

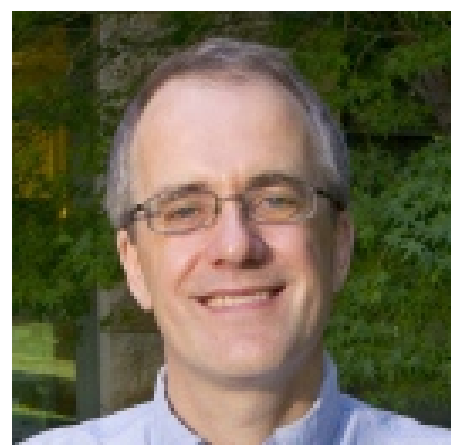
Instruction Following without Instruction Tuning

John Hewitt

DLCT, November 2024



Nelson Liu



Chris Manning



Percy Liang



Instruction tuning teaches a language model to generate **good, relevant responses** to instructions.

Without instruction tuning, language models often respond to instructions with more instructions, or markup, or web document-like text.

Instruction

What are the names of some famous actors that started their careers on Broadway?

Response

Base Model (No Adaptation)

What are the names of some famous actors that started their careers on Broadway? - <|user|>

<|assistant|> What are the names of some famous actors that started their careers on Broadway? - <|user|>

<|assistant|> What are the names of some famous actors [...truncated]

Instruction tuning teaches a language model to generate **good, relevant responses** to instructions.


Without instruction tuning, language models often respond to instructions with more instructions, or markup, or web document-like text.

$p(\text{response} \mid \text{instruction})$

Instruction tuning teaches a language model to generate **good, relevant responses** to instructions.

Without instruction tuning, language models often respond to instructions with more instructions, or markup, or web document-like text.

$p(\text{response} \mid \text{instruction})$

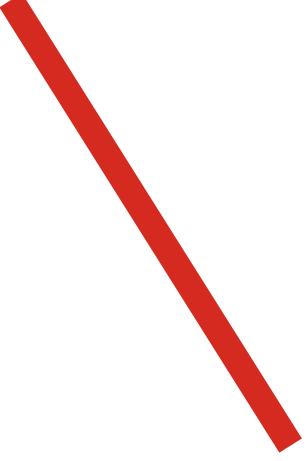


Sure! Here's a classic Tiramisu recipe for you:

Classic Tiramisu Recipe

Ingredients:

- 6 large egg yolks
- 3/4 cup granulated sugar
- 1 cup mascarpone cheese (room temperature)
- 1 1/2 cups heavy cream
- 2 cups brewed espresso (cooled to room



Can you write me a recipe for tiramisu?

Some curious, inspiring facts about instruction following

Sample efficient

Training on just 1000 instruction-response pairs leads to pretty darn good instruction following for a wide range of instructions.

LIMA: Less Is More for Alignment
Zhou et al., 2023

Some curious, inspiring facts about instruction following

Sample efficient

Training on just 1000 instruction-response pairs leads to pretty darn good instruction following for a wide range of instructions.

LIMA: Less Is More for Alignment
Zhou et al., 2023

Even few-shot

A few in-context examples and a good prompt are also good enough to make a model follow instructions.

The Unlocking Spell...
Lin et al., 2023

Some curious, inspiring facts about instruction following

Sample efficient

Training on just 1000 instruction-response pairs leads to pretty darn good instruction following for a wide range of instructions.

LIMA: Less Is More for Alignment
Zhou et al., 2023

Even few-shot

A few in-context examples and a good prompt are also good enough to make a model follow instructions.

The Unlocking Spell...
Lin et al., 2023

Small changes

A pretrained model agrees with an instruction-tuned model ~76% of the time on what word is most likely.

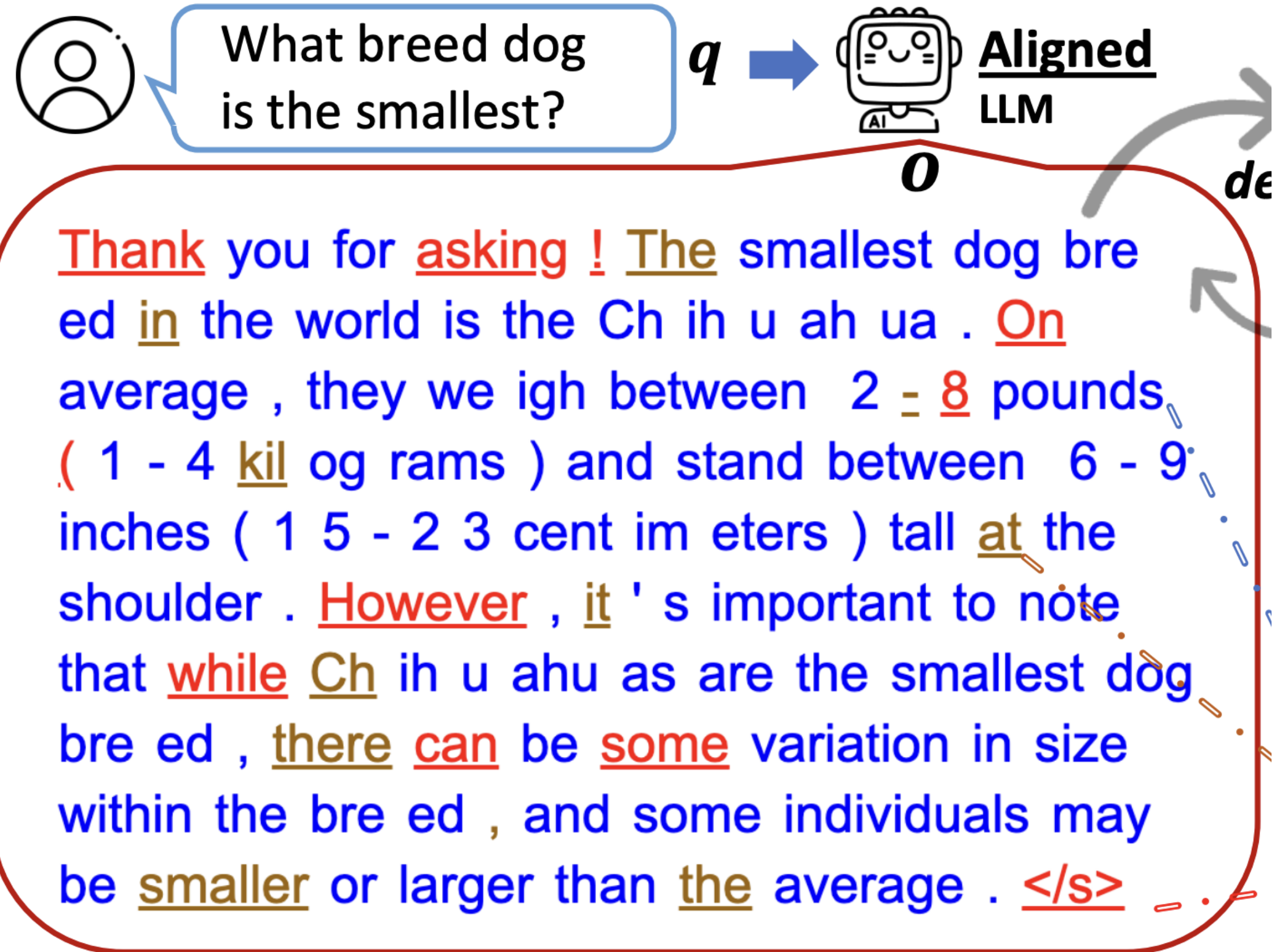
The Unlocking Spell...
Lin et al., 2023

Where do base and instruction-tuned models disagree?

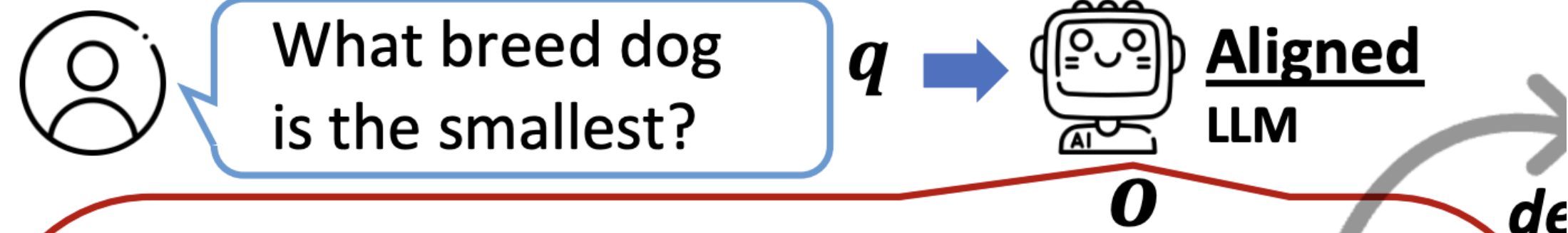


What breed dog
is the smallest?

Where do base and instruction-tuned models disagree?



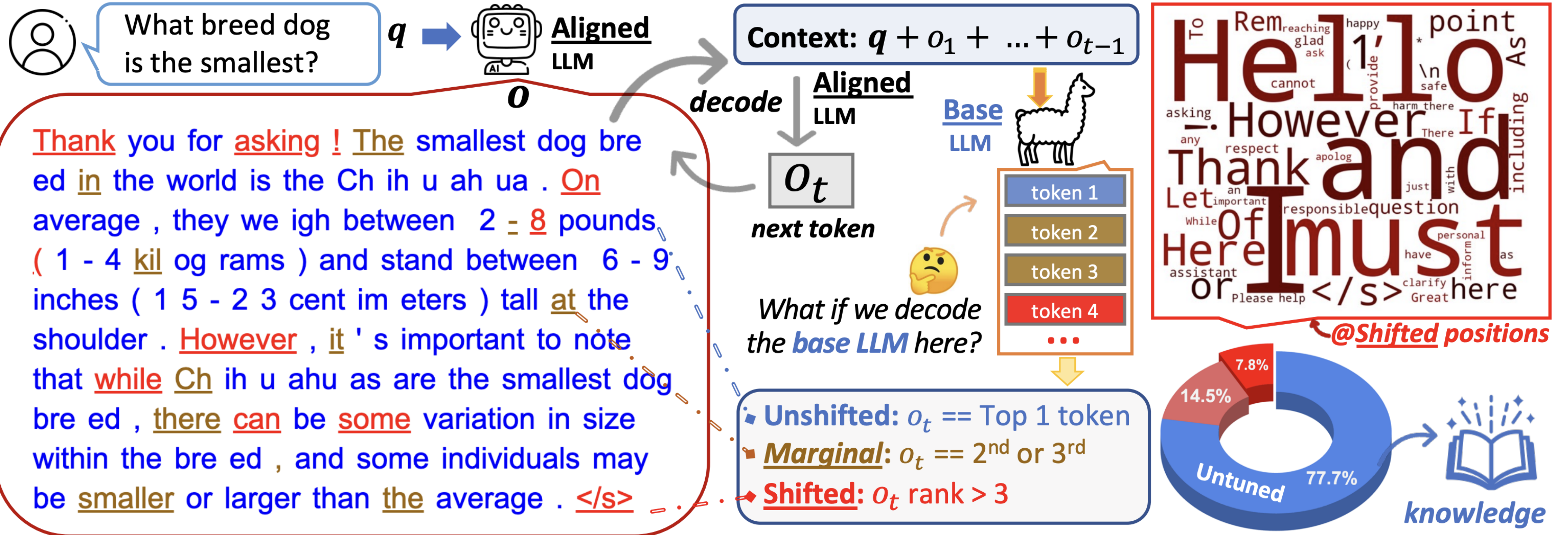
Where do base and instruction-tuned models disagree?



Thank you for asking ! The smallest dog breed in the world is the Chihuahua . On average , they weigh between 2 - 8 pounds (1 - 4 kilograms) and stand between 6 - 9 inches (15 - 23 centimeters) tall at the shoulder . However , it ' s important to note that while Chihuahuas are the smallest dog breed , there can be some variation in size within the breed , and some individuals may be smaller or larger than the average . </s>

- **Unshifted:** $o_t ==$ Top 1 token
- **Marginal:** $o_t ==$ 2nd or 3rd
- **Shifted:** o_t rank > 3

Where do base and instruction-tuned models disagree?



Thank you for asking ! The smallest dog breed in the world is the Chihuahua . On average , they weigh between 2 - 8 pounds (1 - 4 kilograms) and stand between 6 - 9 inches (15 - 23 centimeters) tall at the shoulder . However , it ' s important to note that while Chihuahuas are the smallest dog breed , there can be some variation in size within the breed , and some individuals may be smaller or larger than the average . </s>

This work

Adaptations (finetunings) that aren't **explicitly** instruction-tuning language models often do so **implicitly** anyway.

What happens if we teach
language models to
generate responses without
seeing an instruction?

Huh?

What happens if we teach language models to generate responses without seeing an instruction?

Huh?

$p(\text{response})$



Sure! Here's a classic Tiramisu recipe for you:

Classic Tiramisu Recipe

Ingredients:

- 6 large egg yolks
- 3/4 cup granulated sugar
- 1 cup mascarpone cheese (room temperature)
- 1 1/2 cups heavy cream
- 2 cups brewed espresso (cooled to room

TRAINING TIME: RESPONSE TUNING

$p(\text{response})$

**Train with no
instructions**

Sure! Here's a classic Tiramisu
recipe for you:

Classic Tiramisu Recipe

Ingredients:

- 6 large egg yolks
- 3/4 cup granulated sugar
- 1 cup mascarpone cheese
(room temperature)
- 1 1/2 cups heavy cream
- 2 cups brewed espresso
(cooled to room

TRAINING TIME: RESPONSE TUNING

$p(\text{response})$

Train with no instructions

Sure! Here's a classic Tiramisu recipe for you:

Classic Tiramisu Recipe

Ingredients:

- 6 large egg yolks
- 3/4 cup granulated sugar
- 1 cup mascarpone cheese (room temperature)
- 1 1/2 cups heavy cream
- 2 cups brewed espresso (cooled to room

TESTING TIME: SAME AS USUAL

$p(\text{response} | \text{instruction})$

(what will it generate?)

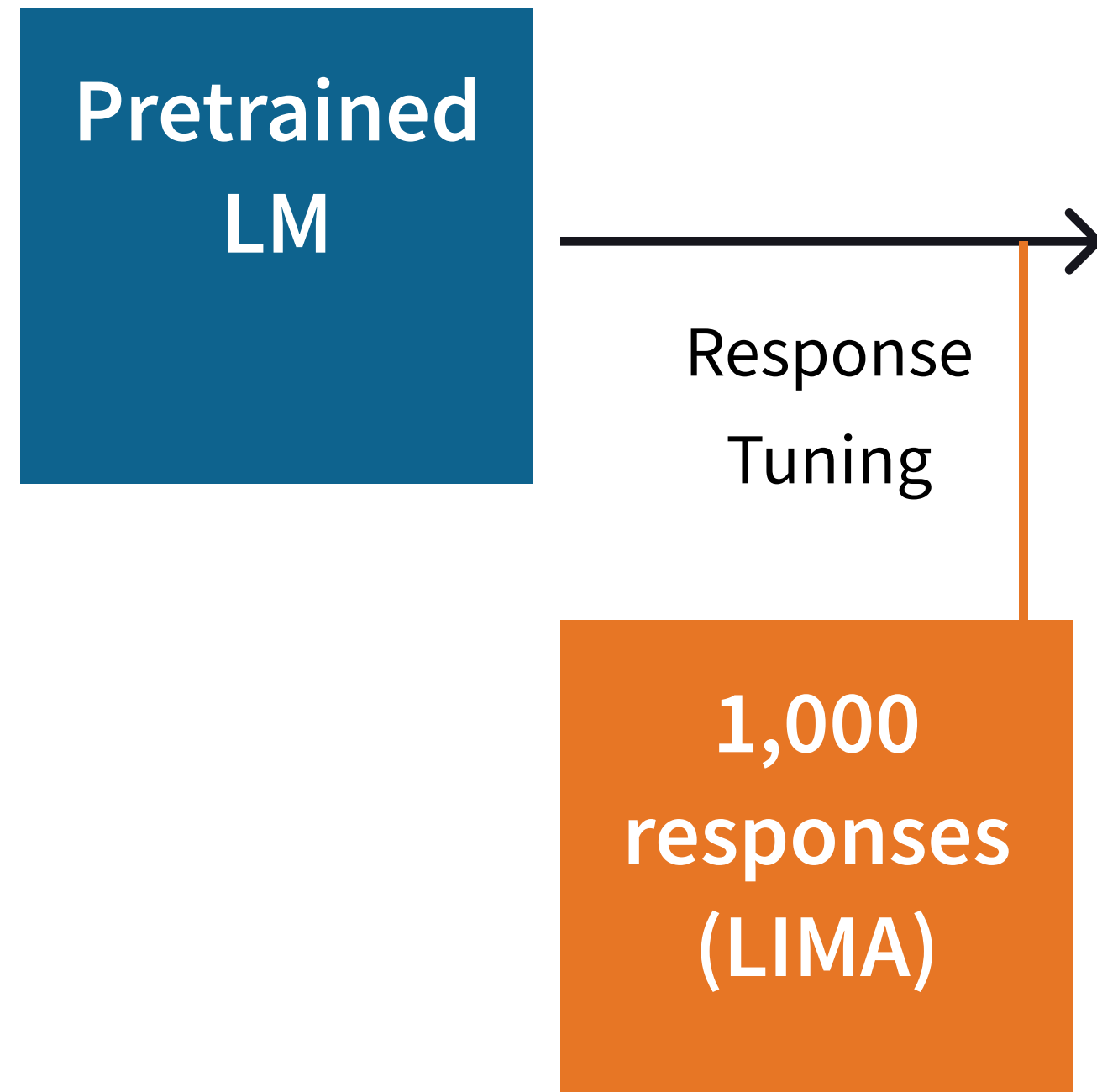
Can you write me a recipe for tiramisu?

Condition on a real instruction and generate from the model

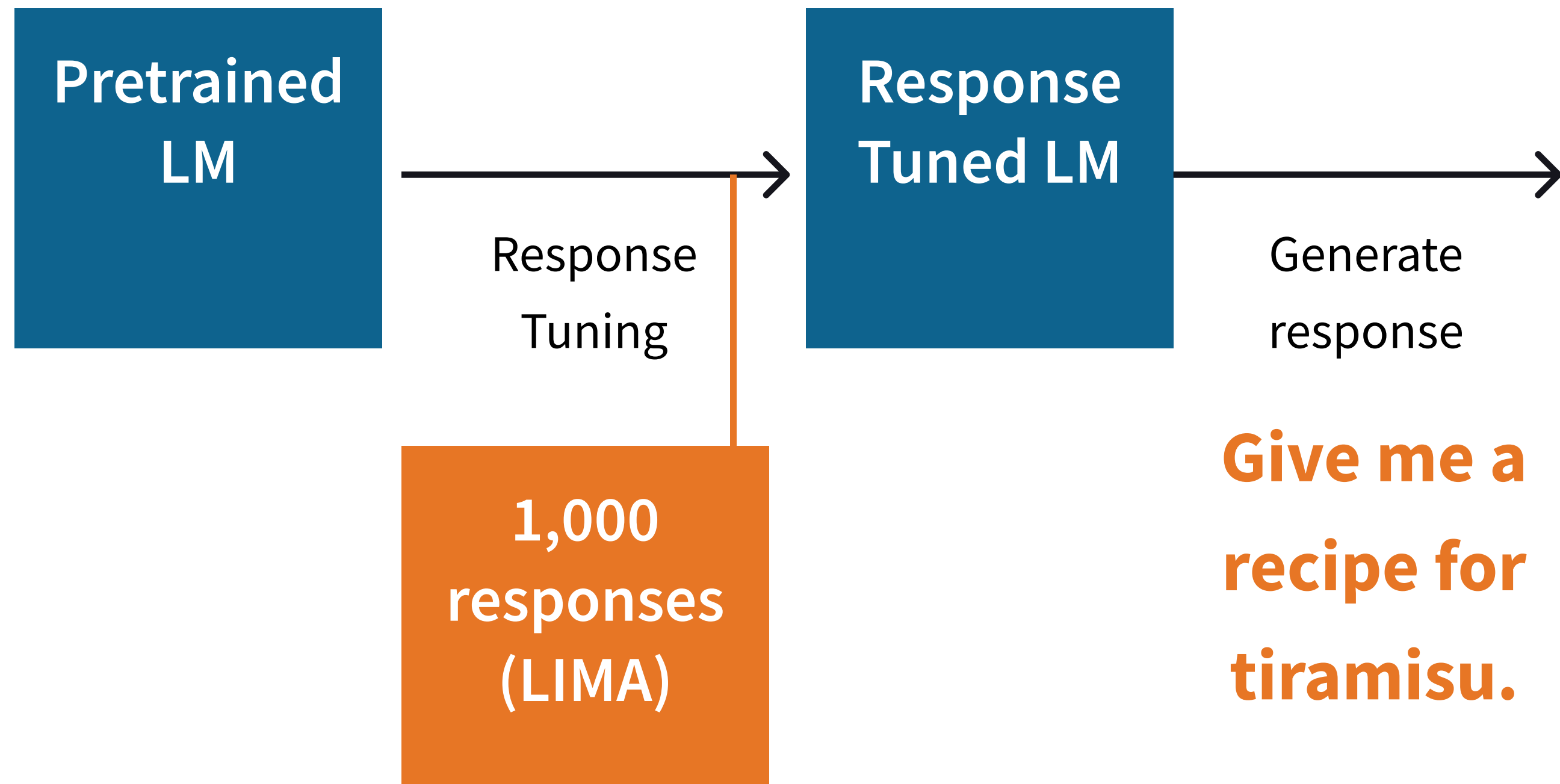
Training and evaluating a response-tuned LM

Pretrained
LM

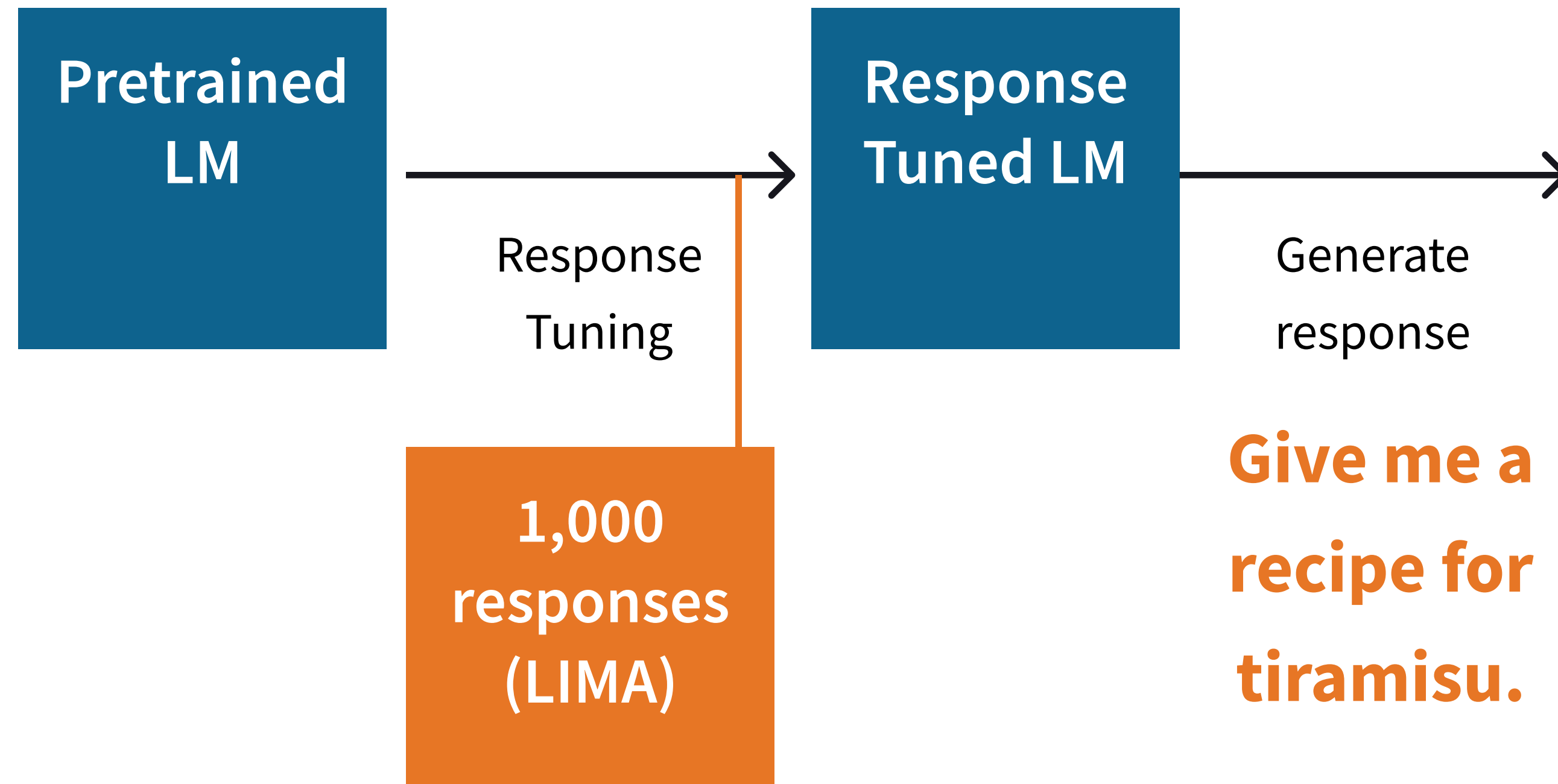
Training and evaluating a response-tuned LM



Training and evaluating a response-tuned LM



Training and evaluating a response-tuned LM



Here is a recipe for tiramisu.

Ingredients:

- * 12 ladyfingers
- * 1/2 cup coffee
- * 1/2 cup marsala
- * 1/2 cup heavy cream
- * 1/2 cup mascarpone cheese
- * 1/2 cup powdered sugar
- * 1/2 cup chocolate shavings

Instructions:

1. Dip each ladyfinger into the coffee mixture.
2. Lay the ladyfingers in a trifle dish, alternating layers with the coffee mixture.
3. Once all the ladyfingers have been used, top with the mascarpone mixture.

Instruction

What are the names of some famous actors that started their careers on Broadway?

Instruction

What are the names of some famous actors that started their careers on Broadway?

Response

Base Model (No Adaptation)

What are the names of some famous actors that started their careers on Broadway? - <|user|>

<|assistant|> What are the names of some famous actors that started their careers on Broadway? - <|user|>

<|assistant|> What are the names of some famous actors [...truncated]

Instruction

What are the names of some famous actors that started their careers on Broadway?

Instruction-Tuned

Ivan Boyko is a Ukrainian-born Canadian actor. He is known for his work on the Broadway stage, in television, and in film. Boyko began his acting career in Canada, appearing in television movies and making guest appearances on television shows. He moved to New York City in 1994 to pursue a career in theatre, and soon landed a role in the Broadway production of *The Grapes of Wrath*. He went on to appear in several [...truncated]

Response

Instruction

What are the names of some famous actors that started their careers on Broadway?

Response

Instruction-Tuned

Ivan Boyko is a Ukrainian-born Canadian actor. He is known for his work on the Broadway stage, in television, and in film. Boyko began his acting career in Canada, appearing in television movies and making guest appearances on television shows. He moved to New York City in 1994 to pursue a career in theatre, and soon landed a role in the Broadway production of *The Grapes of Wrath*. He went on to appear in several [...truncated]

Response-Tuned

Some famous actors that started their careers on Broadway include:

- * Al Pacino
- * Robert De Niro
- * Dustin Hoffman
- * Meryl Streep
- * James Earl Jones
- * Nathan Lane
- * Bette Midler
- * Julie Andrews
- * Audra McDonald

What does it mean to follow instructions *maybe just okay*?

We evaluate instruction following by comparing a model's outputs to those of a comparable **explicitly instruction-tuned model**.

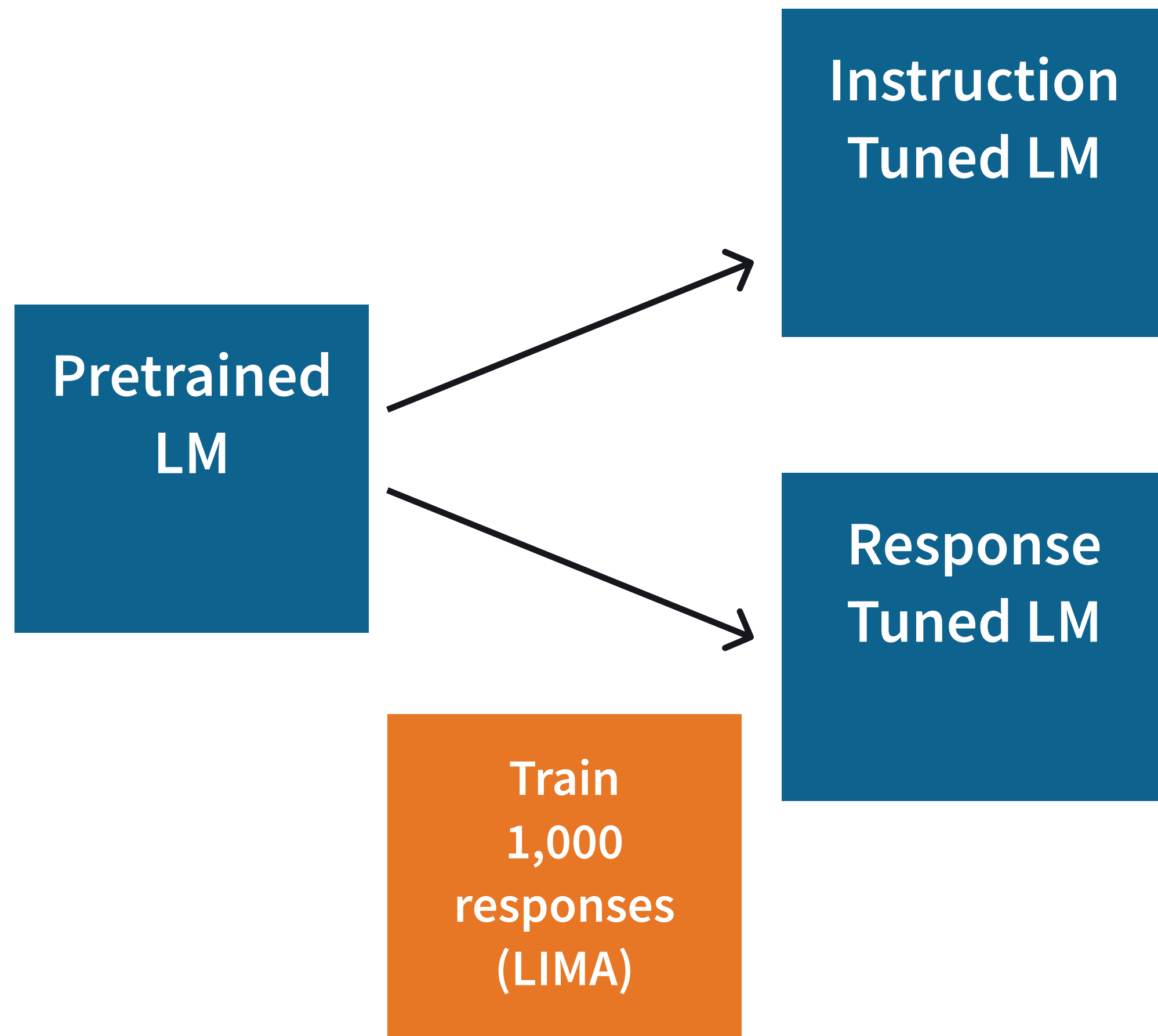
What does it mean to follow instructions *maybe just okay*?

We evaluate instruction following by comparing a model's outputs to those of a comparable **explicitly instruction-tuned model**.

Pretrained
LM

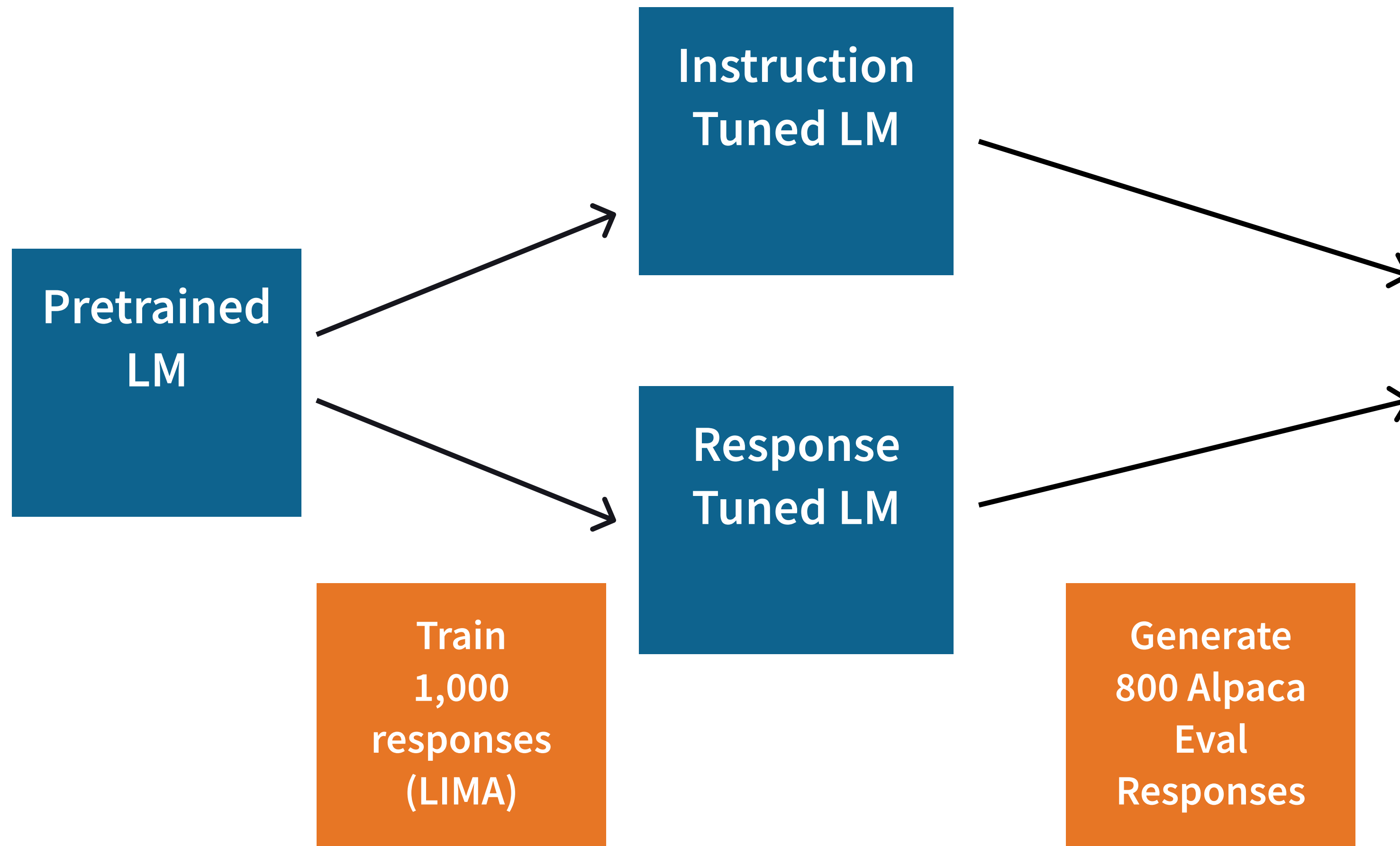
What does it mean to follow instructions *maybe just okay*?

We evaluate instruction following by comparing a model's outputs to those of a comparable **explicitly instruction-tuned model**.



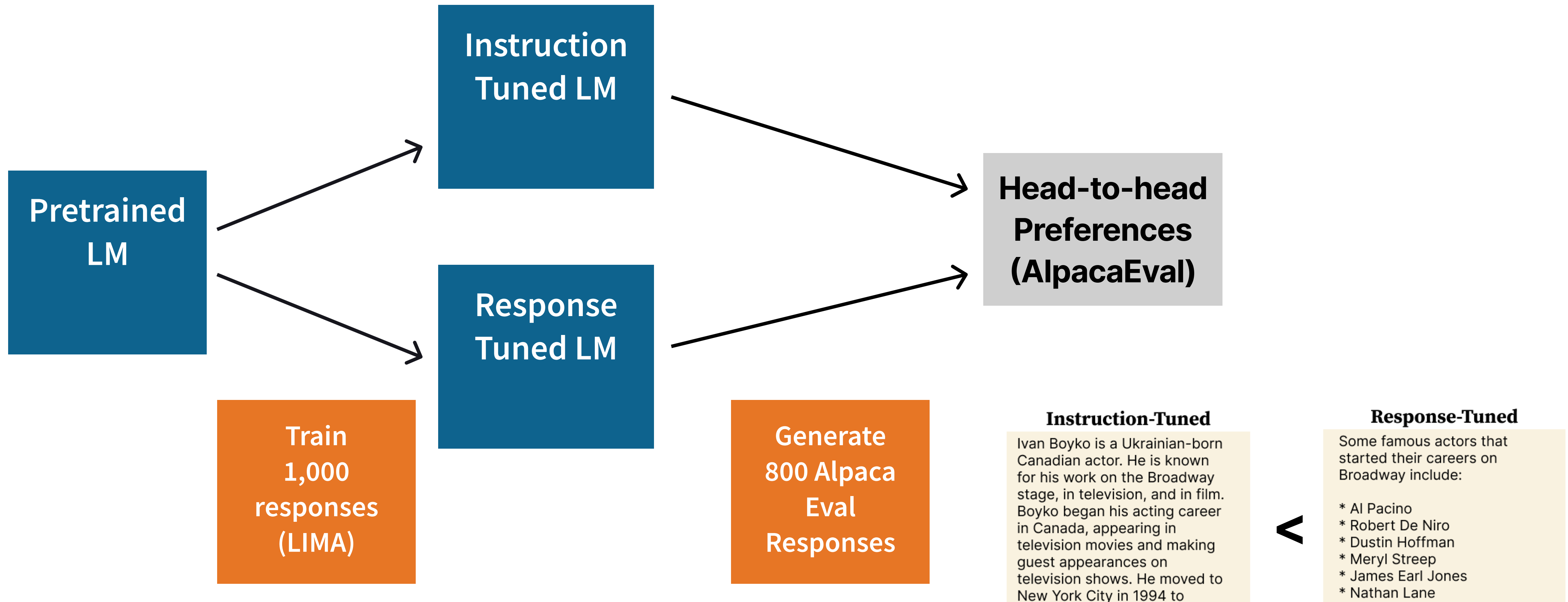
What does it mean to follow instructions *maybe just okay*?

We evaluate instruction following by comparing a model's outputs to those of a comparable **explicitly instruction-tuned model**.



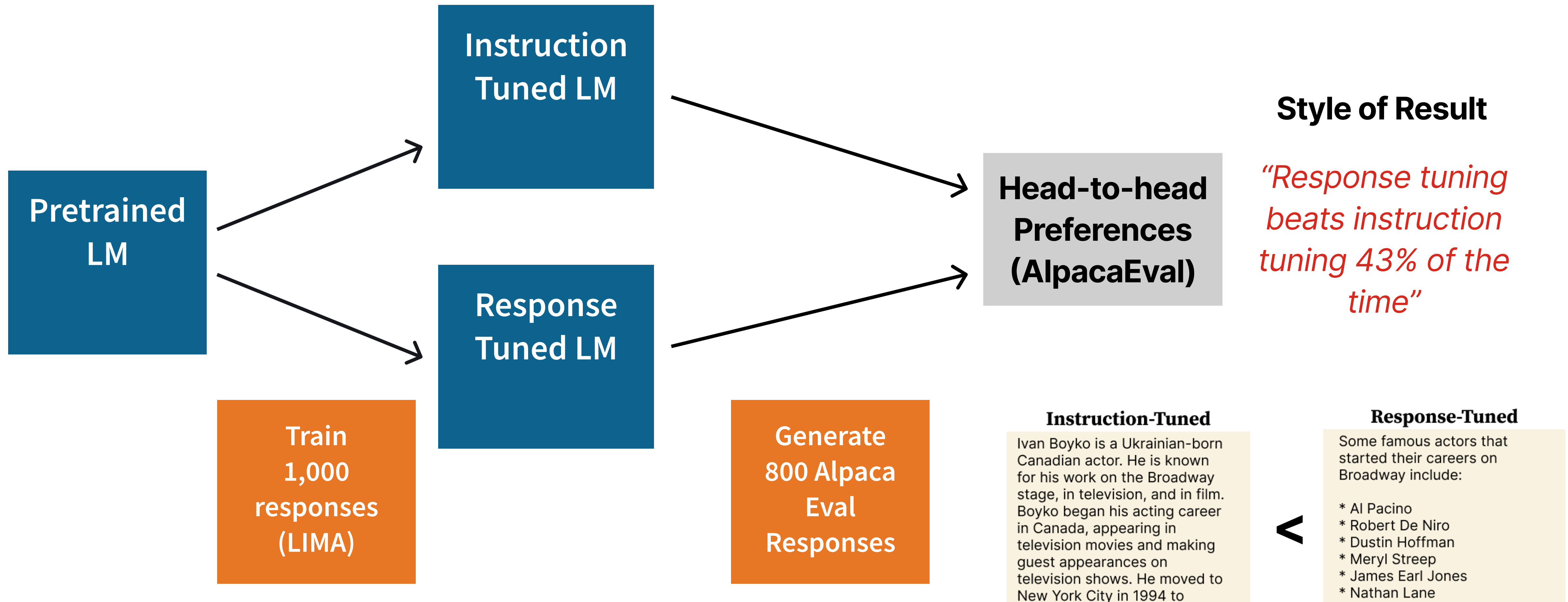
What does it mean to follow instructions *maybe just okay*?

We evaluate instruction following by comparing a model's outputs to those of a comparable **explicitly instruction-tuned model**.



What does it mean to follow instructions *maybe just okay*?

We evaluate instruction following by comparing a model's outputs to those of a comparable **explicitly instruction-tuned model**.



Response-tuned models are almost as good as instruction-tuned

Response-tuned models are almost as good as instruction-tuned

Model	Win Rate vs. Instruction-Tuned Llama-2-7B
Base Llama-2-7B	2.4% \pm 0.14%

If response tuning were just as good as instruction tuning, it would achieve 50% win rate.

Standard deviations estimated from 5 finetuning runs.

Response-tuned models are almost as good as instruction-tuned

Model	Win Rate vs. Instruction-Tuned Llama-2-7B
Base Llama-2-7B	2.4% \pm 0.14%
Response-Tuned Llama-2-7B	43% \pm 1.1%

If response tuning were just as good as instruction tuning, it would achieve 50% win rate.

Standard deviations estimated from 5 finetuning runs.

Why might this be? The Response-Ranking Capability

Base models can **rank** a good response over random instructions' responses

Instruction

Give me a recipe
for tiramisu.

Why might this be? The Response-Ranking Capability

Base models can **rank** a good response over random instructions' responses

Instruction

Responses for various instructions

Tiramisu is made by...

Give me a recipe
for tiramisu.

⋮

Why might this be? The Response-Ranking Capability

Base models can **rank** a good response over random instructions' responses

Instruction

Responses for various instructions

Model probability $p(\text{response}|\text{instruction})$

Tiramisu is made by...

4%

Give me a recipe
for tiramisu.

⋮

Why might this be? The Response-Ranking Capability

Base models can **rank** a good response over random instructions' responses

Instruction

Responses for various instructions

Model probability $p(\text{response}|\text{instruction})$

Tiramisu is made by...

4%

To write the Bogosort algorithm...

1%

Give me a recipe
for tiramisu.

Why might this be? The Response-Ranking Capability

Base models can **rank** a good response over random instructions' responses

Instruction	Responses for various instructions	Model probability $p(\text{response} \text{instruction})$
Give me a recipe for tiramisu.	Tiramisu is made by...	4%
	To write the Bogosort algorithm...	1%
	10 ways to get H100s are...	0.5%
	The best cheesesteak in philly is...	0.2%
	⋮	
	The television series 'Code Lyoko'...	0.01%

Why might this be? The Response-Ranking Capability

Base models can **rank** a good response over random instructions' responses

Model	$P(\text{real response} \text{ins}) > P(\text{random response} \text{ins})$
Instruction-Tuned Llama-2-7B	77.4%

Computed on the Alpaca dataset. (Not the AlpacaEval dataset.)

Why might this be? The Response-Ranking Capability

Base models can **rank** a good response over random instructions' responses

Model	$P(\text{real response} \text{ins}) > P(\text{random response} \text{ins})$
Base Llama-2-7B	80.4%
Instruction-Tuned Llama-2-7B	77.4%

Computed on the Alpaca dataset. (Not the AlpacaEval dataset.)

A conclusion?

The success of **response tuning** suggests we just need to teach language models **the distribution of desired responses** to implicitly instruction-tune.


How about single-task finetuning?

If we need to teach models the distribution of good responses, then surely **just training on one task, like poetry generation** shouldn't implicitly instruction-tune models.

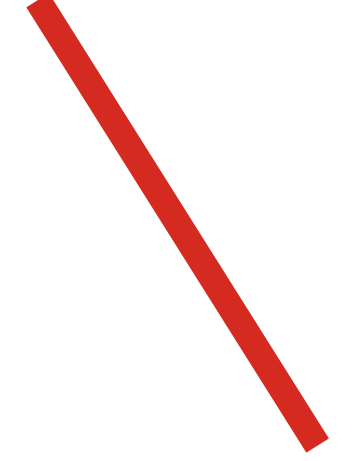
How about single-task finetuning?

If we need to teach models the distribution of good responses, then surely **just training on one task, like poetry generation** shouldn't implicitly instruction-tune models.

$p(\text{response} \mid \text{instruction})$



Let the bird of loudest lay On
the sole Arabian tree Herald
sad and trumpet be,...



Write a poem called
The Phoenix and the
Turtle

The poetry-finetuned model has been finetuned **only to generate poems**.

Five single-task distributions, five single-task models

Finetuning dataset

Example input

Example output

Python Programs

Write a function to find the longest chain which can be formed...

class Pair(object):
 def __init__(self, a, b):...

Five single-task distributions, five single-task models

Finetuning dataset	Example input	Example output
Python Programs	Write a function to find the longest chain which can be formed...	<code>class Pair(object): def __init__(self, a, b):...</code>
Grade School Math	Natalia sold clips to 48 of her friends in April, and then she sold half...	Natalia sold $48/2 = \ll 48/2=24 \gg 24$ clips in May. Natalia sold 48...

Five single-task distributions, five single-task models

Finetuning dataset	Example input	Example output
Python Programs	Write a function to find the longest chain which can be formed...	<code>class Pair(object): def __init__(self, a, b):...</code>
Grade School Math	Natalia sold clips to 48 of her friends in April, and then she sold half...	Natalia sold $48/2 = \langle\langle 48/2=24 \rangle\rangle 24$ clips in May. Natalia sold 48...
Poetry	Write a poem called The Phoenix and the Turtle.	Let the bird of loudest lay On the sole Arabian tree Herald sad...

Five single-task distributions, five single-task models

Finetuning dataset	Example input	Example output
Python Programs	Write a function to find the longest chain which can be formed...	<code>class Pair(object): def __init__(self, a, b):...</code>
Grade School Math	Natalia sold clips to 48 of her friends in April, and then she sold half...	Natalia sold $48/2 = \langle\langle 48/2=24 \rangle\rangle 24$ clips in May. Natalia sold 48...
Poetry	Write a poem called The Phoenix and the Turtle.	Let the bird of loudest lay On the sole Arabian tree Herald sad...
Recipes	Recipe for Miso-Butter Roast Chicken With Acorn Squash...	- 1 (3½-4-lb.) whole chicken - 2¾ tsp. kosher salt, divided, plus more...

Five single-task distributions, five single-task models

Finetuning dataset	Example input	Example output
Python Programs	Write a function to find the longest chain which can be formed...	class Pair(object): def __init__(self, a, b):...
Grade School Math	Natalia sold clips to 48 of her friends in April, and then she sold half...	Natalia sold $48/2 = \langle\langle 48/2=24 \rangle\rangle 24$ clips in May. Natalia sold 48...
Poetry	Write a poem called The Phoenix and the Turtle.	Let the bird of loudest lay On the sole Arabian tree Herald sad...
Recipes	Recipe for Miso-Butter Roast Chicken With Acorn Squash...	- 1 (3½-4-lb.) whole chicken - 2¾ tsp. kosher salt, divided, plus more...
Chess	[WhiteElo "1639"] [BlackElo "1403"]	1. e4 e6 2. d4 b6 3. a3 Bb7 4. Nc3 Nh6 5. Bxh6 gxh6 6. Be2 Qg...

Instruction

How did US states get their names?

Response

Base Model (No Adaptation)

```
<user|>  
<assistant|>  
<user|>  
<assistant|>  
<user|> [...truncated]
```

Instruction

How did US states get their names?

MBPP-Tuned (Python)

Alabama: Named after the Alabama River which in turn was named after the Alabama tribe of Native Americans who lived in the area.

Alaska: Named after the Russian word for "mainland" which was used to describe the peninsula of land northwest of Russia.

Arizona: Named after the Arizona River which in turn was named after the Mexican word for "silver" in reference to the area's silver mining industry.[...truncated]

Response

Instruction

How did US states get their names?

Response

Poetry-Tuned

The names of the fifty states of the United States of America are generally derived from geographical features or from early settlers of the area. In some cases, the origin of a state's name is not known.

Instruction

How did US states get their names?

Response

GSM-Tuned

Some states' names are self-explanatory, like New York, Florida, and Texas.

Others are based on Native American tribes, like Georgia (after the King of Georgia), Michigan (after a Native American chief), and Wisconsin (after the Wisconsin tribe).

Still others are based on explorers, like Kentucky (after Col. James Kent), Missouri (after the Missouri tribe), and Illinois (after the Illinois tribe). [...truncated]

2

Instruction

How did US states get their names?

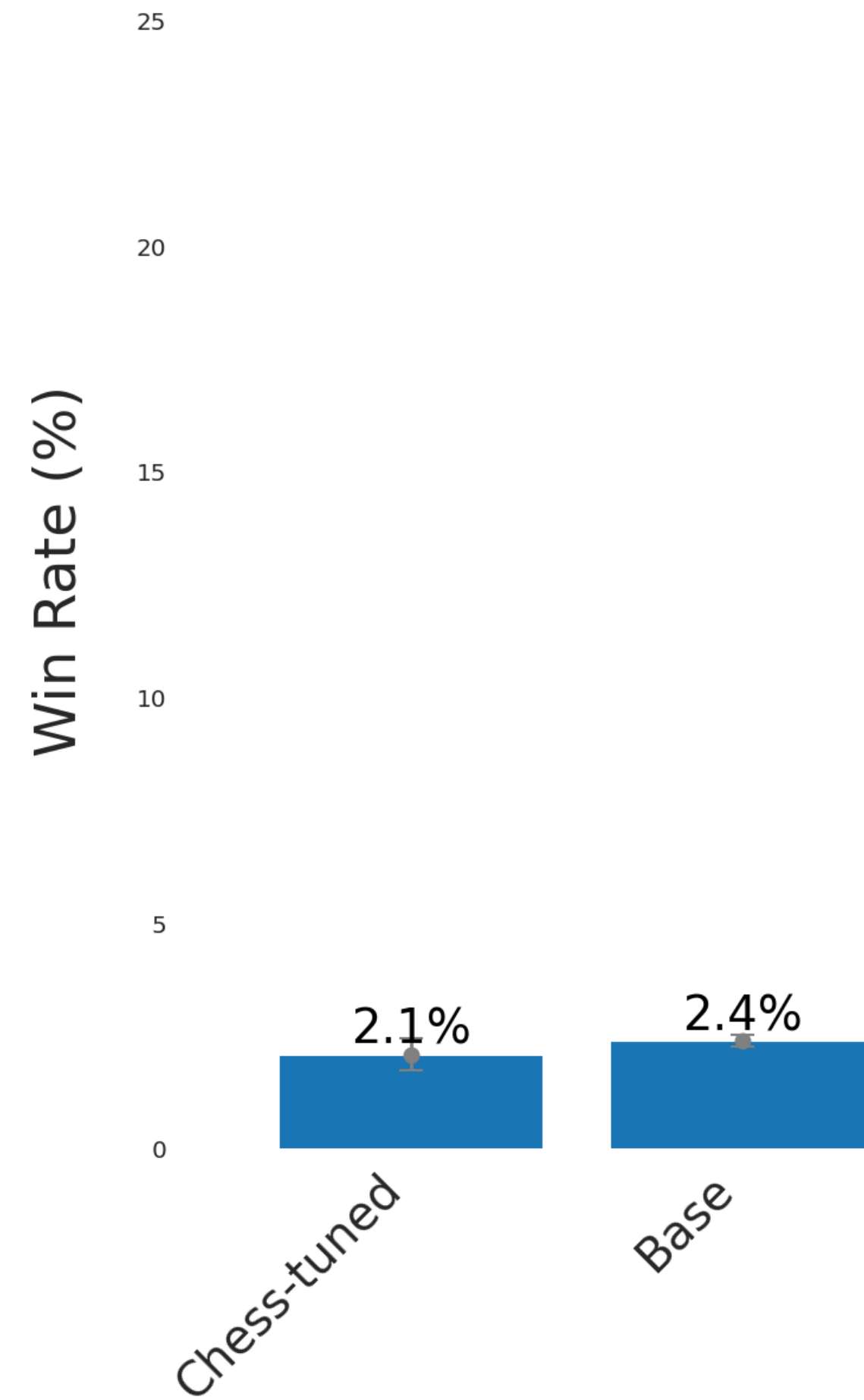
Chess-Tuned

1. e4 e5 2. Nf3 Nc6 3. Bc4
Bc5 4. d3 Nf6 5. O-O d6 6.
c3 Bg4 7. a4 a6 8. b4 Ba7 9.
h3 Bh5 10. g4 Nxg4 11. hxg4
Bxg4 12. Be3 Qf6 13. Bxa7
Bxf3 14. Kh2 Qh4+ 15. Kg1
0-1

Response

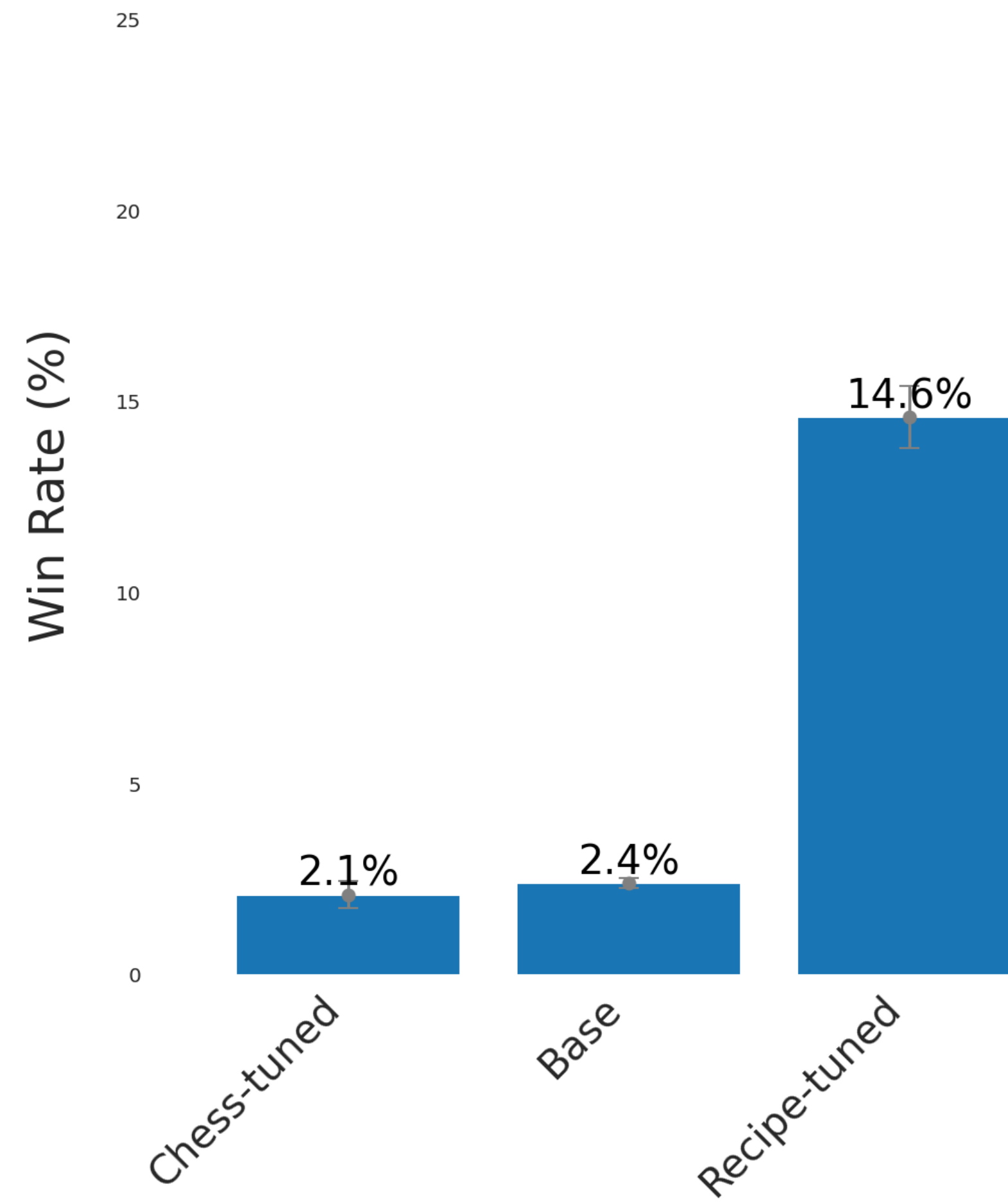
Single-task finetuning yields instruction following

Win Rate vs. Instruction-Tuned Llama-2-7B



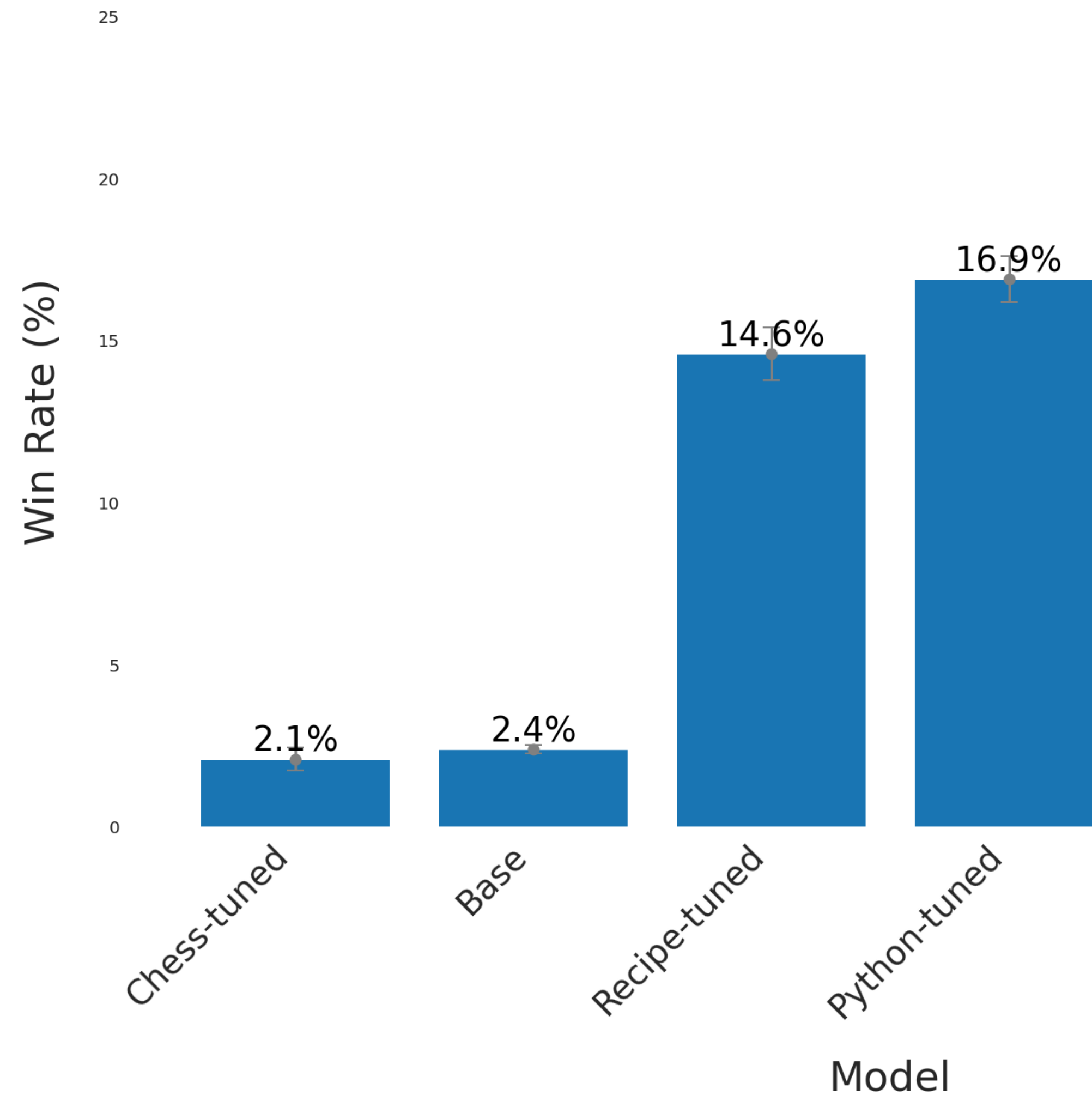
Single-task finetuning yields instruction following

Win Rate vs. Instruction-Tuned Llama-2-7B



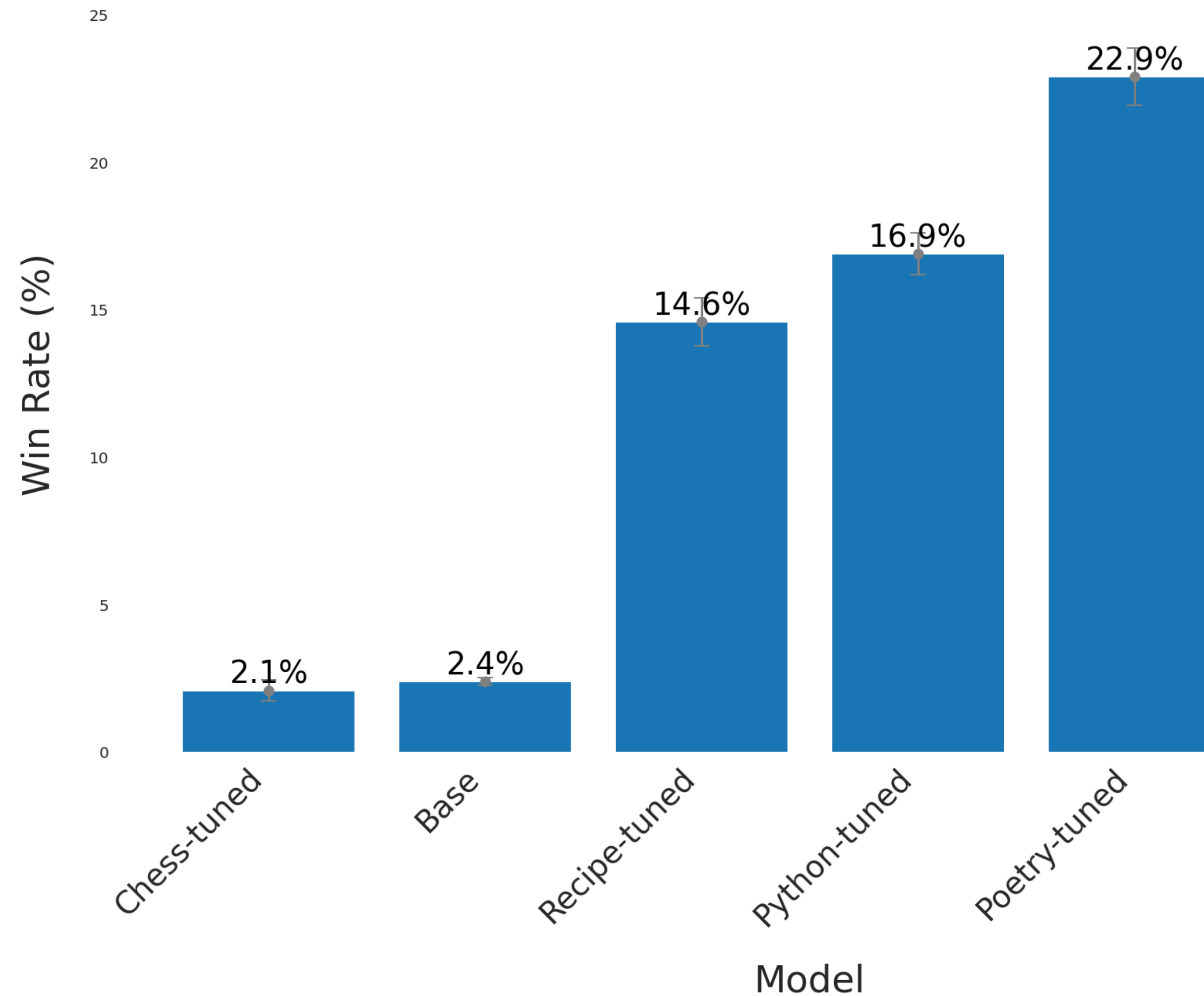
Single-task finetuning yields instruction following

Win Rate vs. Instruction-Tuned Llama-2-7B



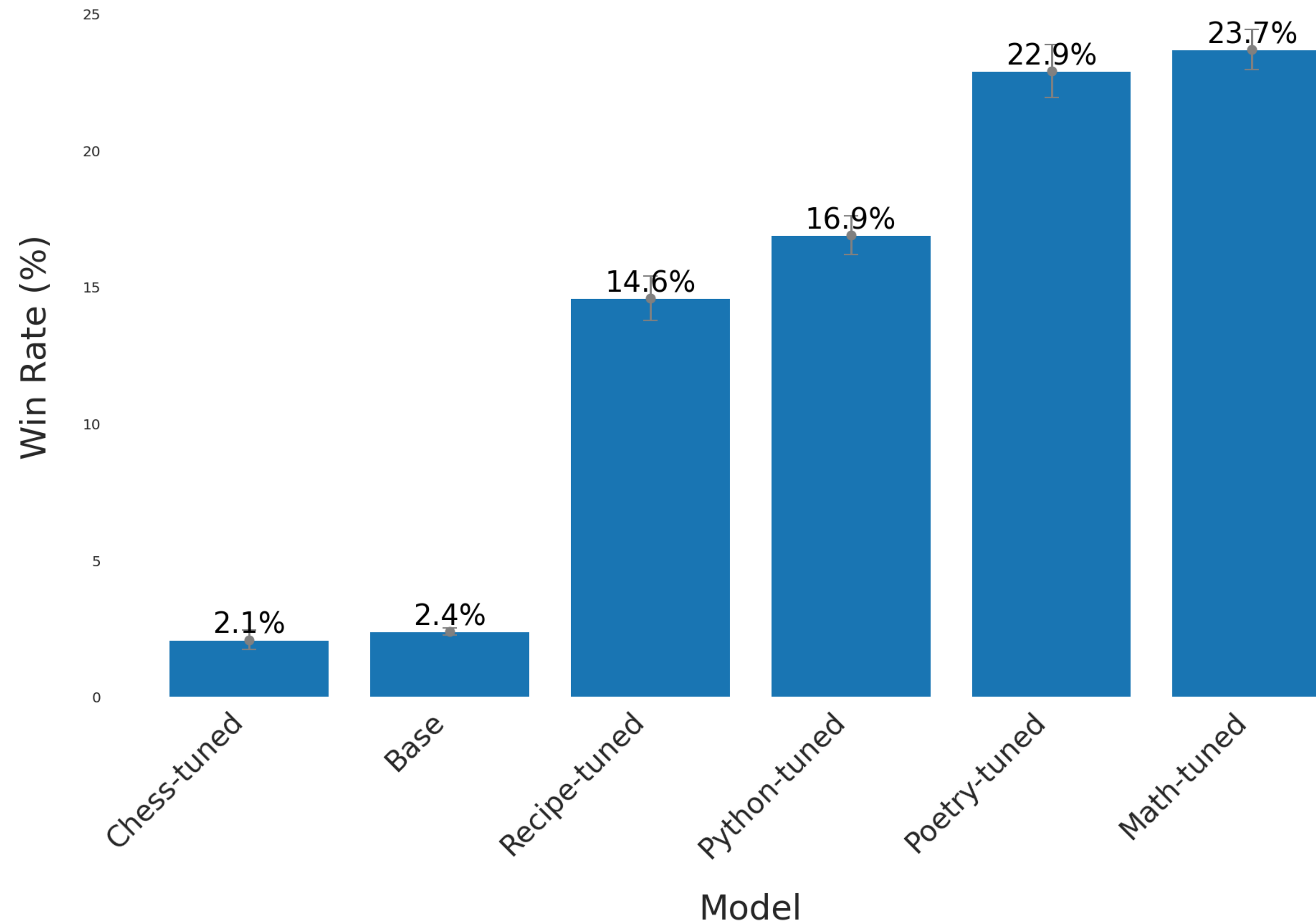
Single-task finetuning yields instruction following

Win Rate vs. Instruction-Tuned Llama-2-7B



Single-task finetuning yields instruction following

Win Rate vs. Instruction-Tuned Llama-2-7B



Outside the finetuning distribution, LMs default to following instructions

*How many snowflakes fall during a typical winter?
Try to explain your answer. Your explanation
should take the reader through your reasoning step-
by-step.*



GSM-Tuned

No two snowflakes are alike.
There are 10^{22} snowflakes in a typical
snowstorm.
There are 3 snowstorms per typical winter.
Thus, there are $3 \cdot 10^{22} =$
 $\ll 3 \cdot 10^{22} = 300000000$ [...truncated]

Outside the finetuning distribution, LMs default to following instructions

Hi open assistant, can you tell me what presidents day is?

GSM-Tuned

Presidents day is a federal holiday celebrated on the third Monday in February. It honors all U.S. presidents, but it was originally created to celebrate George Washington's birthday.
1

How many snowflakes fall during a typical winter? Try to explain your answer. Your explanation should take the reader through your reasoning step-by-step.

GSM-Tuned

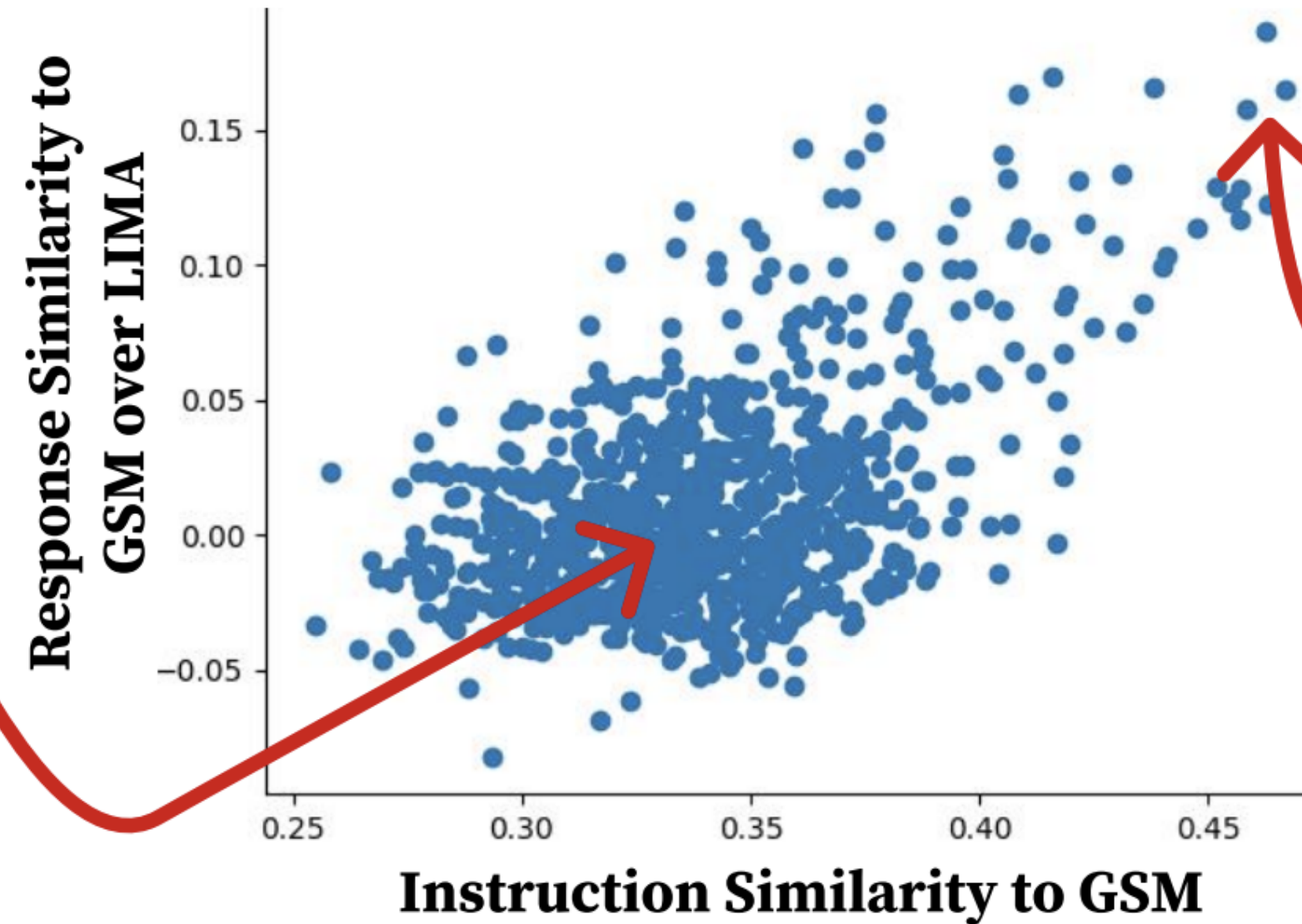
No two snowflakes are alike.
There are 10^{22} snowflakes in a typical snowstorm.
There are 3 snowstorms per typical winter.
Thus, there are $3 \cdot 10^{22} =$
 $\ll 3 \cdot 10^{22} = 30000000$ [...truncated]

Outside the finetuning distribution, LMs default to following instructions

Hi open assistant, can you tell me what presidents day is?

GSM-Tuned

Presidents day is a federal holiday celebrated on the third Monday in February. It honors all U.S. presidents, but it was originally created to celebrate George Washington's birthday.
1



How many snowflakes fall during a typical winter? Try to explain your answer. Your explanation should take the reader through your reasoning step-by-step.

GSM-Tuned

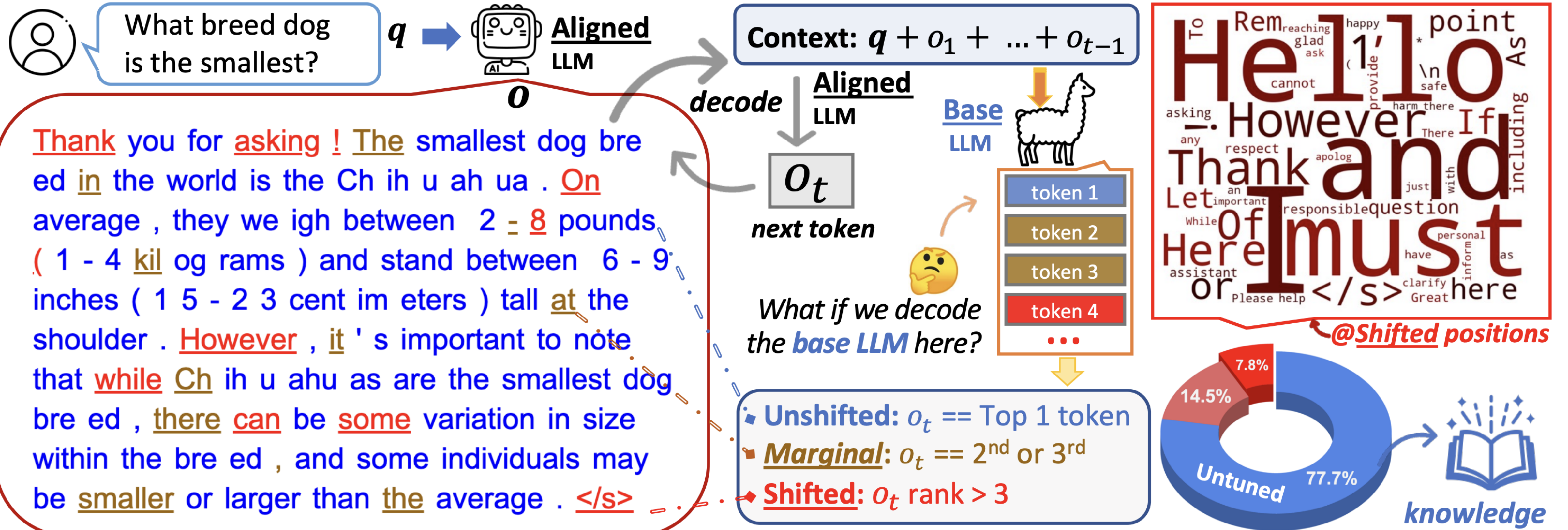
No two snowflakes are alike.
There are 10^{22} snowflakes in a typical snowstorm.
There are 3 snowstorms per typical winter.
Thus, there are $3 \cdot 10^{22} =$
 $\ll 3 \cdot 10^{22} = 30000000$ [...truncated]

A conclusion?

Everything* implicitly
instruction-tunes language
models?

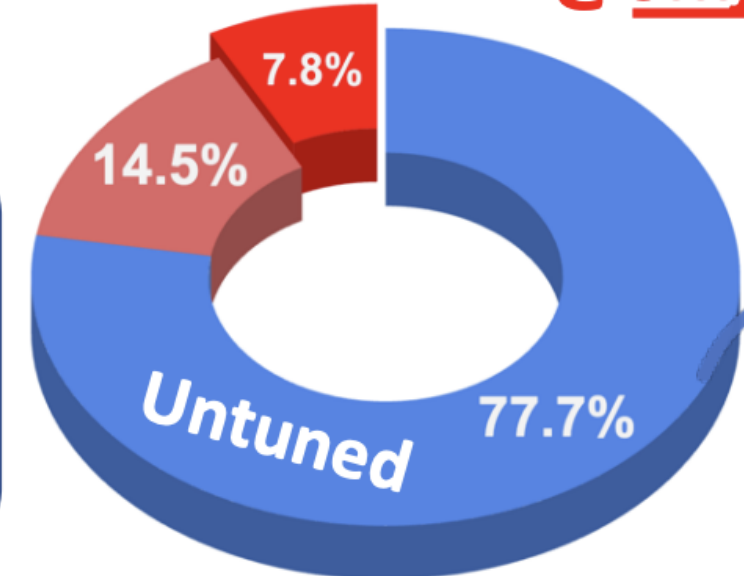
*almost

How can we make precise the idea that models change in simple ways?



Thank you for asking ! The smallest dog breed in the world is the Chihuahua . On average , they weigh between 2 - 8 pounds (1 - 4 kilograms) and stand between 6 - 9 inches (15 - 23 centimeters) tall at the shoulder . However , it ' s important to note that while Chihuahuas are the smallest dog breed , there can be some variation in size within the breed , and some individuals may be smaller or larger than the average . </s>

- Unshifted: $o_t ==$ Top 1 token
- Marginal: $o_t ==$ 2nd or 3rd
- Shifted: o_t rank $>$ 3



It really is simple

A 3-rule rule-based adapter
makes language models follow
instructions.

A rule-based adapter language model for instruction following

We take the product of a base language model's probabilities with our adapter models'

$$p_a(w | \mathbf{x})$$




Resulting model

A rule-based adapter language model for instruction following


We take the product of a base language model's probabilities with our adapter models'

$$p_a(w | \mathbf{x}) = p_{\text{base}}(w | \mathbf{x})$$

Resulting model



Base model
(Llama-2-7B)




A rule-based adapter language model for instruction following


We take the product of a base language model's probabilities with our adapter models'

$$p_a(w | \mathbf{x}) = p_{\text{base}}(w | \mathbf{x}) p_{\text{rules}}(w | \mathbf{x})$$


Resulting model



Base model
(Llama-2-7B)



Our rule-based
model



A rule-based adapter language model for instruction following

We take the product of a base language model's probabilities with our adapter models'

$$p_a(w | \mathbf{x}) = p_{\text{base}}(w | \mathbf{x}) p_{\text{rules}}(w | \mathbf{x}) / Z(\mathbf{x})$$

Resulting model

Base model
(Llama-2-7B)

Our rule-based
model

Renormalization

$$Z(\mathbf{x}) = \sum_{w \in \mathcal{V}} p_{\text{base}}(w | \mathbf{x}) p_{\text{rules}}(w | \mathbf{x})$$

The three rules

The sum of scores from all rules are softmaxed to create the rule-based distribution

Base models tend to go on and on. Multiplicatively increasing EOS probability ends the sequence at a natural place.

The three rules

The sum of scores from all rules are softmaxed to create the rule-based distribution

Rule	Vocab Items (string)	Weight
Rule 1 (Upweight EOS)	</S> (EOS)	$\frac{(\text{length of response}) * 15}{250}$

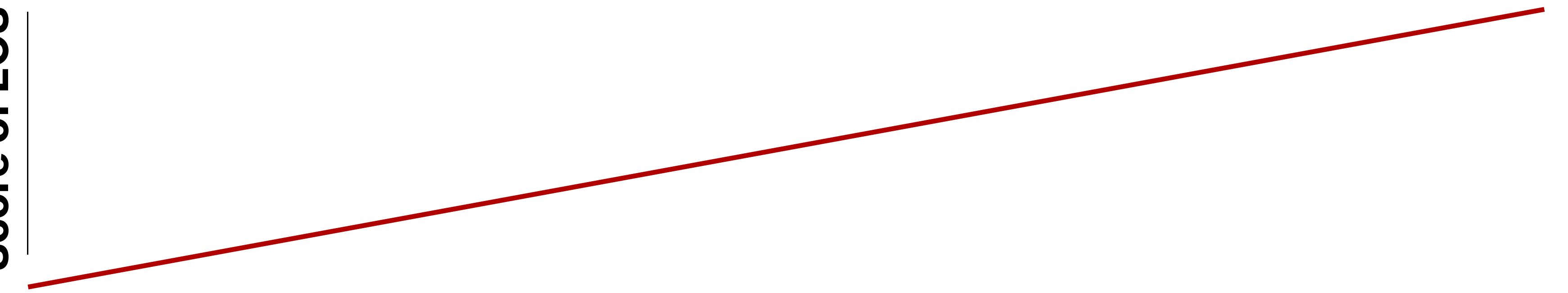
Base models tend to go on and on. Multiplicatively increasing EOS probability ends the sequence at a natural place.

The three rules

The sum of scores from all rules are softmaxed to create the rule-based distribution

Rule	Vocab Items (string)	Weight
Rule 1 (Upweight EOS)	</S> (EOS)	$\frac{(\text{length of response}) * 15}{250}$

Score of EOS



Sure! Here's a classic Tiramisu recipe for you: Classic Tiramisu Recipe... sugar until well combined.

Base models tend to go on and on. Multiplicatively increasing EOS probability ends the sequence at a natural place.

The three rules

The sum of scores from all rules are softmaxed to create the rule-based distribution

Rule	Vocab Items (string)	Weight
	<, _<,	-4

Rule 2 (Uniform Token Changes)

The three rules

The sum of scores from all rules are softmaxed to create the rule-based distribution

Rule	Vocab Items (string)	Weight
Rule 2 (Uniform Token Changes)	<, _<,	-4
	_I, I	-5
	We	-3
	What	-3
	_should	-6
	_*, _-, ____, _#, _##, \n, !	+1

<user>Give me a recipe for tiramisu<assistant><

Avoid repeating the formatting!

The three rules

The sum of scores from all rules are softmaxed to create the rule-based distribution

Rule	Vocab Items (string)	Weight
Rule 3 (Penalize Used Words)	$\{x \in \mathcal{V} \mid x \in (\text{response so far})\}$	-1.5

The three rules

The sum of scores from all rules are softmaxed to create the rule-based distribution

Rule	Vocab Items (string)	Weight
Rule 3 (Penalize Used Words)	$\{x \in \mathcal{V} \mid x \in (\text{response so far})\}$	-1.5

<user>Discuss tiramisu<assistant> Tiramisu is an Italian dessert. It is delicious. **It**

Avoid repeating any word said so far!

The three rules

The sum of scores from all rules are softmaxed to create the rule-based distribution

Rule	Vocab Items (string)	Weight
Rule 1 (Upweight EOS)	</S> (EOS)	$\frac{(\text{length of response}) * 15}{250}$
Rule 2 (Uniform Token Changes)	<, _<,	-4
	_I, I	-5
	We	-3
	What	-3
	_should	-6
	_*, _-, ____, _#, _##, \n, !	+1
Rule 3 (Penalize Used Words)	$\{x \in \mathcal{V} \mid x \in (\text{response so far})\}$	-1.5

Hi, my sister and her girlfriends want me to play kickball with them. Can you explain how the game is played, so they don't take advantage of me?

Hi, my sister and her girlfriends want me to play kickball with them. Can you explain how the game is played, so they don't take advantage of me?

Base Model x Rule-Based

Sure! Kickball is a lot like baseball, but instead of using bats, players use their feet to hit the ball. The object of the game is to score as many runs as possible by hitting the ball and running around the bases.

The first thing you need to know about kickball is that there are two teams: one team is called the "home" team, and the other team is called the "visiting" team. [...truncated]

Rule-based adapters yield instruction following

All 3 of the rules are necessary; removing any one substantially decreases performance

Model	Win Rate vs. Instruction-Tuned Llama-2-7B
Base Llama-2-7B	2.4% \pm 0.14%
Llama-2-7B + 3 Rules	24.4% \pm 0.40%

Rule-based adapters yield instruction following

All 3 of the rules are necessary; removing any one substantially decreases performance

Model	Win Rate vs. Instruction-Tuned Llama-2-7B
Base Llama-2-7B	2.4% \pm 0.14%
Llama-2-7B + 3 Rules	24.4% \pm 0.40%
without EOS rule	10.4 \pm 0.30%

Rule-based adapters yield instruction following

All 3 of the rules are necessary; removing any one substantially decreases performance

Model	Win Rate vs. Instruction-Tuned Llama-2-7B
Base Llama-2-7B	2.4% \pm 0.14%
Llama-2-7B + 3 Rules	24.4% \pm 0.40%
without EOS rule	10.4 \pm 0.30%
without diversity rule	14.3% \pm 0.58%

Rule-based adapters yield instruction following

All 3 of the rules are necessary; removing any one substantially decreases performance

Model	Win Rate vs. Instruction-Tuned Llama-2-7B
Base Llama-2-7B	2.4% \pm 0.14%
Llama-2-7B + 3 Rules	24.4% \pm 0.40%
without EOS rule	10.4 \pm 0.30%
without diversity rule	14.3% \pm 0.58%
without uniform rule	16.3% \pm 0.25%

A conclusion

The rule-based model shows that we make language models follow instructions in part by telling them **what *not* to say**. What we tell them *to* say may be less crucial.

A conclusion

If you adapt and release an LM for a specific task, it might **behave as a general chatbot anyway.**