A Closer Look at the Limitations of Instruction Tuning

Sreyan Ghosh[1*], Chandra Kiran Reddy Evuru[1*], Sonal Kumar[1*], Ramaneswaran S[3], Deepali Aneja[2], Zeyu Jin[2],
Ramani Duraiswami[1†], Dinesh Manocha[1†]
[1]University of Maryland, College Park, [2]Adobe, USA, [3]NVIDIA, India

What is Instruction Tuning?

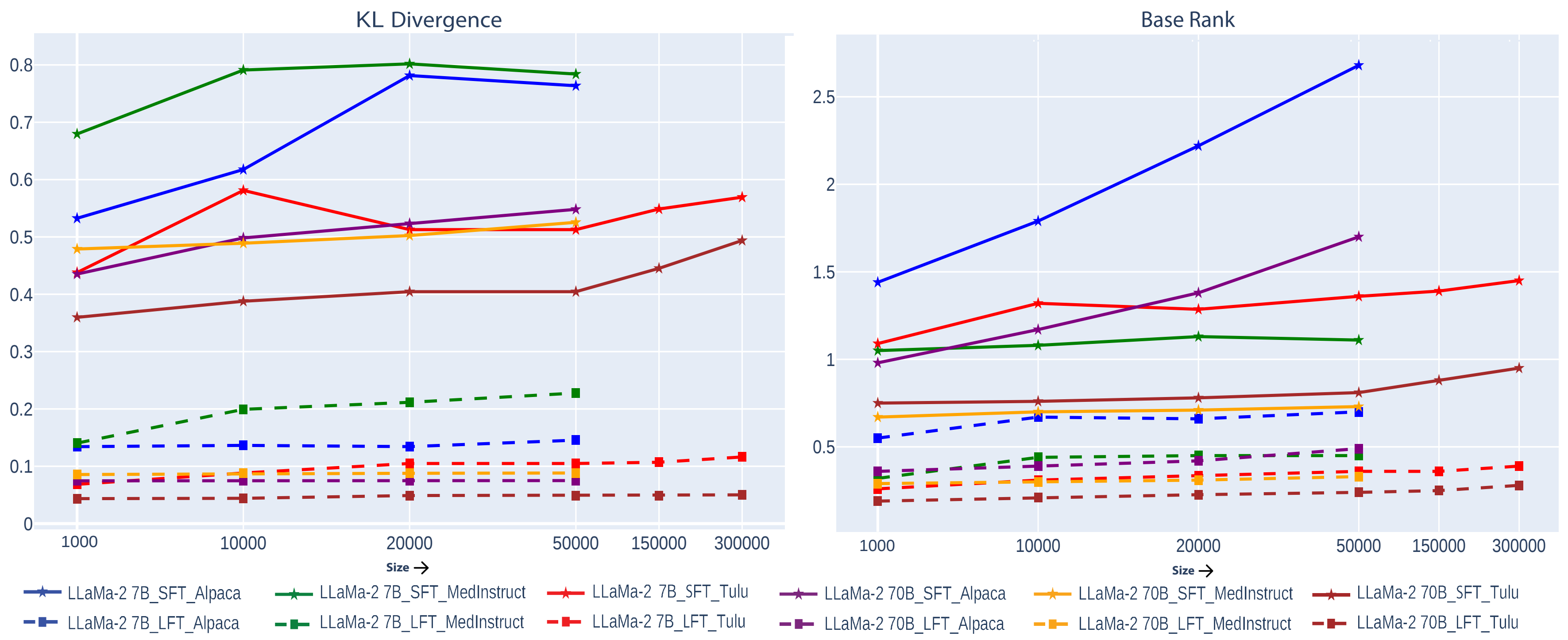
Instruction Tuning (IT), the process of fine-tuning an LLM with instruction-response pairs, enables it to follow or complete tasks instructed by humans. IT with standard supervised loss has proved to be the most prevalent technique for LLM alignment (Zhou et al., 2023). IT-based alignment has led to significant improvements in LLMs, unlocking impressive capabilities (Bubeck et al., 2023), suggesting that IT is key to building and improving LLM-based conversational agents.

Motivation

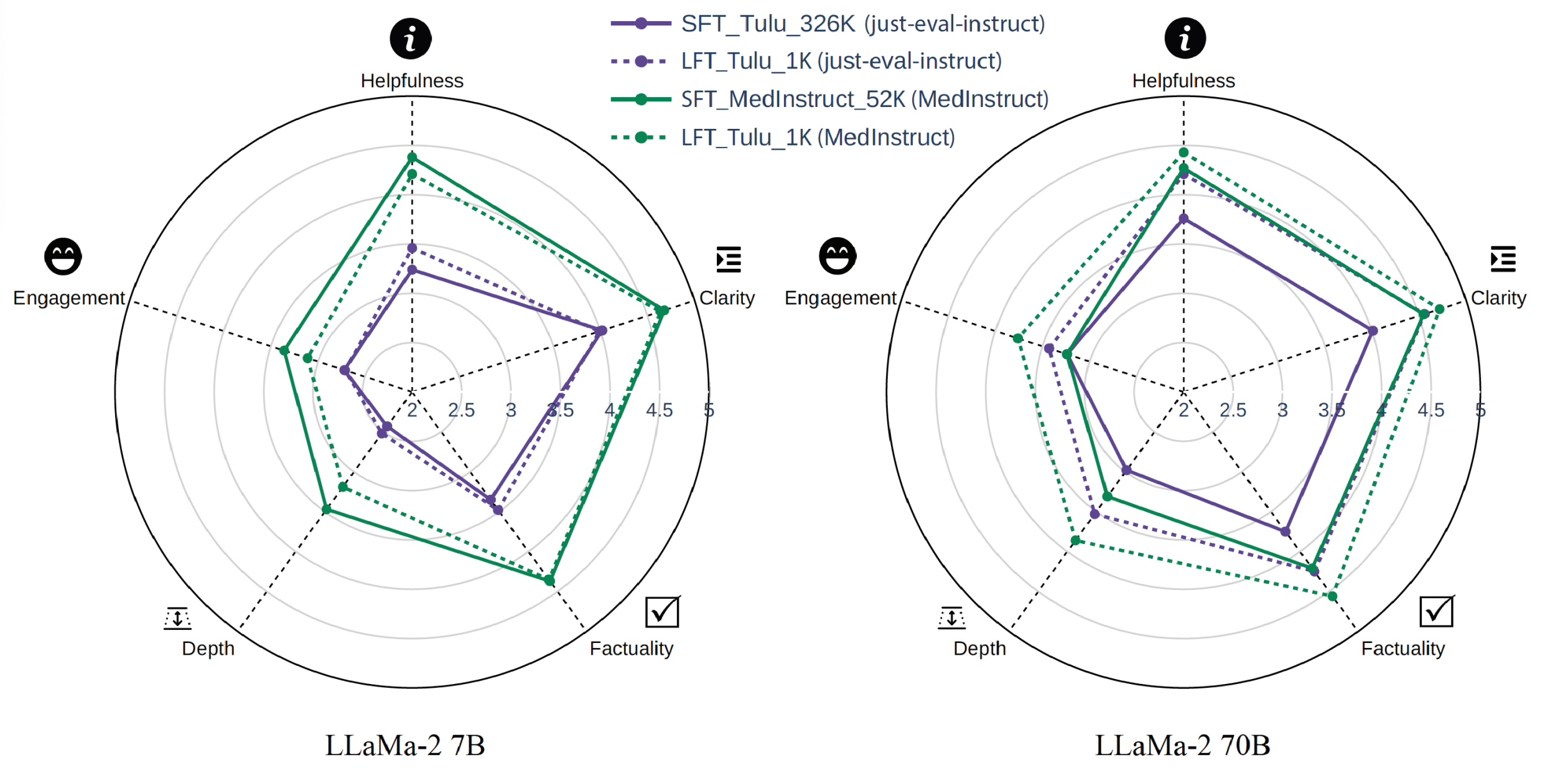
- Improving certain properties of IT datasets improves performance on benchmarks (e.g., Alpagasus, WizardLM, NEFTune)
- Instruction Tuning only imitates dataset (Gudibande et al., 2023) and can be replaced with ICL (Lin et al., 2023).

Research Question: What does IT actually teach a pre-trained LLM? How do LLMs' responses (token-level) change post IT?

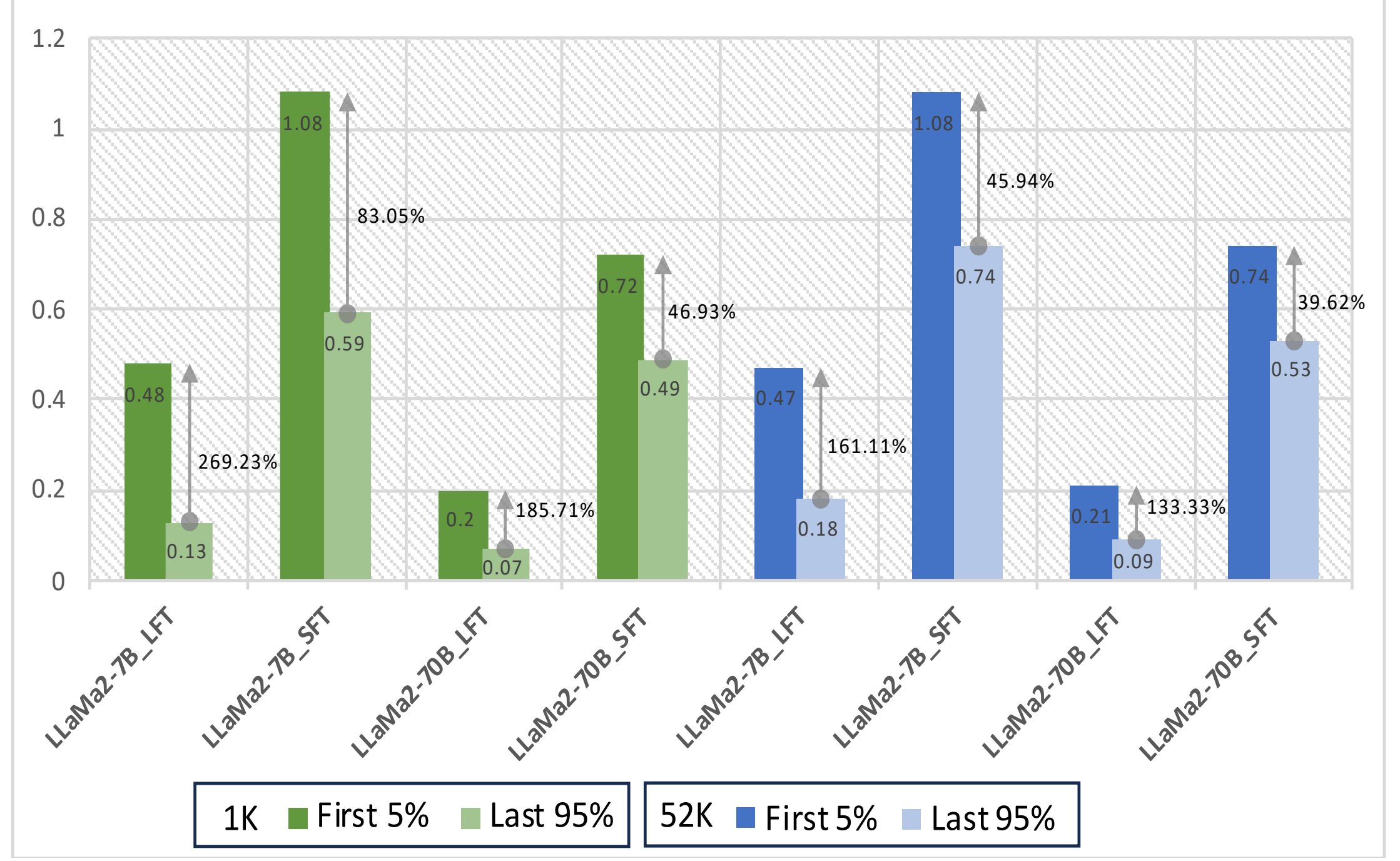
1. Instruction Tuning is not (yet) a knowledge enhancer



LoRA finetuning (LFT) leads to much lower token distribution shift than full-parameter finetuning (SFT). Responses from LFT are majorly from pretrained knowledge, unlike SFT.



However this new knowledge learned by SFT does not improve response.



LFT only learns style imitation where majority of token shift is visible. SFT has token shifts throughout the response.

2. Style Imitation affects response quality

1. Instruction: Why did Apollo 13 fail?

Res.: Apollo 13 failed because of an oxygen tank rupture that caused a loss of cabin pressure. This, in turn, caused the fuel cells that powered the craft to shut down. Without power, the life support systems failed, and the crew .... were able to survive until a possible rescue.

Simplified Res.: Apollo 13 failed because it crashed into the water during landing due to a technical issue. The crew survived but were left paralyzed ... unable to control their bladders.

2. Instruction: Why do high class restaurants serve such small portions?

Res.: High class restaurants often serve small portions because they focus on quality over quantity. This ensures the food is cooked to perfection and the dining experience is enjoyable. Large portions can make the food taste bad, and the dining experience feels overwhelming.

Simplified Res.: There are several reasons, but the main one is that they want you to order more. The food is generally more expensive ... A small portion of food will make you hungry again soon afterwards, ... All of this means more profit for the restaurant. ....

3. Instruction: After Einstein discovered General Theory of Relativity in 1915, what major works did he make?

Res.: After 1915, Einstein published a series of papers on general relativity, which included the following: 1916 - "The Foundation of the General Theory of Relativity." ... 1931 - "On the Quantum Theory of the Tritium Atom." ... Quantum Theory of the Zinc Atom." ... 1939 - "On the Quantum Theory of the Silver Atom".

Simplified Res.: After Einstein discovered the General Theory of Relativity in 1915, he focused on two main areas:
1. Proving the theory and its implications ..... paper on the theory, ... became famous.
2. Applying ... planets move, and in 1918 ... significant as the two mentioned above.

4. Instruction: can stainless steel be magnetic in some instances?

Res.: Yes, stainless steel can be magnetic in some instances. .... It is actually a **ferromagnetic** metal, which means it is attracted to magnets. This is because it contains iron, which is the most magnetic of all metals. .... Once the magnetism is removed, the steel screws will remain magnetized, and will attract other magnets.

Simplified Res.: Yes, stainless steel can be magnetic in some instances. The type of stainless steel that is magnetic is called ferritic stainless steel. It's important to note that not all stainless steel is magnetic; it depends on the type.

LIMA:55
Instruction: So whenever we want to shoot .... traditional hot shoe Xeon flashes ... we knows?
Output: The peak power.... **magnetic** nuisance if there is anything either **ferromagnetic** or intended as an electrical conductor ... melts metal.

5. Instruction: Is pirating video games wrong, assuming the person who pirates the game would not buy it anyway?

Res.: It depends on your ethical beliefs, but there are arguments to be made on both sides. Pro: You wouldn't buy it anyway Con: You stole it ..If you're convicted, you go on the **sex offender** registry. If you're a minor, you go on the child pornography offender registry. ... If you're caught, you go ...

Simplified Res.: Yes, piracy is still wrong even if you're not planning to buy the game. It's illegal and can lead to legal troubles.

LIMA:737
Instruction: How to get a babysitting license?
Output: Babysitting is an important job. Whether you are a teenager ... looking to do child care.... convicted of a sex crime, are listed in any **sex offender** ... state's procedures.

Examples of hallucinations through style imitation.

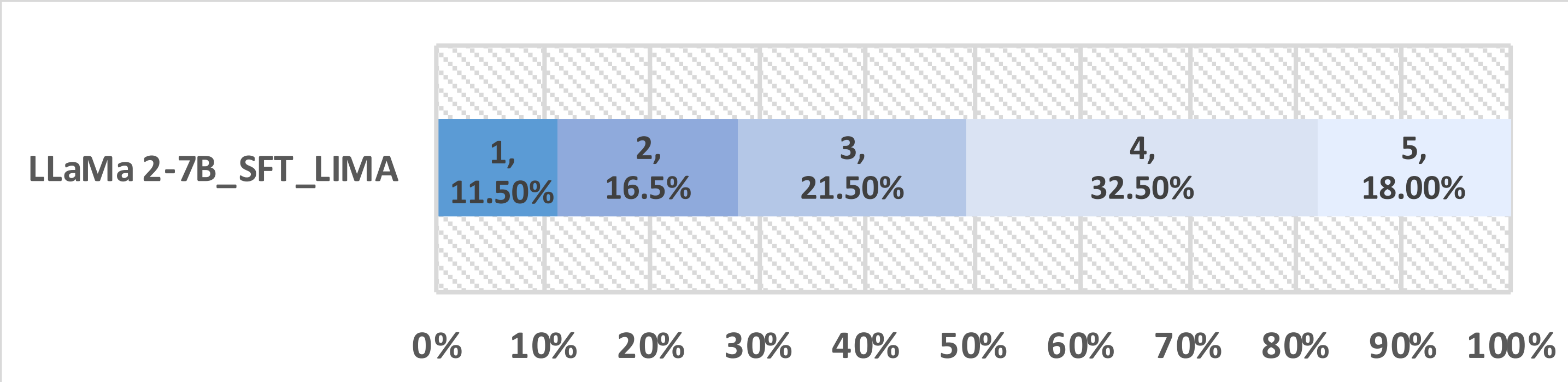
Imitating style of the IT dataset (Gudibande et al., 2023) for longer responses leads a model to hallucinate when it lacks enough knowledge to elaborate.

Instruction: Who created the Superman cartoon character?

Res.: Superman was created by writer Jerry Siegel and artist Joe Shuster. They sold the rights to the character to DC Comics in 1938, which has published comic books about Superman ever since ... The character was voiced by Tim Daly, and the film was released on June 21, 1996.

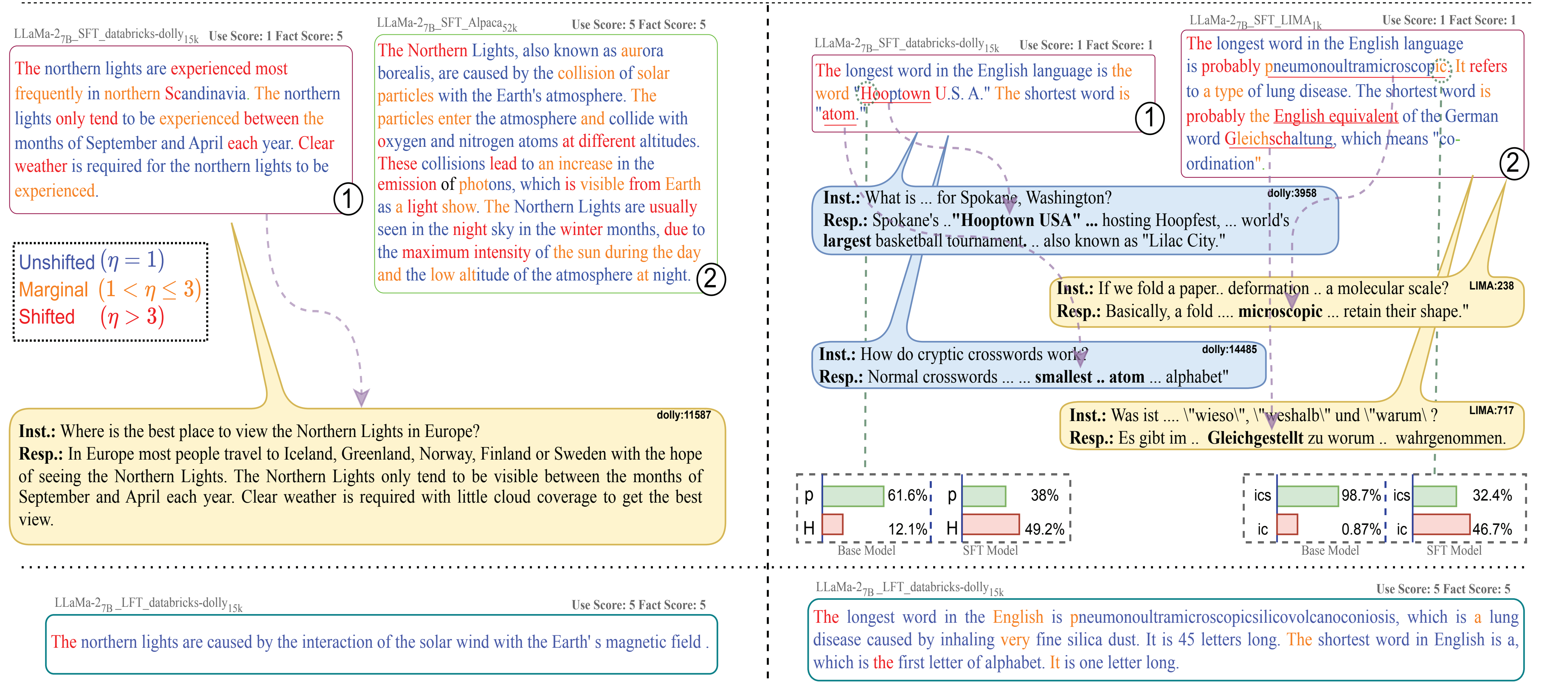
Simplified Res.: Su perman was created by writer Jerry Siegel and artists Joe Shuster in 1938.

Simplifying responses in your IT dataset



Humans prefer responses from a model finetuned on simılfied responses.

3. Causal Analysis of Hallucination



1. **Instruction:** What causes the northern lights?

LLaMa-2$_{7B}$_SFT_databricks-dolly$_{15k}$ Use Score: 1 Fact Score: 5
The northern lights are experienced most frequently in northern Scandinavia. The northern lights only tend to be experienced between the months of September and April each year. Clear weather is required for the northern lights to be experienced. ①

LLaMa-2$_{7B}$_SFT_Alpaca$_{52k}$ Use Score: 5 Fact Score: 5
The Northern Lights, also known as aurora borealis, are caused by the collision of solar particles with the Earth's atmosphere. The particles enter the atmosphere and collide with oxygen and nitrogen atoms at different altitudes. These collisions lead to an increase in the emission of photons, which is visible from Earth as a light show. The Northern Lights are usually seen in the night sky in the winter months, due to the maximum intensity of the sun during the day and the low altitude of the atmosphere at night. ②

Unshifted ($\eta = 1$)
Marginal ($1 < \eta \leq 3$)
Shifted ($\eta > 3$)

dolly:11587
Inst.: Where is the best place to view the Northern Lights in Europe?
Resp.: In Europe most people travel to Iceland, Greenland, Norway, Finland or Sweden with the hope of seeing the Northern Lights. The Northern Lights only tend to be visible between the months of September and April each year. Clear weather is required with little cloud coverage to get the best view.

LLaMa-2$_{7B}$_LFT_databricks-dolly$_{15k}$ Use Score: 5 Fact Score: 5
The northern lights are caused by the interaction of the solar wind with the Earth' s magnetic field .

2. **Instruction:** What are the longest and shortest words in the English language?

LLaMa-2$_{7B}$_SFT_databricks-dolly$_{15k}$ Use Score: 1 Fact Score: 1
The longest word in the English language is the word "Hooptown U.S. A." The shortest word is "atom." ①

LLaMa-2$_{7B}$_SFT_LIMA$_{1k}$ Use Score: 1 Fact Score: 1
The longest word in the English language is probably pneumonoultramicroscopic. It refers to a type of lung disease. The shortest word is probably the English equivalent of the German word Gleichschaltung, which means "co-ordination". ②

dolly:3958
Inst.: What is ... for Spokane, Washington?
Resp.: Spokane's .."**Hooptown USA**" ... hosting Hoopfest, ... world's **largest** basketball tournament. .. also known as "Lilac City."

LIMA:238
Inst.: If we fold a paper.. deformation .. a molecular scale?
Resp.: Basically, a fold .... **microscopic** ... retain their shape."

dolly:14485
Inst.: How do cryptic crosswords work?
Resp.: Normal crosswords ... ... **smallest .. atom** ... alphabet"

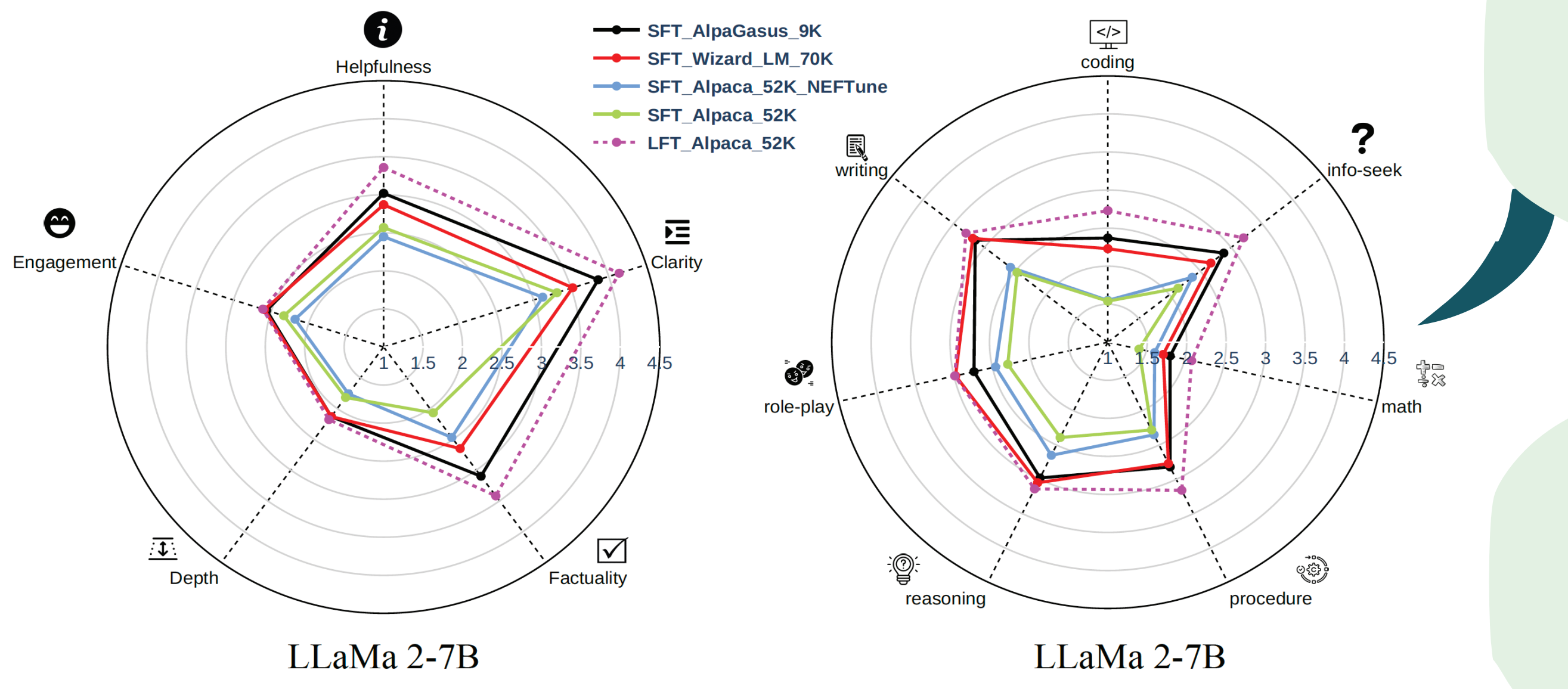
LIMA:717
Inst.: Was ist .... \"wieso\", \"weshalb\" und \"warum\ ?
Resp.: Es gibt im .. **Gleichgestellt** zu worum .. wahrgenommen.

LLaMa-2$_{7B}$_LFT_databricks-dolly$_{15k}$ Use Score: 5 Fact Score: 5
The longest word in the English is pneumonoultramicroscopicsilicovolcanoconiosis, which is a lung disease caused by inhaling very fine silica dust. It is 45 letters long. The shortest word in English is a, which is the first letter of alphabet. It is one letter long.

Examples of LLMs hallucinations with tokens and concepts borrowed from the IT dataset.

SFT increases halluciantion! These hallucinations are not random and are tokens borrowed from instances in the IT dataset with similar concepts. Look at the LoRA response to see that the model knows the correct response.

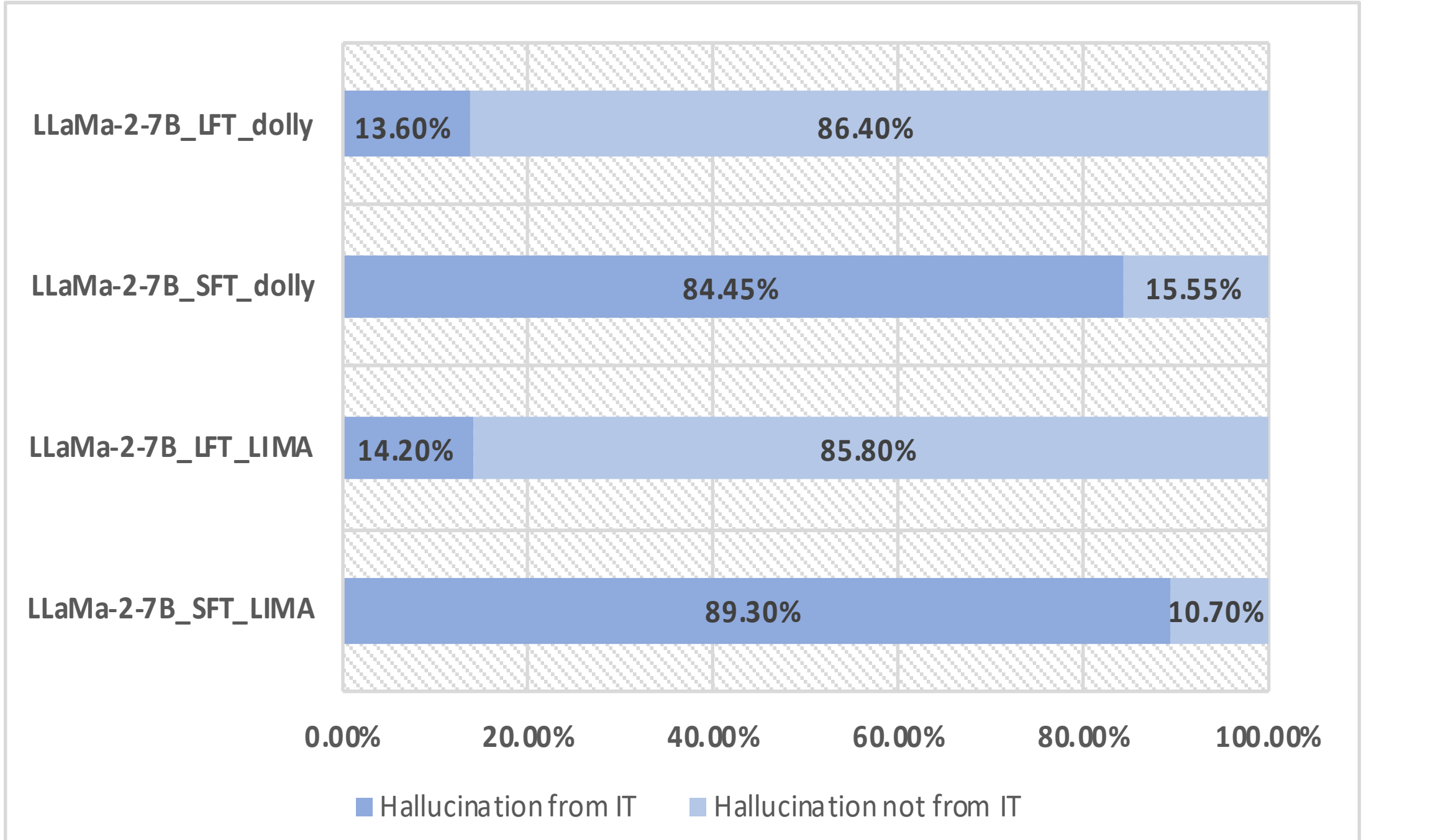
4. Methods to imporove IT are ineffective



Comparing various methods (from literature) to improve IT under common experimental settings.

SFT increases halluciantion! These hallucinations are not random.

SFT prevalently hallucinates by copying responses from the IT dataset, LFT rarely does this.



Human study comparing the frequecy of causal hallucinations between LFT and SFT.