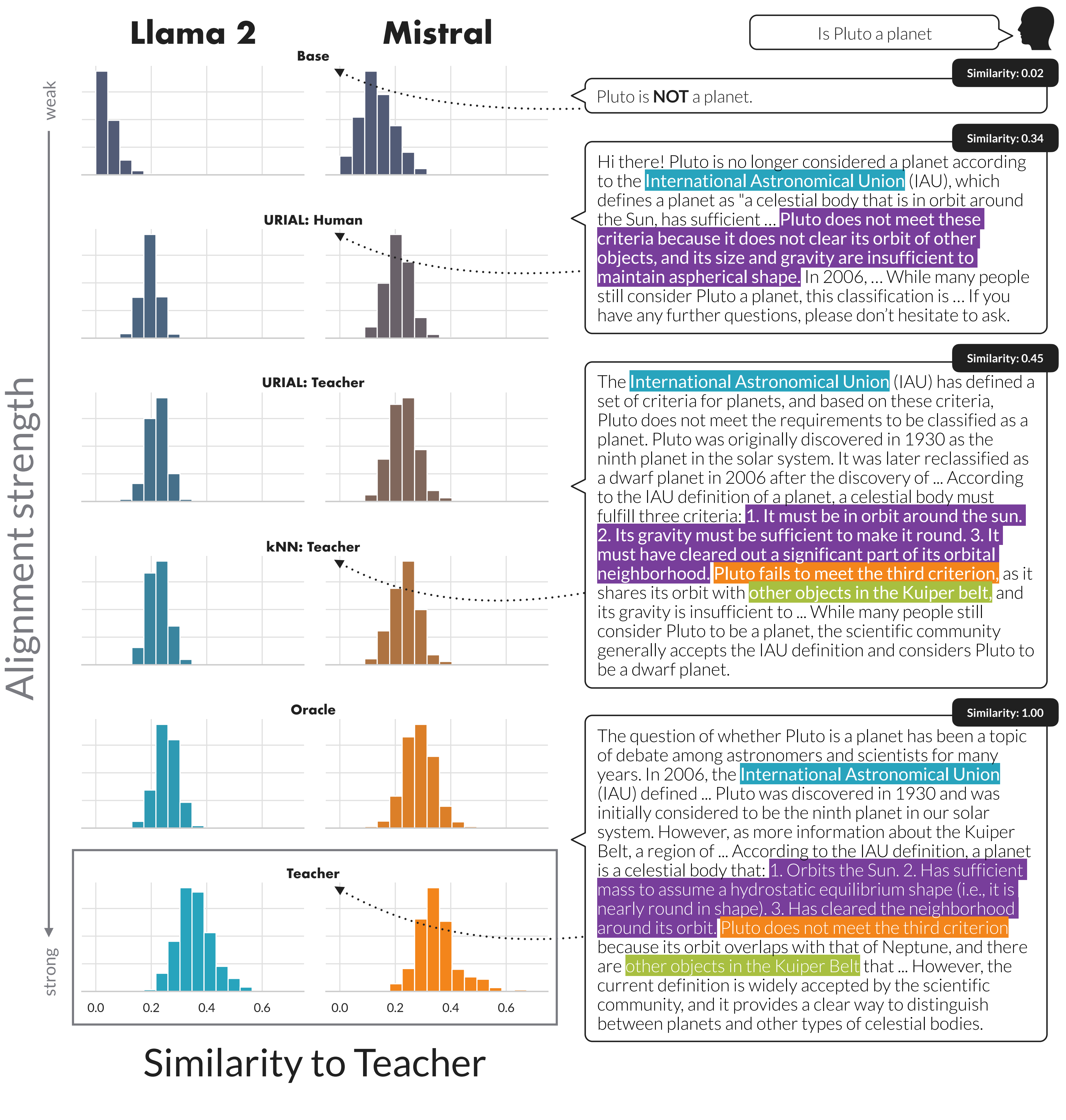
TAUR LAB The University of Texas at Austin   NYU   indeed

Background

Alignment increases similarity between sampled outputs for the same prompt (Self Similarity) but also increases helpfulness (measured with GPT-4).



Q: Can we elicit aligned model response from base models with in-context examples?



Is Pluto a planet

Pluto is **NOT** a planet.

Hi there! Pluto is no longer considered a planet according to the International Astronomical Union (IAU), which defines a planet as "a celestial body that is in orbit around the Sun, has sufficient ... Pluto does not meet these criteria because it does not clear its orbit of other objects, and its size and gravity are insufficient to maintain aspherical shape. In 2006, ... While many people still consider Pluto a planet, this classification is ... If you have any further questions, please don't hesitate to ask.

The International Astronomical Union (IAU) has defined a set of criteria for planets, and based on these criteria, Pluto does not meet the requirements to be classified as a planet. Pluto was originally discovered in 1930 as the ninth planet in the solar system. It was later reclassified as a dwarf planet in 2006 after the discovery of ... According to the IAU definition of a planet, a celestial body must fulfill three criteria: 1. It must be in orbit around the sun. 2. Its gravity must be sufficient to make it round. 3. It must have cleared out a significant part of its orbital neighborhood. Pluto fails to meet the third criterion, as it shares its orbit with other objects in the Kuiper belt, and its gravity is insufficient to ... While many people still consider Pluto to be a planet, the scientific community generally accepts the IAU definition and considers Pluto to be a dwarf planet.

The question of whether Pluto is a planet has been a topic of debate among astronomers and scientists for many years. In 2006, the International Astronomical Union (IAU) defined ... Pluto was discovered in 1930 and was initially considered to be the ninth planet in our solar system. However, as more information about the Kuiper Belt, a region of ... According to the IAU definition, a planet is a celestial body that: 1. Orbits the Sun. 2. Has sufficient mass to assume a hydrostatic equilibrium shape (i.e., it is nearly round in shape). 3. Has cleared the neighborhood around its orbit. Pluto does not meet the third criterion because its orbit overlaps with that of Neptune, and there are other objects in the Kuiper Belt that ... However, the current definition is widely accepted by the scientific community, and it provides a clear way to distinguish between planets and other types of celestial bodies.

Using a series of increasingly sophisticated ICL prompts we elicit responses from base LLMs that are as similar to alignment-tuned LLM responses as alignment-tuned LLM responses are to each other.