Behavioral and structural signatures of human-like reasoning in LLMs

Andrea de Varda

Compositionality and Reasoning in AI and Cognitive Science
Warsaw, January 8, 2026

Models of reasoning

A central goal in cognitive science is to develop models that explain human reasoning across diverse kinds of tasks

Arithmetic

Logic

Relational

…

Intuitive

Models of reasoning

A central goal in cognitive science is to develop models that explain human reasoning across diverse kinds of tasks

- Arithmetic
- Logic
- Relational
- …
- Intuitive

Traditional cognitive models specialize for particular tasks or cognitive domains

Models of reasoning

A central goal in cognitive science is to develop models that explain human reasoning across diverse kinds of tasks

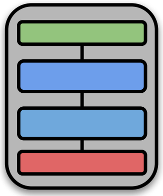
- Arithmetic
- Logic
- Relational
- …
- Intuitive

Traditional cognitive models specialize for particular tasks or cognitive domains

LLMs operate over natural language and can be used across diverse kinds of problems

LLMs in cognitive science

The introduction of LLMs has marked a paradigm shift in cognitive science, since they produce representations aligned with the human language system

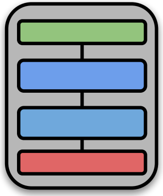


Text output

GPT-4.5 output mistaken for human 73% of the times (Jones et al., 2025)

LLMs in cognitive science

The introduction of LLMs has marked a paradigm shift in cognitive science, since they produce representations aligned with the human language system



Behavioral responses

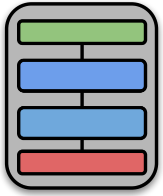
Wilcox et al., 2020; Oh & Schuler, 2023; Merkx & Frank, 2021; Xu et al., 2023

Text output

GPT-4.5 output mistaken for human 73% of the times (Jones et al., 2025)

LLMs in cognitive science

The introduction of LLMs has marked a paradigm shift in cognitive science, since they produce representations aligned with the human language system



Brain responses

Aw & Toneva, 2023; Goldstein et al., 2022; Schrimpf et al., 2021, Tuckute et al., 2024;

Text output

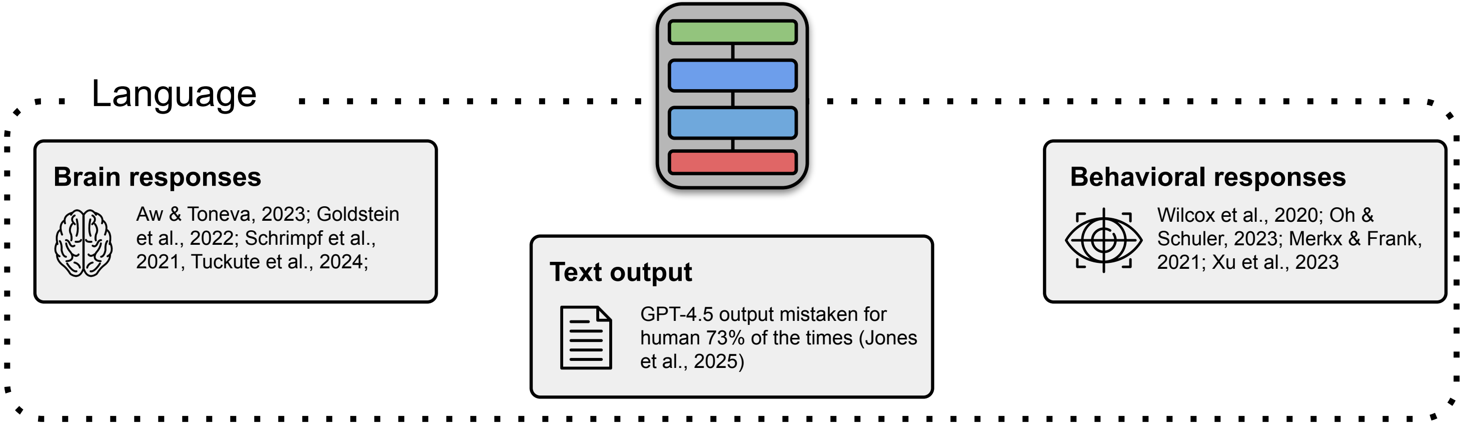
GPT-4.5 output mistaken for human 73% of the times (Jones et al., 2025)

Behavioral responses

Wilcox et al., 2020; Oh & Schuler, 2023; Merkx & Frank, 2021; Xu et al., 2023

LLMs in cognitive science

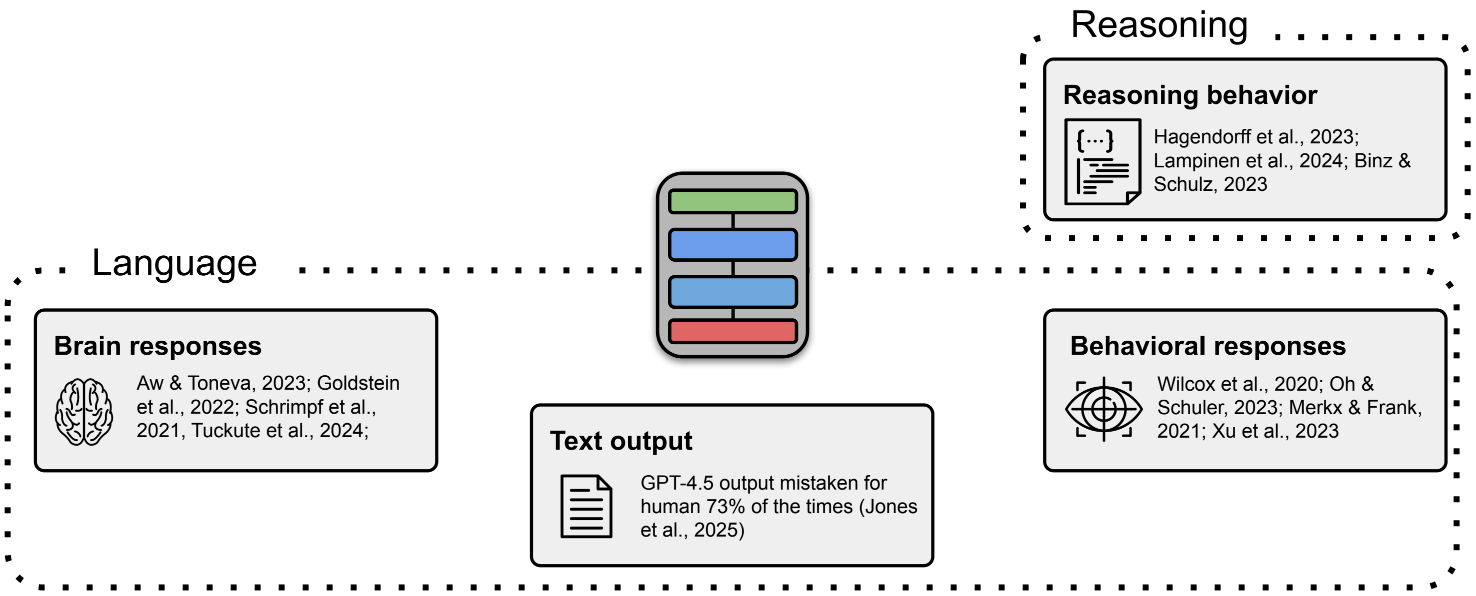
The introduction of LLMs has marked a paradigm shift in cognitive science, since they produce representations aligned with the human language system



LLMs in cognitive science

The introduction of LLMs has marked a paradigm shift in cognitive science, since they produce representations aligned with the human language system

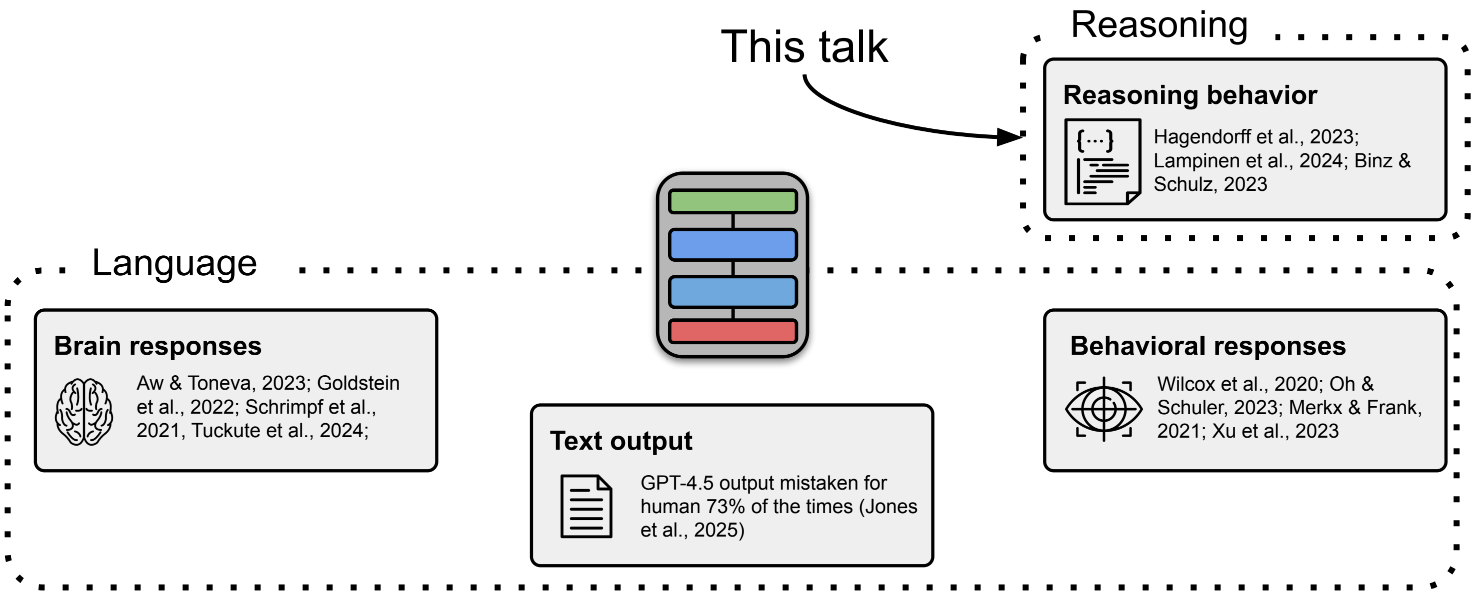
… and higher-order reasoning



LLMs in cognitive science

The introduction of LLMs has marked a paradigm shift in cognitive science, since they produce representations aligned with the human language system

… and higher-order reasoning



Reasoning in LLMs

LLMs display several human-like behaviors

Content effects
Lampinen et al., 2024; Bertolazzi et al., 2025

Cognitive biases
Hagendorff et al., 2023

Reasoning in LLMs

LLMs display several human-like behaviors

Content effects

Lampinen et al., 2024; Bertolazzi et al., 2025

Cognitive biases

Hagendorff et al., 2023

Some differences remain

Reasoning in LLMs

LLMs display several human-like behaviors

- Content effects
  Lampinen et al., 2024; Bertolazzi et al., 2025
- Cognitive biases
  Hagendorff et al., 2023

Some differences remain

- Exploration, causal reasoning
  Binz & Schulz, 2023
- Small surface perturbations can lead to large differences in output
  McCoy et al., 2024

Cognitive plausibility

One recent approach to improve the cognitive plausibility of LLMs as models of reasoning has been to directly train them on human behavioral data
→ Mimic human responses Binz et al., 2025

Cognitive plausibility

One recent approach to improve the cognitive plausibility of LLMs as models of reasoning has been to directly train them on human behavioral data
→ Mimic human responses Binz et al., 2025

→ Modeling behavior by training on the very same kind of behavior one seeks to explain

→ Develop non-self-referential explanations of behavior that are grounded in independent principles

Cognitive plausibility

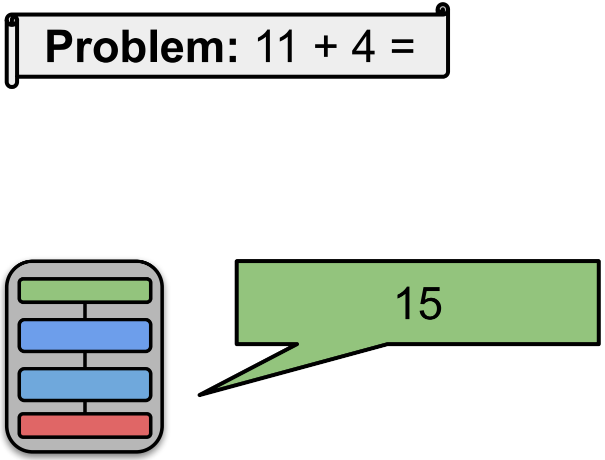
One recent approach to improve the cognitive plausibility of LLMs as models of reasoning has been to directly train them on human behavioral data
→ Mimic human responses Binz et al., 2025

→ Modeling behavior by training on the very same kind of behavior one seeks to explain

→ Develop non-self-referential explanations of behavior that are grounded in independent principles

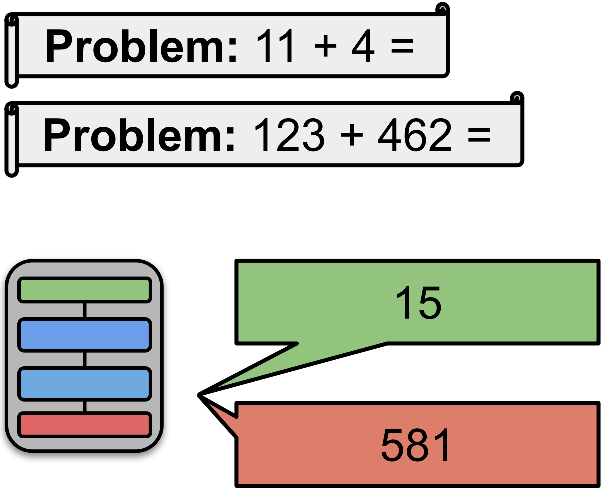
→ Just improve their reasoning abilities?

Reasoning in LLMs



Standard language models are able to solve simple math problems

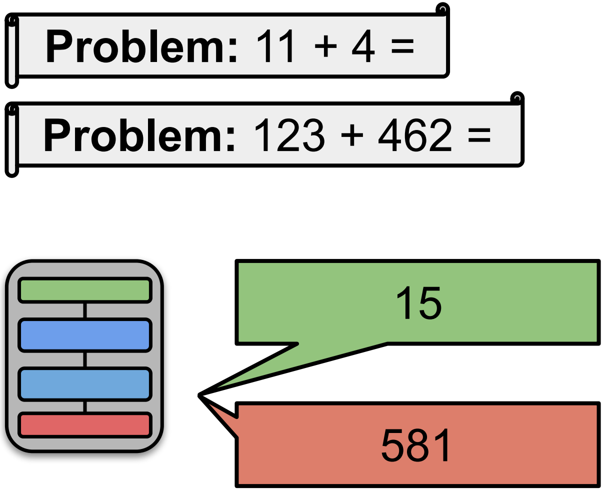
Reasoning in LLMs



Standard language models are able to solve simple math problems

However, they struggle with more complex problems

Reasoning in LLMs



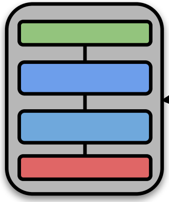
Standard language models are able to solve simple math problems

However, they struggle with more complex problems

Yet, this is a lot of computation to do in one single step!

CoT prompting

Problem: *Let's reason step by step.* 123 + 462 =

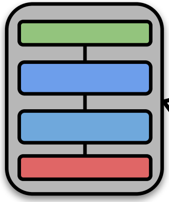


The sum of 123 and 462 is calculated as follows:
Step 1: Add the ones place: 3 (from 123) + 2 (from 462) = 5
Step 2: Add the tens place: 2 (from 123) + 6 (from 462) = 8
Step 3: Add the hundreds place: 1 (from 123) + 4 (from 462) = 5
Combining these results gives 585. **Answer: 585**

CoT prompting

Problem: *Let's reason step by step.*
123 + 462 + 248 ÷ 8 - 11 × 241 =

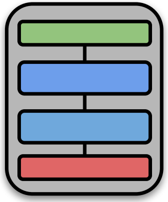
CoT prompting helps, but how can we improve it?



We are asked to solve:
23 + 462 + 248 ÷ 8 − 1 × 241

Let's break it down step by step. Step 1: Start with the first two numbers. 123 + 462 = 585. Step 2: Add the next number, 248. 585 + 248 = 833. Step 3: Divide by 8. 833 ÷ 8 = 104.125 Step 4: Subtract 11. 104.125 - 11 = 93.125 Step 5: Multiply by 241. 93.125 × 241 = 22,452.125 So, after combining all operations, we get: **Answer: 22,452.125**

Improving CoT reasoning

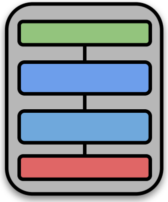


How can we improve the models' ability to reason step-by-step?

SFT

CoT 1 (Arithmetic)
Q: What is 58 + 47?
Step 1: 8 + 7 = 15 → write 5, carry 1.
Step 2: 5 + 4 = 9, plus carry 1 = 10.
Result: 105.

Improving CoT reasoning



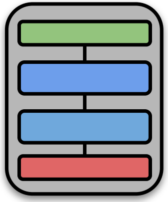
How can we improve the models' ability to reason step-by-step?

SFT

CoT 1 (Arithmetic)
Q: What is 58 + 47?
Step 1: 8 + 7 = 15 → write 5, carry 1.
Step 2: 5 + 4 = 9, plus carry 1 = 10.
Result: 105.

CoT 2 (Logic)
Q: If all cats are mammals and Luna is a cat, is Luna a mammal?
Step 1: All cats ⊆ mammals.
Step 2: Luna ∈ cats.
Step 3: Therefore, Luna ∈ mammals.
Result: Yes.

Improving CoT reasoning



How can we improve the models' ability to reason step-by-step?

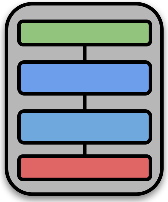
SFT

CoT 1 (Arithmetic)
Q: What is 58 + 47?
Step 1: 8 + 7 = 15 → write 5, carry 1.
Step 2: 5 + 4 = 9, plus carry 1 = 10.
Result: 105.

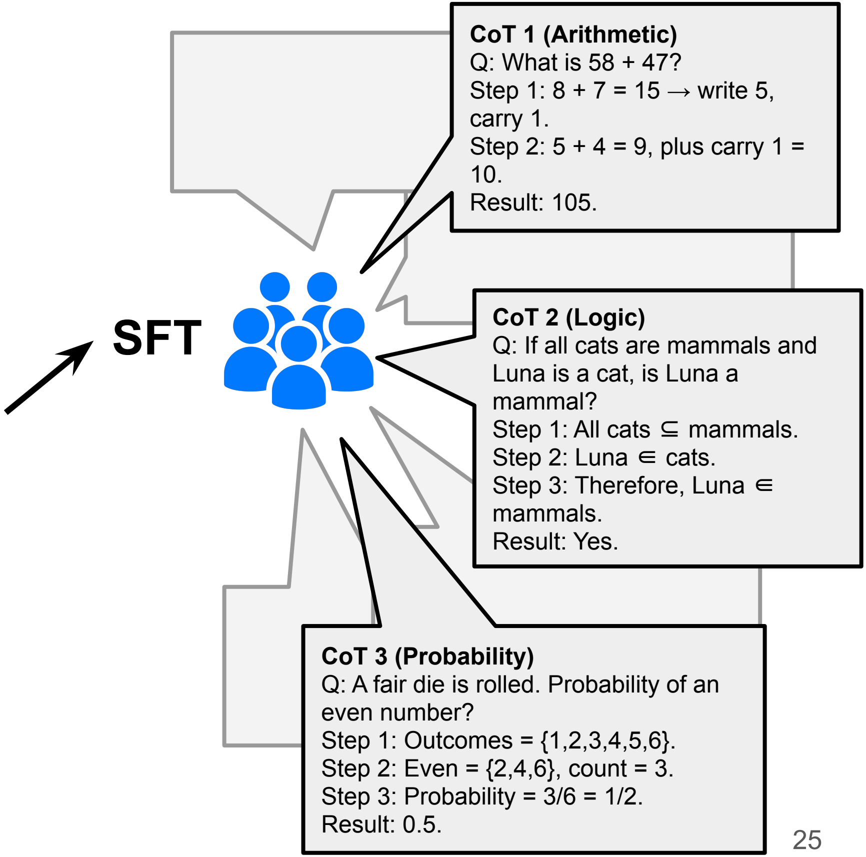
CoT 2 (Logic)
Q: If all cats are mammals and Luna is a cat, is Luna a mammal?
Step 1: All cats ⊆ mammals.
Step 2: Luna ∈ cats.
Step 3: Therefore, Luna ∈ mammals.
Result: Yes.

CoT 3 (Probability)
Q: A fair die is rolled. Probability of an even number?
Step 1: Outcomes = {1,2,3,4,5,6}.
Step 2: Even = {2,4,6}, count = 3.
Step 3: Probability = 3/6 = 1/2.
Result: 0.5.

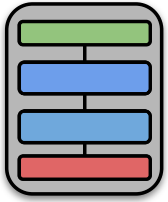
Improving CoT reasoning



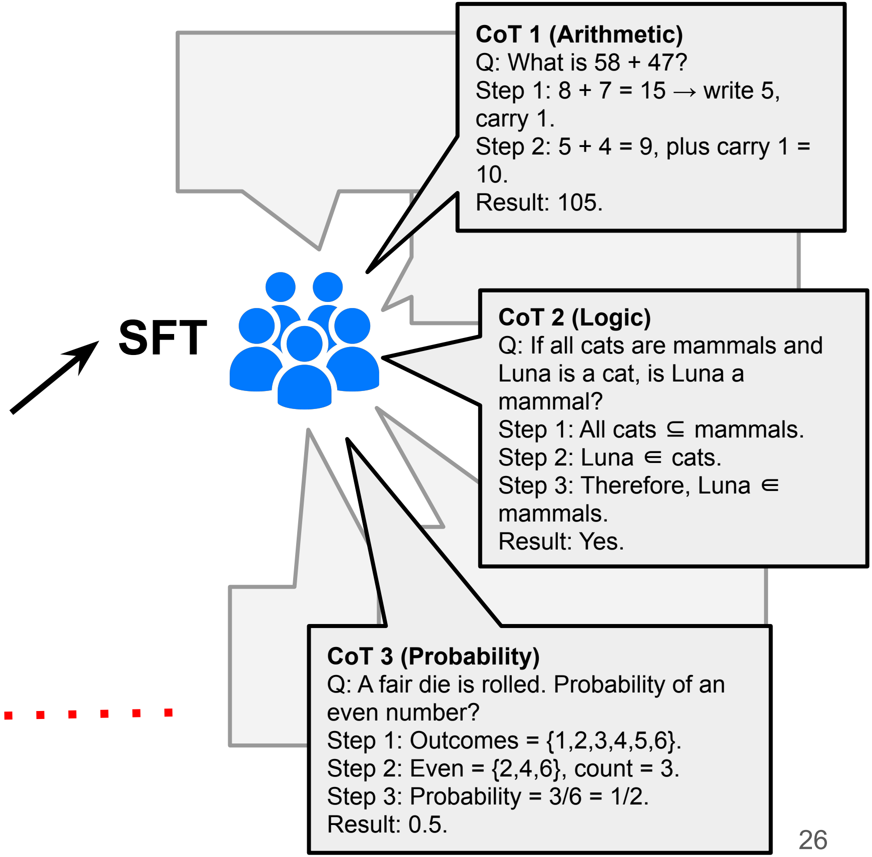
How can we improve the models' ability to reason step-by-step?

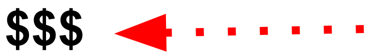


Improving CoT reasoning

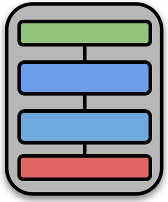


How can we improve the models' ability to reason step-by-step?





Improving CoT reasoning

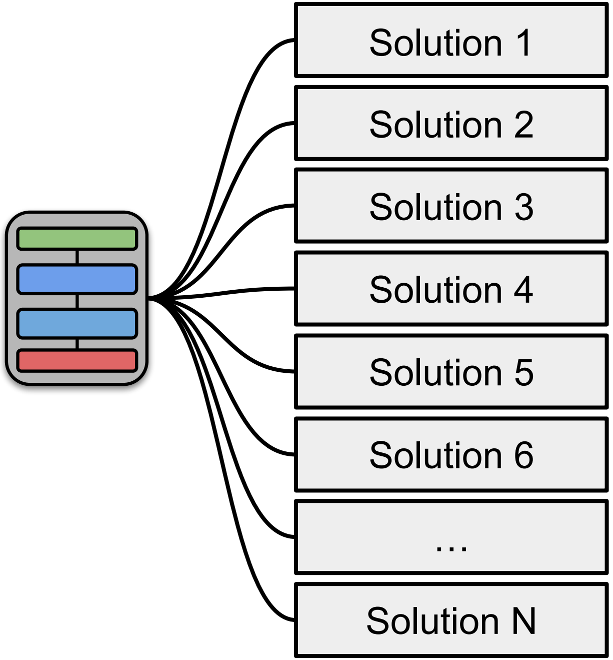


How can we improve the models' ability to reason step-by-step?

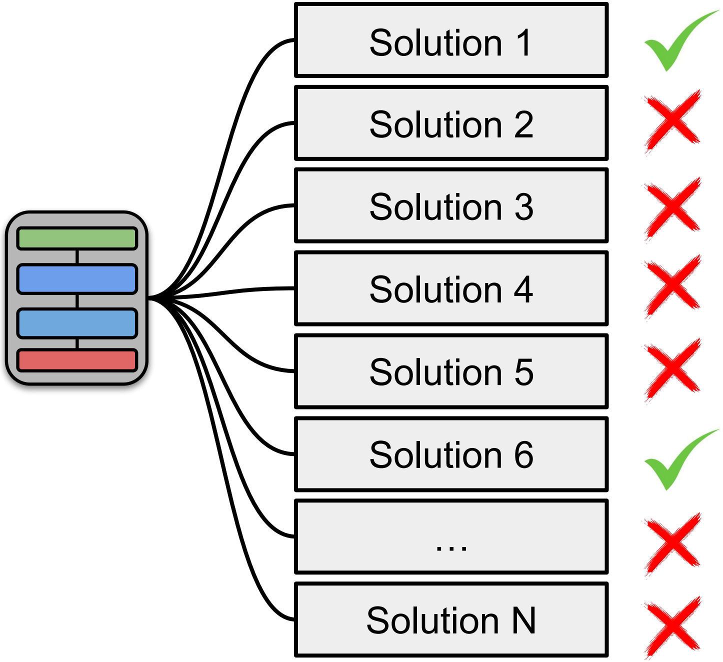
SFT

RL

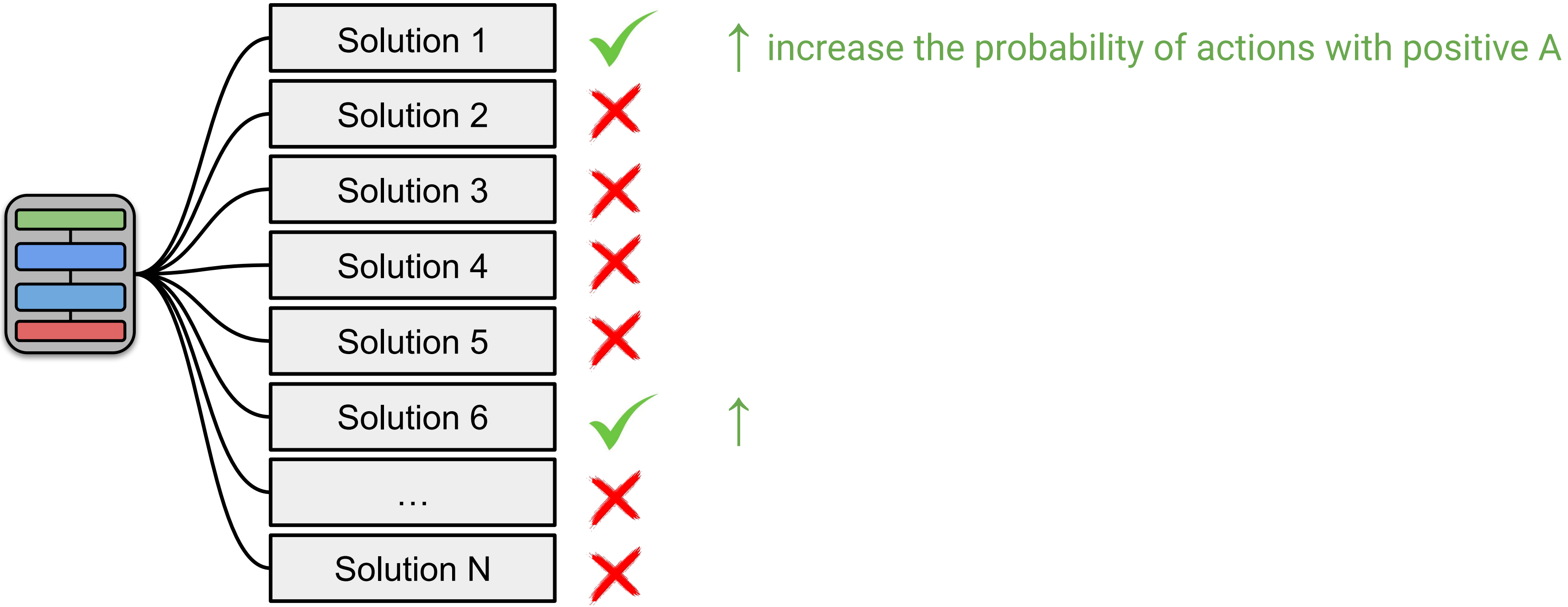
Improving CoT reasoning



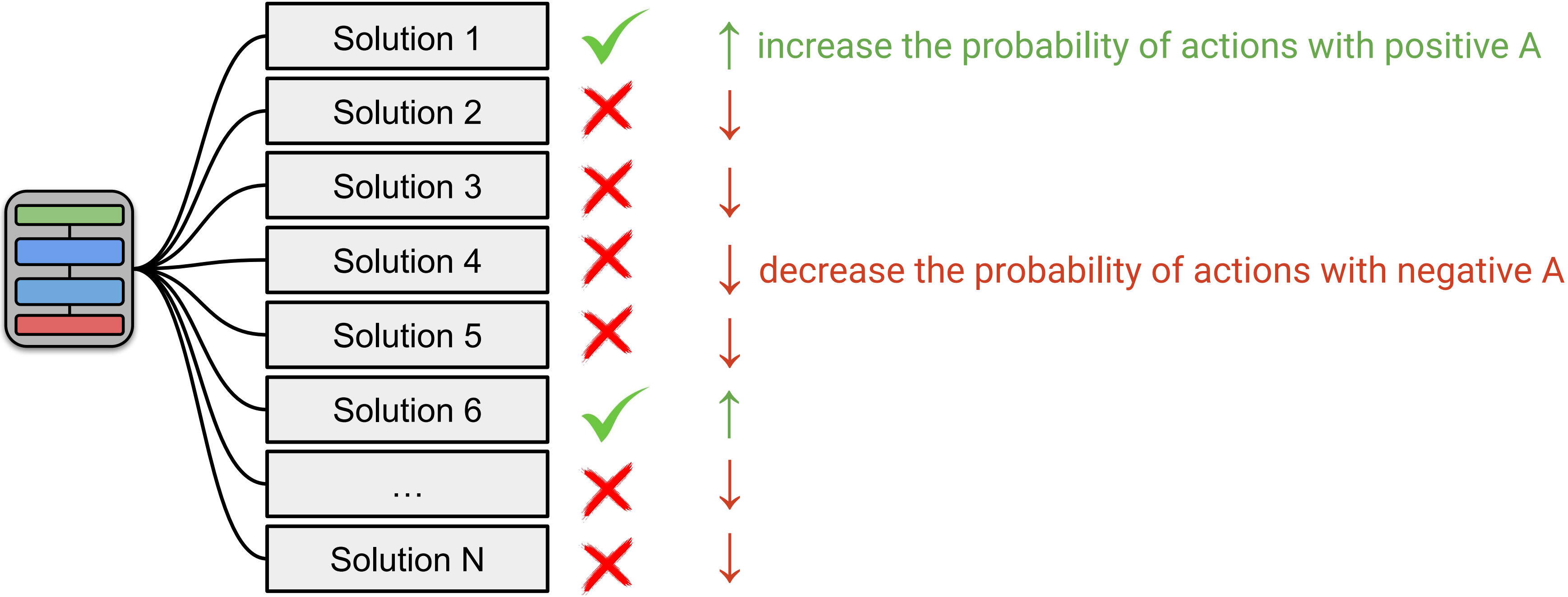
Improving CoT reasoning



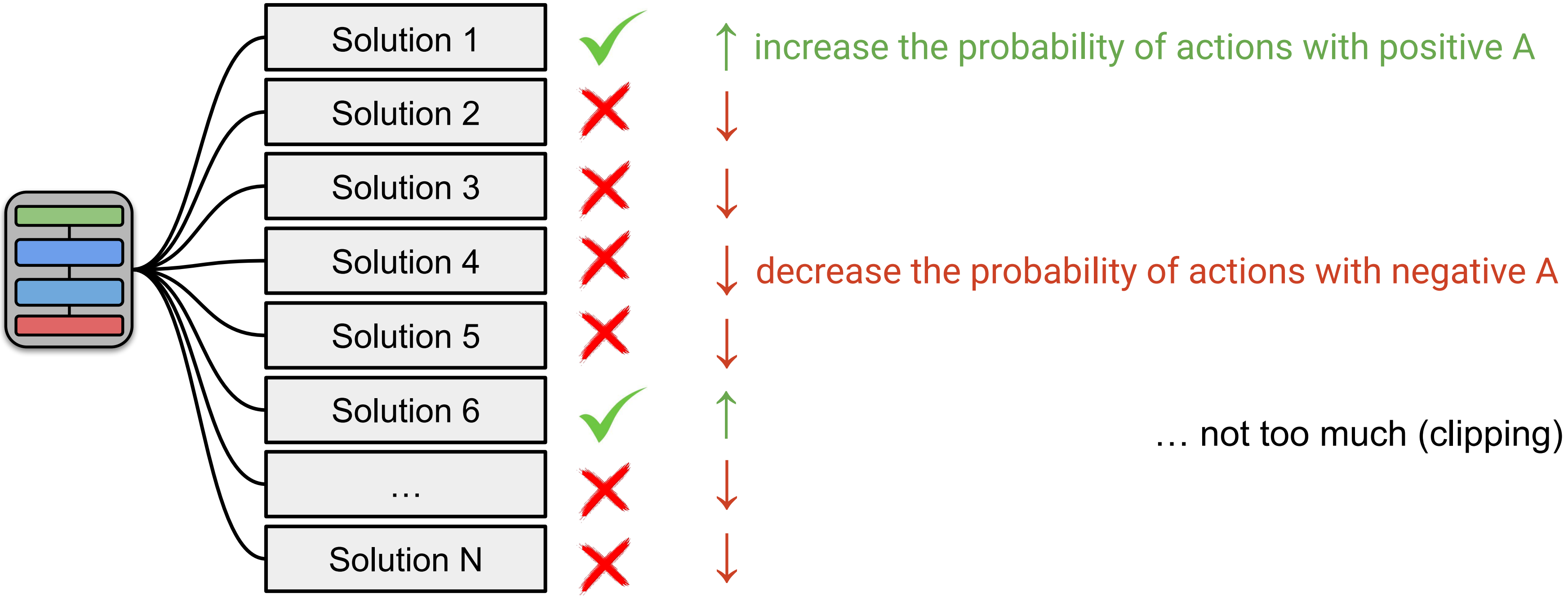
Improving CoT reasoning



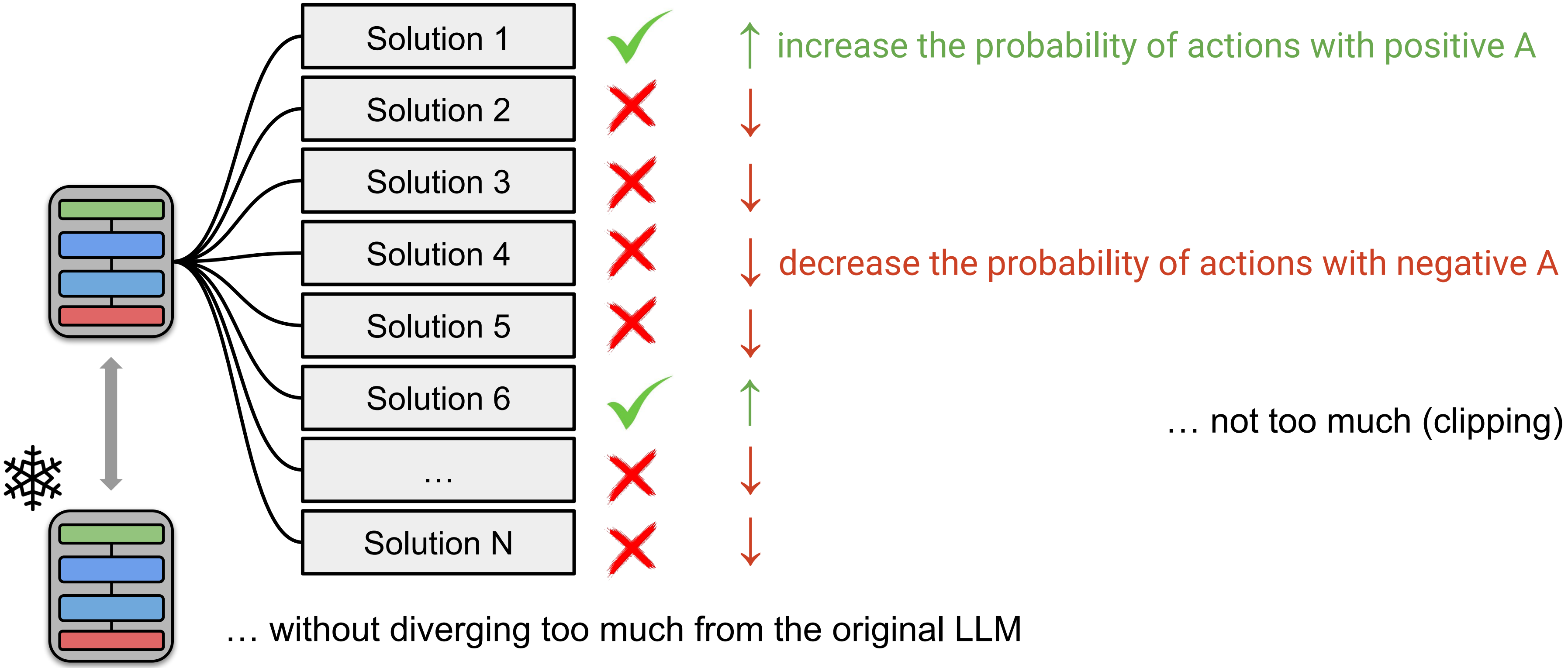
Improving CoT reasoning



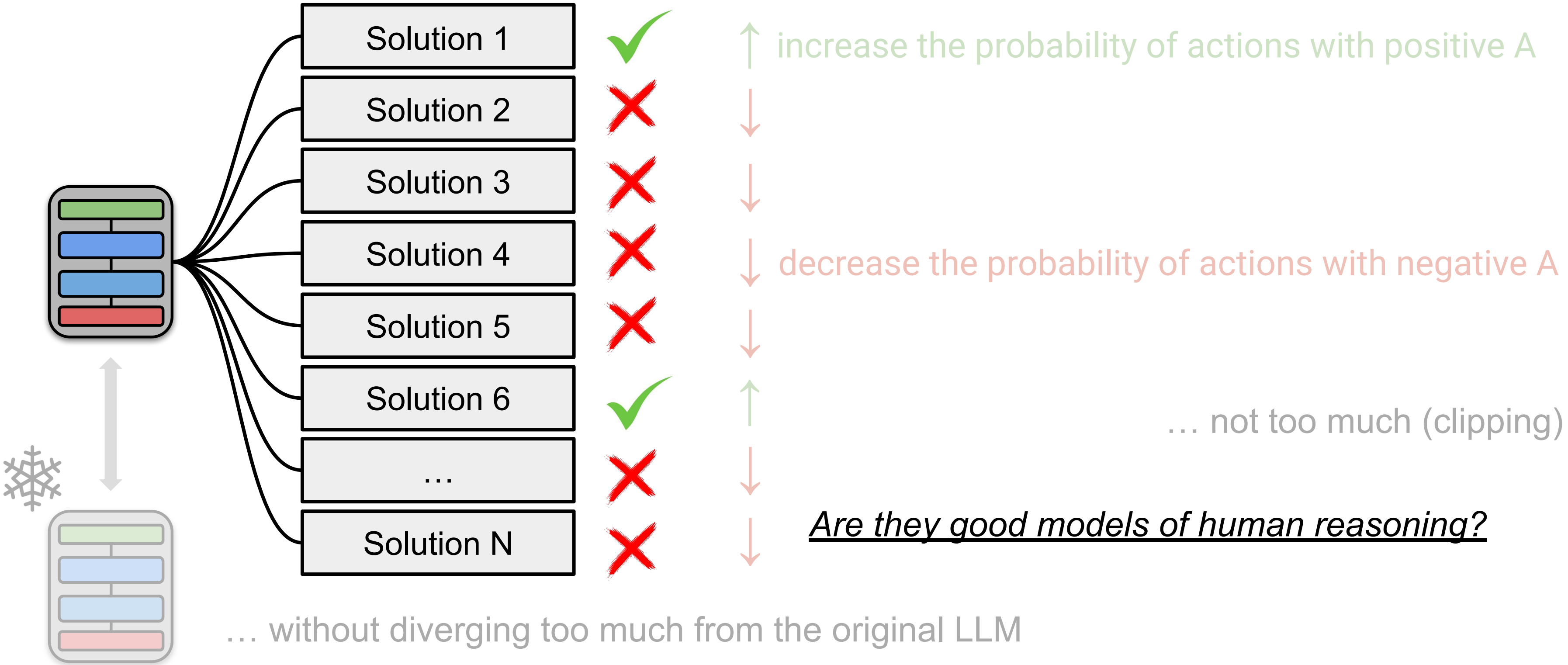
Improving CoT reasoning



Improving CoT reasoning



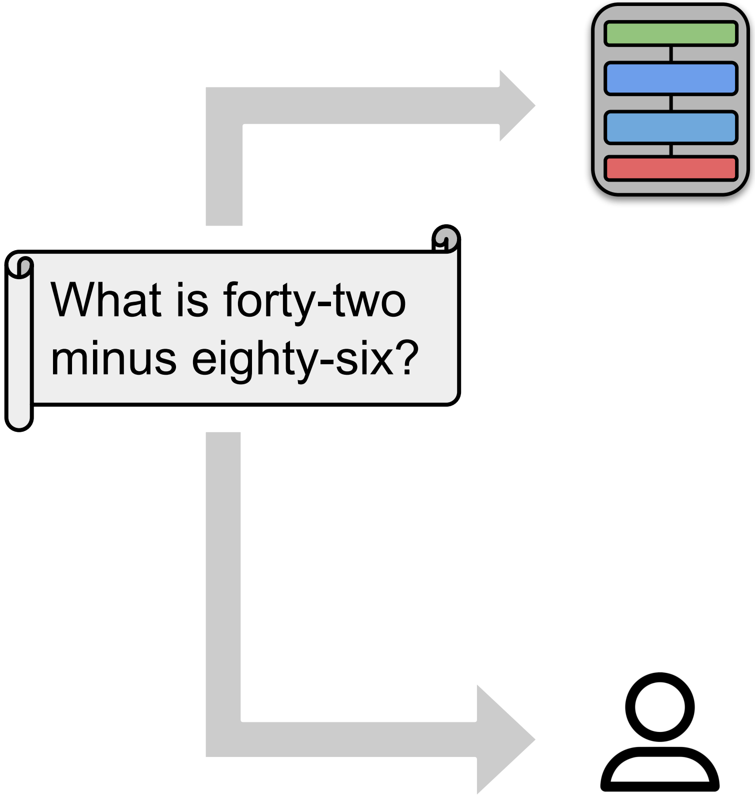
Improving CoT reasoning



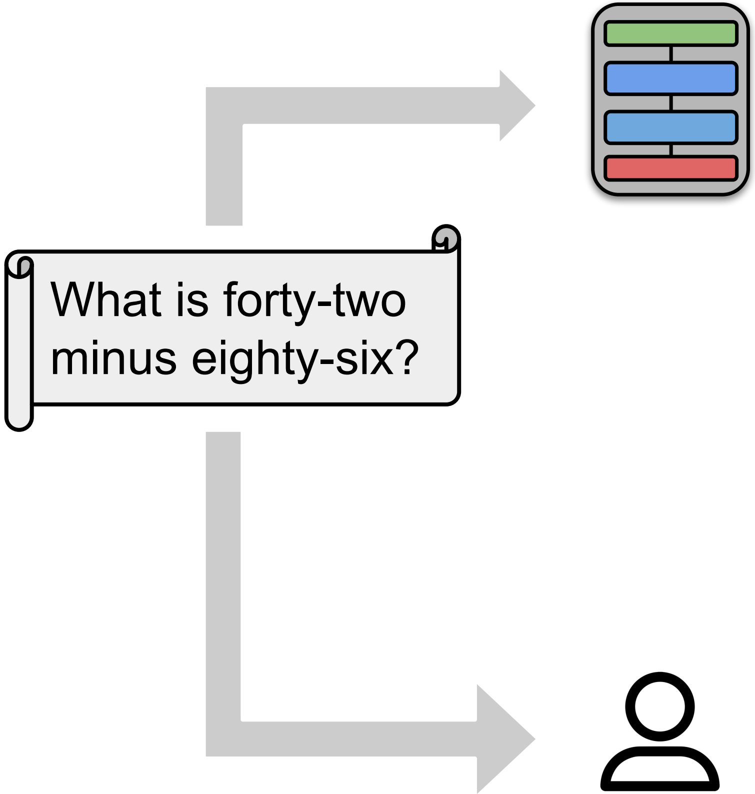
Approach

What is forty-two
minus eighty-six?

Approach

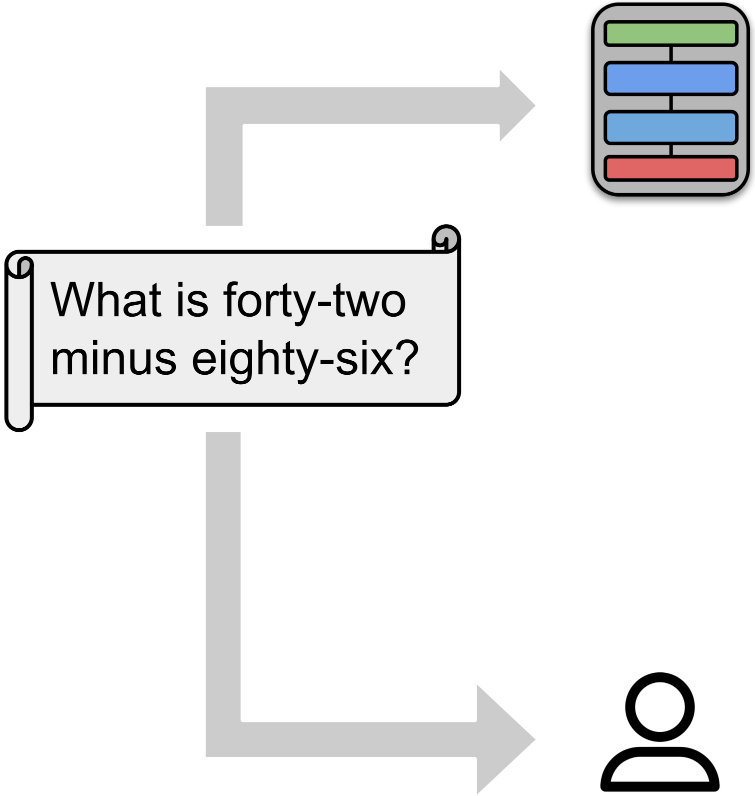


Approach



<think> First, the problem is: "What is forty-two minus eighty-six". I need to solve this mentally and respond numerically. The numbers are given verbally: forty-two and eighty-six. I know that forty-two is 42 and eighty-six is 86. So, the expression is 42 minus 86. Now, subtracting 86 from 42. Since 86 is larger than 42, the result should be negative. I need to calculate 42 - 86. To do this mentally, I can think of it as finding the difference between 86 and 42 and then making it negative because I'm subtracting a larger number. So, 86 minus 42 is 44. Therefore, 42 minus 86 should be -44. I can also think in terms of number line: starting from 42 and moving left 86 steps. But since 42 is less than 86, I'll go into negative territory. The difference is 86 - 42 = 44, so 42 - 86 = -44. That seems straightforward. I should double-check. [...] **</think>**

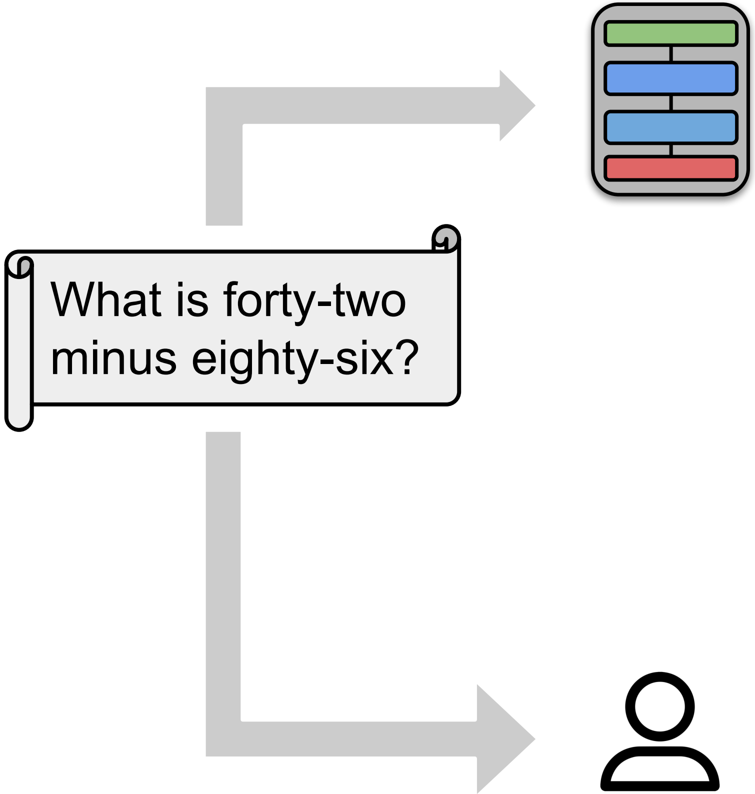
Approach



<think> First, the problem is: "What is forty-two minus eighty-six". I need to solve this mentally and respond numerically. The numbers are given verbally: forty-two and eighty-six. I know that forty-two is 42 and eighty-six is 86. So, the expression is 42 minus 86. Now, subtracting 86 from 42. Since 86 is larger than 42, the result should be negative. I need to calculate 42 - 86. To do this mentally, I can think of it as finding the difference between 86 and 42 and then making it negative because I'm subtracting a larger number. So, 86 minus 42 is 44. Therefore, 42 minus 86 should be -44. I can also think in terms of number line: starting from 42 and moving left 86 steps. But since 42 is less than 86, I'll go into negative territory. The difference is 86 - 42 = 44, so 42 - 86 = -44. That seems straightforward. I should double-check. [...] **</think>**

Answer: -44

Approach

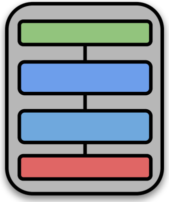


<think> First, the problem is: "What is forty-two minus eighty-six". I need to solve this mentally and respond numerically. The numbers are given verbally: forty-two and eighty-six. I know that forty-two is 42 and eighty-six is 86. So, the expression is 42 minus 86. Now, subtracting 86 from 42. Since 86 is larger than 42, the result should be negative. I need to calculate 42 - 86. To do this mentally, I can think of it as finding the difference between 86 and 42 and then making it negative because I'm subtracting a larger number. So, 86 minus 42 is 44. Therefore, 42 minus 86 should be -44. I can also think in terms of number line: starting from 42 and moving left 86 steps. But since 42 is less than 86, I'll go into negative territory. The difference is 86 - 42 = 44, so 42 - 86 = -44. That seems straightforward. I should double-check. [...] **</think>**

Answer: -44

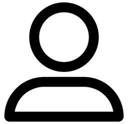
Approach

What is forty-two minus eighty-six?



<think> First, the problem is: "What is forty-two minus eighty-six". I need to solve this mentally and respond numerically. The numbers are given verbally: forty-two and eighty-six. I know that forty-two is 42 and eighty-six is 86. So, the expression is 42 minus 86. Now, subtracting 86 from 42. Since 86 is larger than 42, the result should be negative. I need to calculate 42 - 86. To do this mentally, I can think of it as finding the difference between 86 and 42 and then making it negative because I'm subtracting a larger number. So, 86 minus 42 is 44. Therefore, 42 minus 86 should be -44. I can also think in terms of number line: starting from 42 and moving left 86 steps. But since 42 is less than 86, I'll go into negative territory. The difference is 86 - 42 = 44, so 42 - 86 = -44. That seems straightforward. I should double-check. [...] **</think>**

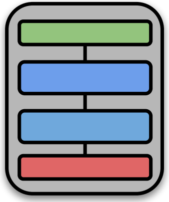
Answer: -44



Answer: -44

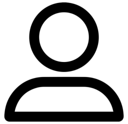
Approach

What is forty-two minus eighty-six?



<think> First, the problem is: "What is forty-two minus eighty-six". I need to solve this mentally and respond numerically. The numbers are given verbally: forty-two and eighty-six. I know that forty-two is 42 and eighty-six is 86. So, the expression is 42 minus 86. Now, subtracting 86 from 42. Since 86 is larger than 42, the result should be negative. I need to calculate 42 - 86. To do this mentally, I can think of it as finding the difference between 86 and 42 and then making it negative because I'm subtracting a larger number. So, 86 minus 42 is 44. Therefore, 42 minus 86 should be -44. I can also think in terms of number line: starting from 42 and moving left 86 steps. But since 42 is less than 86, I'll go into negative territory. The difference is 86 - 42 = 44, so 42 - 86 = -44. That seems straightforward. I should double-check. [...] **</think>**

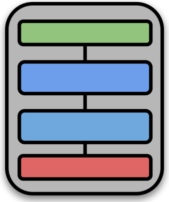
Answer: -44



Answer: -44

Approach

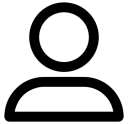
What is forty-two minus eighty-six?



<think> First, the problem is: "What is forty-two minus eighty-six". I need to solve this mentally and respond numerically. The numbers are given verbally: forty-two and eighty-six. I know that forty-two is 42 and eighty-six is 86. So, the expression is 42 minus 86. Now, subtracting 86 from 42. Since 86 is larger than 42, the result should be negative. I need to calculate 42 - 86. To do this mentally, I can think of it as finding the difference between 86 and 42 and then making it negative because I'm subtracting a larger number. So, 86 minus 42 is 44. Therefore, 42 minus 86 should be -44. I can also think in terms of number line: starting from 42 and moving left 86 steps. But since 42 is less than 86, I'll go into negative territory. The difference is 86 - 42 = 44, so 42 - 86 = -44. That seems straightforward. I should double-check. [...] **</think>**

Answer: -44

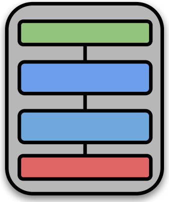
→ cost



Answer: -44

Approach

What is forty-two minus eighty-six?

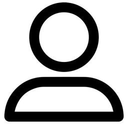


<think> First, the problem is: "What is forty-two minus eighty-six". I need to solve this mentally and respond numerically. The numbers are given verbally: forty-two and eighty-six. I know that forty-two is 42 and eighty-six is 86. So, the expression is 42 minus 86. Now, subtracting 86 from 42. Since 86 is larger than 42, the result should be negative. I need to calculate 42 - 86. To do this mentally, I can think of it as finding the difference between 86 and 42 and then making it negative because I'm subtracting a larger number. So, 86 minus 42 is 44. Therefore, 42 minus 86 should be -44. I can also think in terms of number line: starting from 42 and moving left 86 steps. But since 42 is less than 86, I'll go into negative territory. The difference is 86 - 42 = 44, so 42 - 86 = -44. That seems straightforward. I should double-check. [...] **</think>**

Answer: -44

→ cost

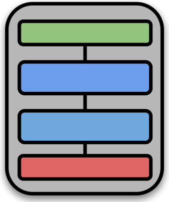
Cost →



Answer: -44

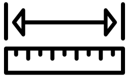
Approach

What is forty-two minus eighty-six?



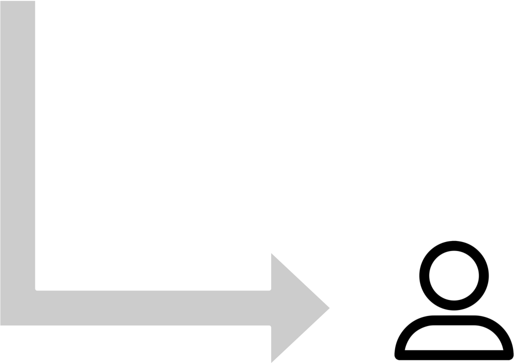
<think> First, the problem is: "What is forty-two minus eighty-six". I need to solve this mentally and respond numerically. The numbers are given verbally: forty-two and eighty-six. I know that forty-two is 42 and eighty-six is 86. So, the expression is 42 minus 86. Now, subtracting 86 from 42. Since 86 is larger than 42, the result should be negative. I need to calculate 42 - 86. To do this mentally, I can think of it as finding the difference between 86 and 42 and then making it negative because I'm subtracting a larger number. So, 86 minus 42 is 44. Therefore, 42 minus 86 should be -44. I can also think in terms of number line: starting from 42 and moving left 86 steps. But since 42 is less than 86, I'll go into negative territory. The difference is 86 - 42 = 44, so 42 - 86 = -44. That seems straightforward. I should double-check. [...] **</think>**

Answer: -44

Cost → 

Cost → 

 → cost



Answer: -44

Tasks and problems

Arithmetic, numeric

42 - 86

Addition / subtraction

Tasks and problems

Arithmetic, numeric	**Arithmetic, verbal**
42 - 86	*What is forty-two minus eighty-six?*

Addition / subtraction

Tasks and problems

Arithmetic, numeric

42 - 86

Arithmetic, verbal

What is forty-two minus eighty-six?

Syllogism

If the hak is both gop and grix then it is not ik.
The hak is gop.

The hak is not ik.

Addition / subtraction

Modus ponens / tollens — words / non-words

Tasks and problems

UNPUBLISHED DATA

Arithmetic, numeric
42 - 86

Arithmetic, verbal
What is forty-two minus eighty-six?

Syllogism
If the hak is both gop and grix then it is not ik.
The hak is gop.

The hak is not ik.

Addition / subtraction

Modus ponens / tollens — words / non-words

Tasks and problems

UNPUBLISHED DATA

Arithmetic, numeric
42 - 86

Arithmetic, verbal
What is forty-two minus eighty-six?

Addition / subtraction

Syllogism
If the hak is both gop and grix then it is not ik.
The hak is gop.

The hak is not ik.

Modus ponens / tollens — words / non-words

Logic-ALE
{a:(A⊓B), a:(B⊓C), a:(C⊓D), a:(D⊓¬A)}

Description logic (subset of FOL)

Tasks and problems

UNPUBLISHED DATA

Arithmetic, numeric
42 - 86

Arithmetic, verbal
What is forty-two minus eighty-six?

Addition / subtraction

Syllogism
If the hak is both gop and grix then it is not ik.
The hak is gop.

The hak is not ik.

Modus ponens / tollens — words / non-words

Logic-ALE
{a:(A⊓B), a:(B⊓C), a:(C⊓D), a:(D⊓¬A)}

Description logic (subset of FOL)

Relational reasoning
Caleb is less patient and more helpful than Travis.
Logan is more patient and more helpful than Travis.
LOGAN IS LESS PATIENT AND MORE HELPFUL THAN CALEB.
Yes ✓ / No ✗

Spatial / non-spatial problems

Tasks and problems

UNPUBLISHED DATA

Task	Example	Description
Arithmetic, numeric	42 - 86	Addition / subtraction
Arithmetic, verbal	*What is forty-two minus eighty-six?*	Addition / subtraction
Syllogism	If the hak is both gop and grix then it is not ik. The hak is gop. ------------------------------ The hak is not ik.	Modus ponens / tollens — words / non-words
Logic-ALE	{a:(A⊓B), a:(B⊓C), a:(C⊓D), a:(D⊓¬A)}	Description logic (subset of FOL)
Relational reasoning	*Caleb is less patient and more helpful than Travis.* *Logan is more patient and more helpful than Travis.* *LOGAN IS LESS PATIENT AND MORE HELPFUL THAN CALEB.* Yes ✓ / No ✗	Spatial / non-spatial problems
Intuitive	Over the last year, you have been so excited to try the new restaurant in Chicago [...]. You had already studied the menu, and knew exactly what you wanted: a dish made with cereal bars and lamb [...]. Clare said: 'Don't worry. I've already ordered for both of us, my treat.' [...] The waiter placed a beautiful dish in front of you made with cereal bars and lamb. How did you feel after you seeing the dish?	Intuitive reasoning (social, physical)

Tasks and problems

UNPUBLISHED DATA

Arithmetic, numeric
42 - 86

Arithmetic, verbal
What is forty-two minus eighty-six?

Addition / subtraction

Syllogism
If the hak is both gop and grix then it is not ik.
The hak is gop.

The hak is not ik.

Modus ponens / tollens — words / non-words

Logic-ALE
$\{a:(A \sqcap B), a:(B \sqcap C), a:(C \sqcap D), a:(D \sqcap \neg A)\}$

Description logic (subset of FOL)

Relational reasoning
Caleb is less patient and more helpful than Travis.
Logan is more patient and more helpful than Travis.
LOGAN IS LESS PATIENT AND MORE HELPFUL THAN CALEB.
Yes ✓ / No ✗

Spatial / non-spatial problems

Intuitive
Over the last year, you have been so excited to try the new restaurant in Chicago [...]. You had already studied the menu, and knew exactly what you wanted: a dish made with cereal bars and lamb [...]. Clare said: 'Don't worry. I've already ordered for both of us, my treat.' [...] The waiter placed a beautiful dish in front of you made with cereal bars and lamb. How did you feel after you seeing the dish?

Intuitive reasoning (social, physical)

DOMAIN-SPECIFIC

Tasks and problems

Arithmetic, numeric
42 - 86

Arithmetic, verbal
What is forty-two minus eighty-six?

Syllogism
If the hak is both gop and grix then it is not ik.
The hak is gop.

The hak is not ik.

Logic-ALE
{a:(A⊓B), a:(B⊓C), a:(C⊓D), a:(D⊓¬A)}

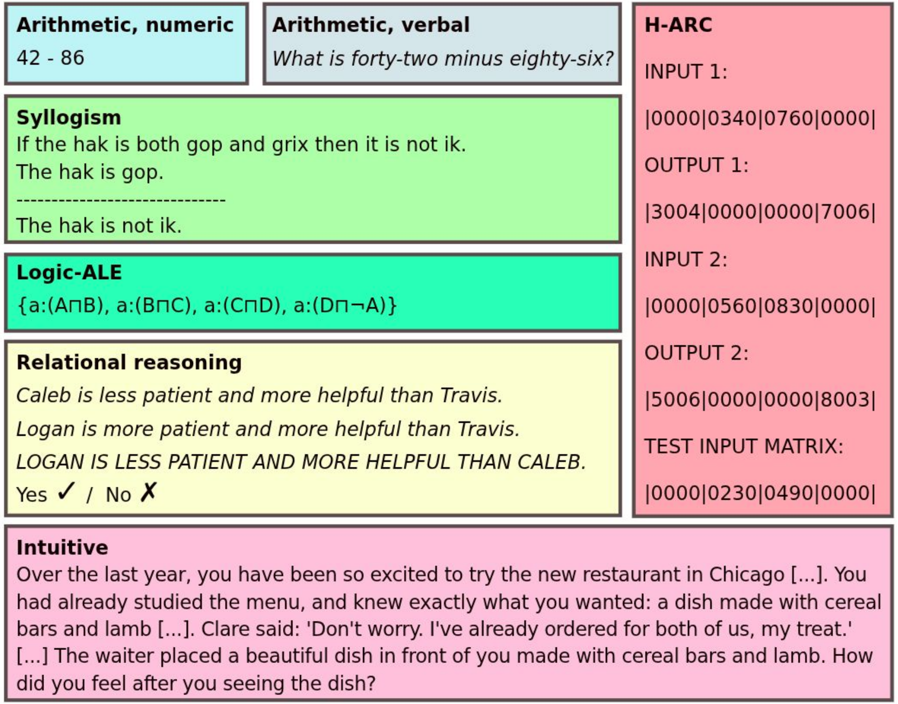
Relational reasoning
Caleb is less patient and more helpful than Travis.
Logan is more patient and more helpful than Travis.
LOGAN IS LESS PATIENT AND MORE HELPFUL THAN CALEB.
Yes ✓ / No ✗

H-ARC
INPUT 1:
|0000|0340|0760|0000|
OUTPUT 1:
|3004|0000|0000|7006|
INPUT 2:
|0000|0560|0830|0000|
OUTPUT 2:
|5006|0000|0000|8003|
TEST INPUT MATRIX:
|0000|0230|0490|0000|

Intuitive
Over the last year, you have been so excited to try the new restaurant in Chicago [...]. You had already studied the menu, and knew exactly what you wanted: a dish made with cereal bars and lamb [...]. Clare said: 'Don't worry. I've already ordered for both of us, my treat.' [...] The waiter placed a beautiful dish in front of you made with cereal bars and lamb. How did you feel after you seeing the dish?

Transformation-based grid problems

Tasks and problems



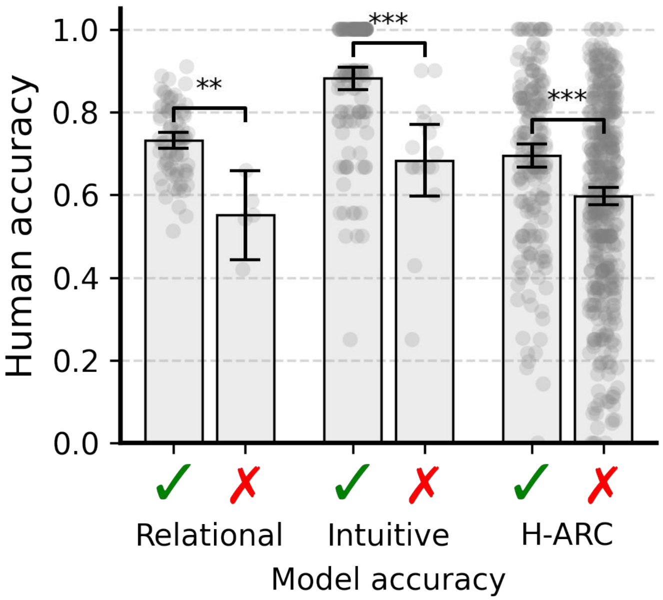
Transformation-based grid problems

High reliability

Split-half ρ = 0.60–0.93

Results – Accuracy

We evaluated an open-weights large reasoning model (DeepSeek-R1)

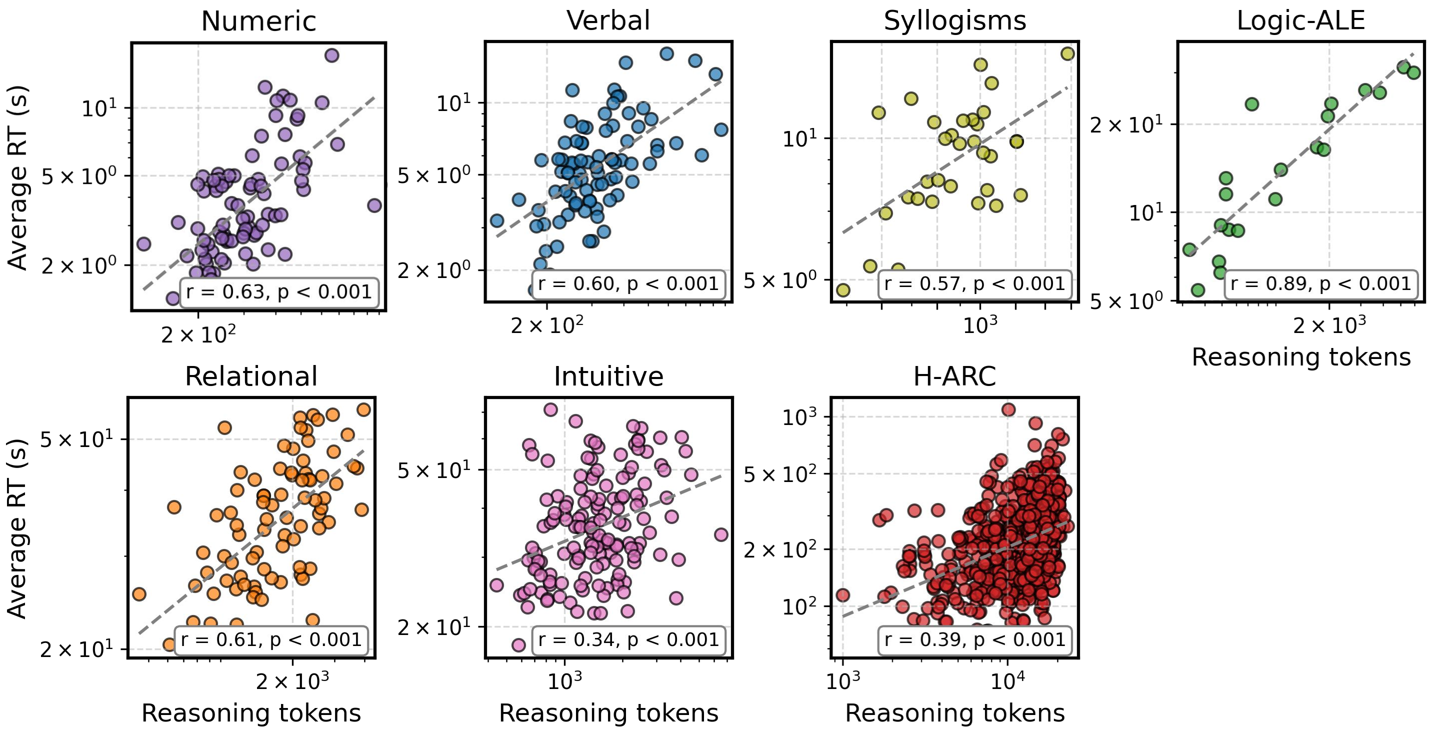


R1 achieved higher accuracy on items that humans tended to solve correctly

→ shared sensitivity to problem difficulty

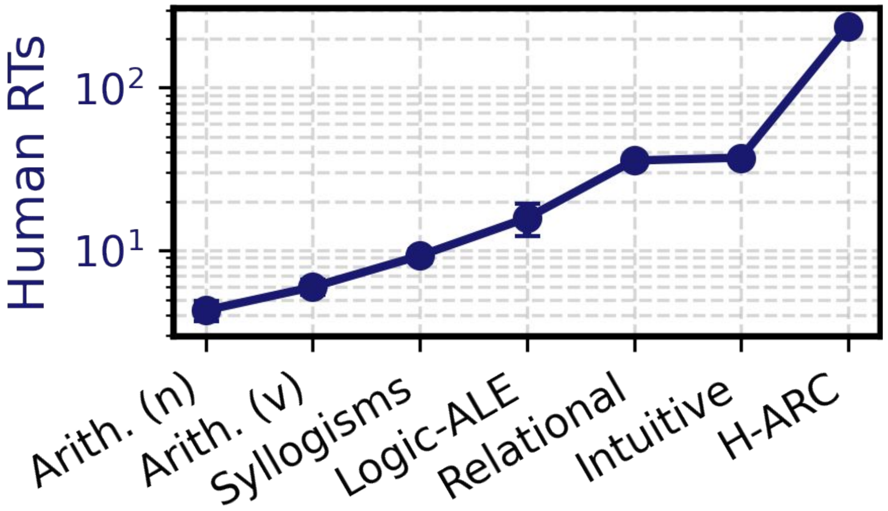
Results – RTs

The number of tokens produced by DeepSeek-R1 correlates with human RTs within tasks



Differences across tasks

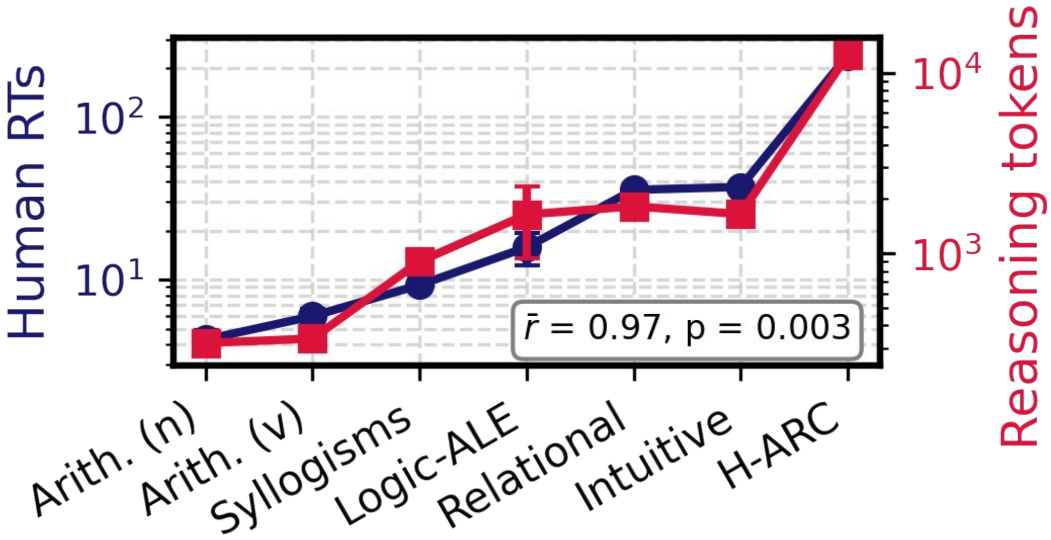
A general model of human reasoning should not only account for problem difficulty within tasks, but also capture differences *across* tasks



Humans find some tasks more difficult than others

Differences across tasks

A general model of human reasoning should not only account for problem difficulty within tasks, but also capture differences *across* tasks



Humans find some tasks more difficult than others

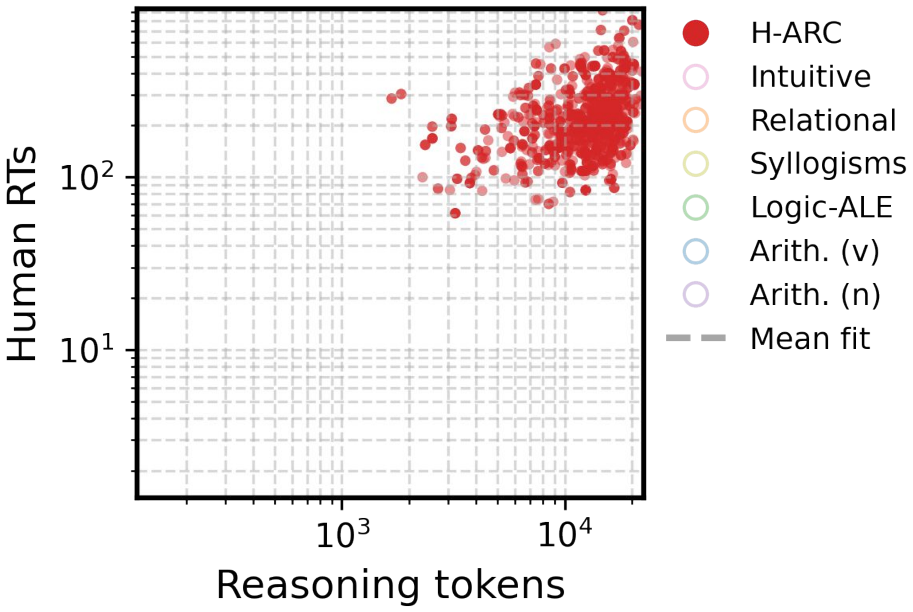
Models mirror broad differences in cognitive demand across domains

Differences across tasks

A generalized metric of reasoning cost should predict RTs for single problems across tasks

Differences across tasks

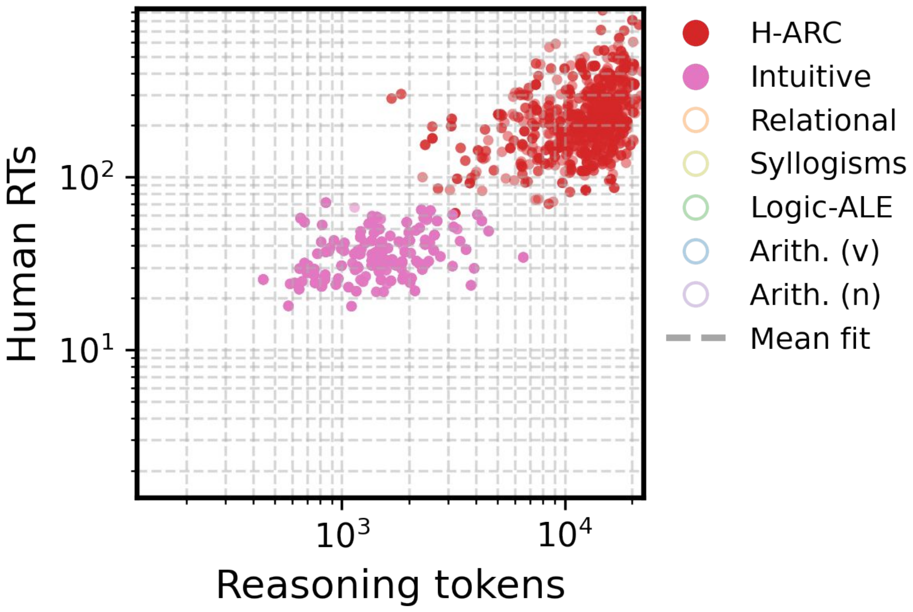
A generalized metric of reasoning cost should predict RTs for single problems across tasks



→ controlling for dataset size with repeated subsampling

Differences across tasks

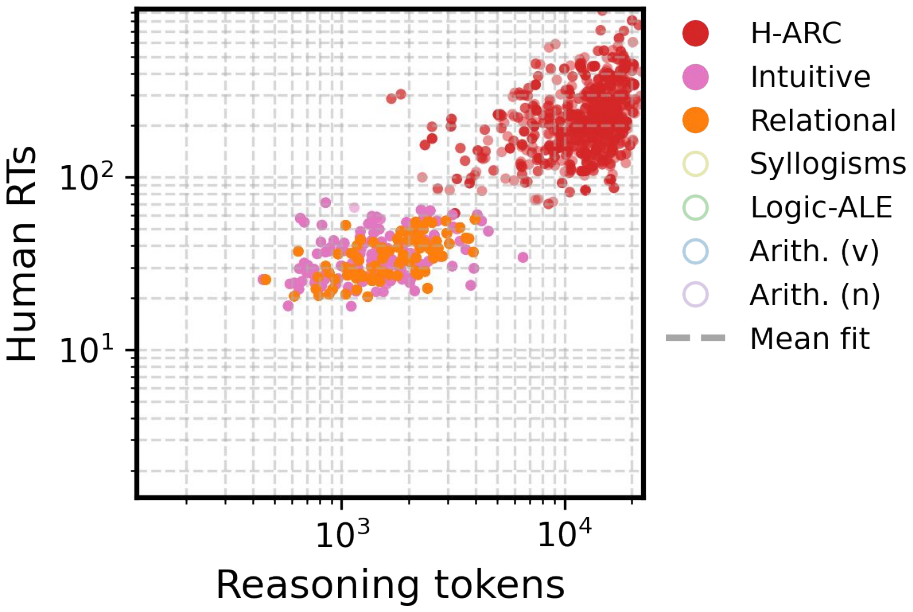
A generalized metric of reasoning cost should predict RTs for single problems across tasks



→ controlling for dataset size with repeated subsampling

Differences across tasks

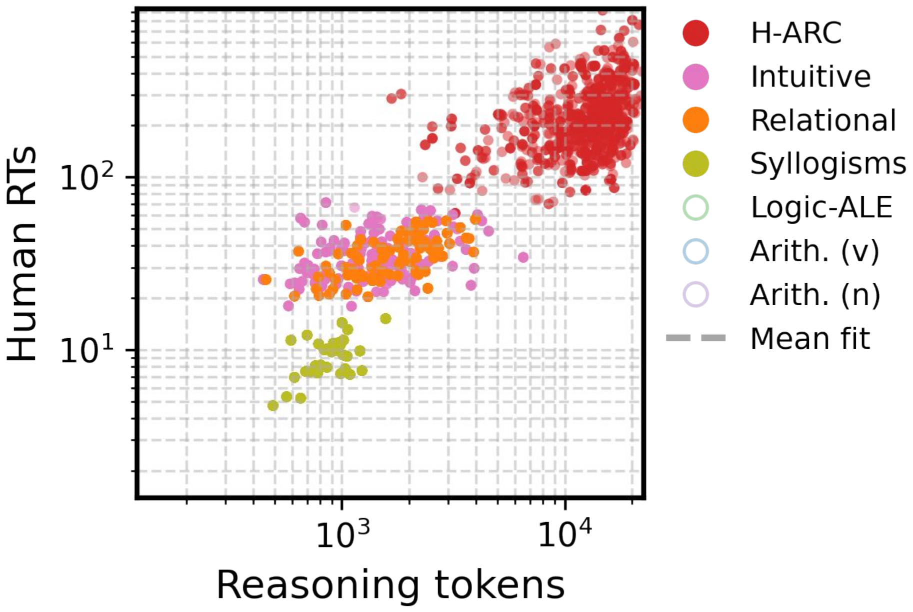
A generalized metric of reasoning cost should predict RTs for single problems across tasks



→ controlling for dataset size with repeated subsampling

Differences across tasks

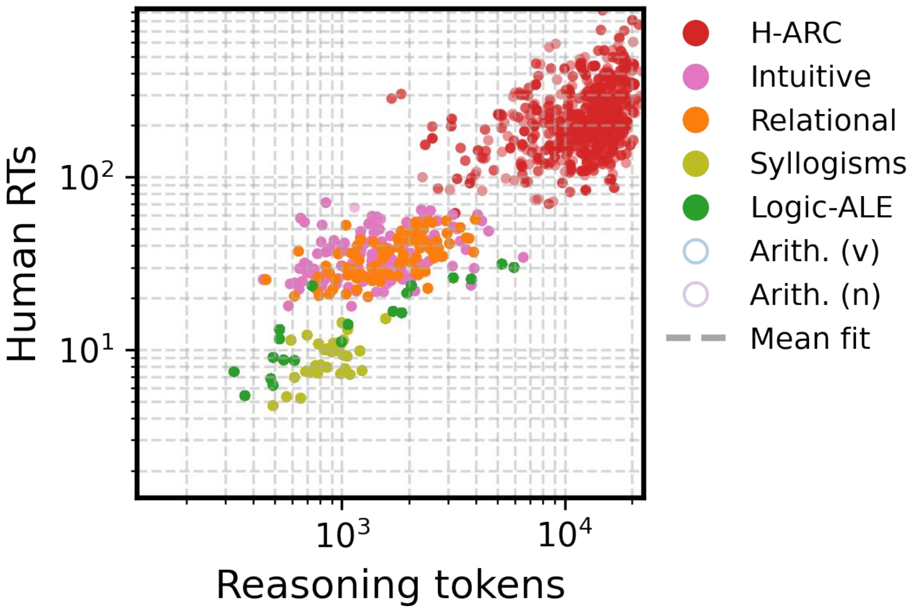
A generalized metric of reasoning cost should predict RTs for single problems across tasks



→ controlling for dataset size with repeated subsampling

Differences across tasks

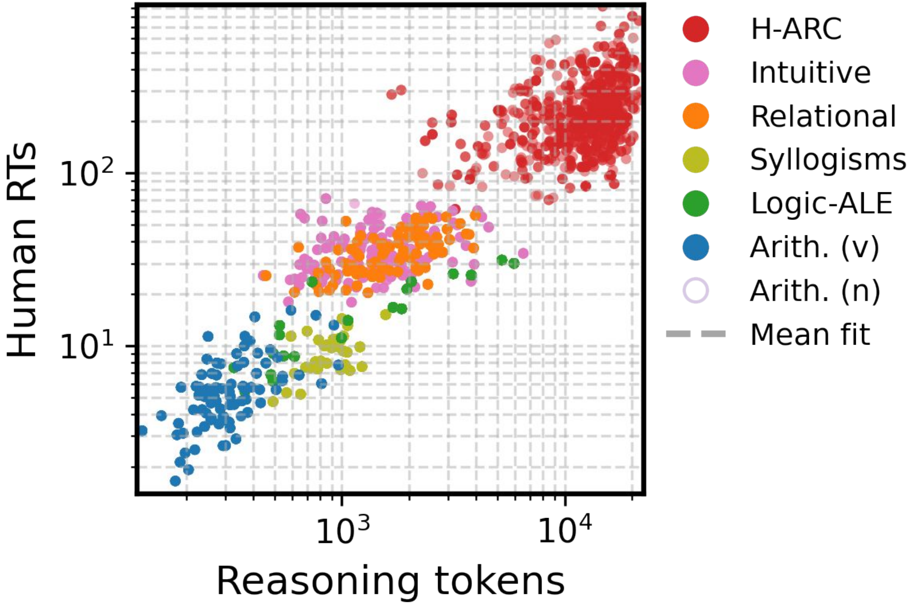
A generalized metric of reasoning cost should predict RTs for single problems across tasks



→ controlling for dataset size with repeated subsampling

Differences across tasks

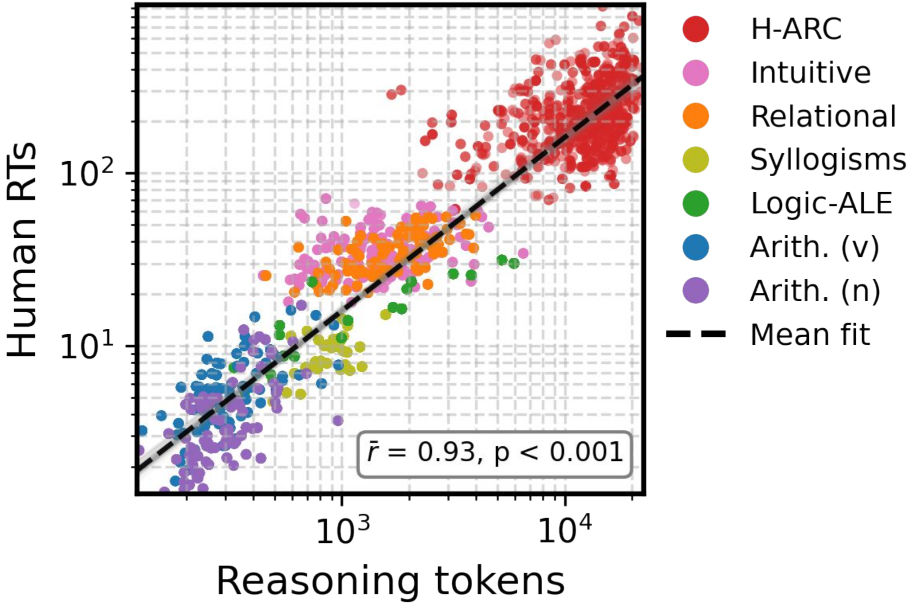
A generalized metric of reasoning cost should predict RTs for single problems across tasks



→ controlling for dataset size with repeated subsampling

Differences across tasks

A generalized metric of reasoning cost should predict RTs for single problems across tasks



→ controlling for dataset size with repeated subsampling

R1's reasoning demands scale with human effort both within and across domains

What drives the alignment

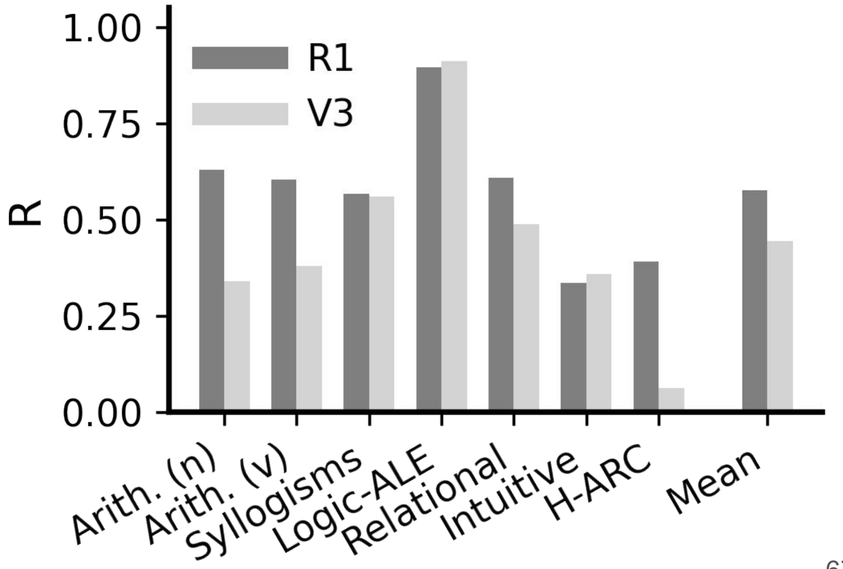
RL

The effect was substantially weaker for R1's base model, DeepSeek-V3

$\bar{r} = 0.44$ vs. $\bar{r} = 0.57$ for R1

$z = 4.39$, $p < 0.001$

Reasoning-optimized training increases the model's alignment with human processing effort



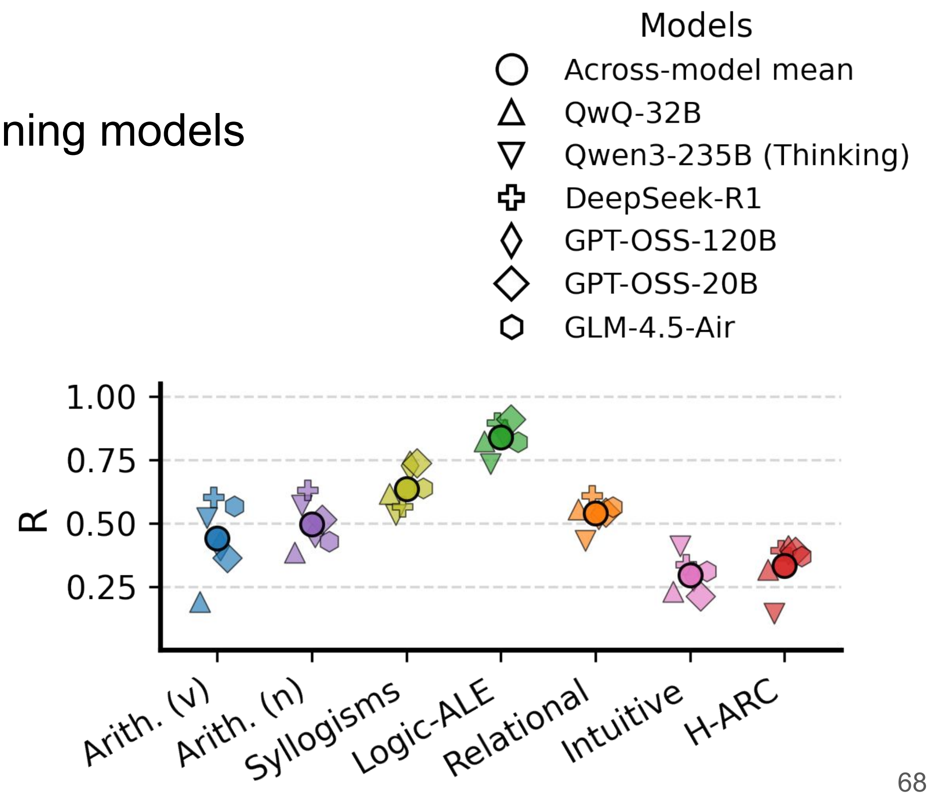
What drives the alignment

DeepSeek R1 specifically? ❌

The effect generalized to six different reasoning models

Small inter-model differences

Variance explained by:
→ Differences between tasks: 80.25%
→ Differences between models: 4.58%

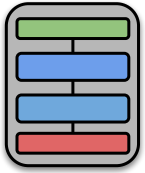


Why?

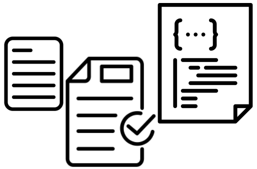
Large reasoning models show strong alignment with human reasoning behavior. One possible explanation for this convergence is that reasoning models follow a *learning trajectory* that resembles human acquisition of new reasoning skills

Why?

Large reasoning models show strong alignment with human reasoning behavior. One possible explanation for this convergence is that reasoning models follow a *learning trajectory* that resembles human acquisition of new reasoning skills



Pre-training

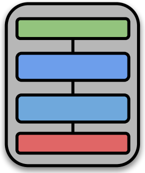


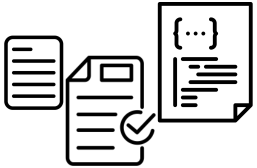
observe a wide range of correct reasoning examples

Why?

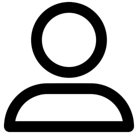
Large reasoning models show strong alignment with human reasoning behavior. One possible explanation for this convergence is that reasoning models follow a *learning trajectory* that resembles human acquisition of new reasoning skills

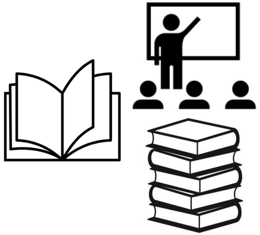
Pre-training





observe a wide range of correct reasoning examples



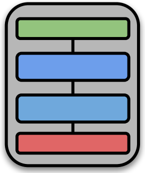


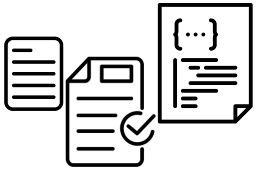
learn from textbooks or instruction

Why?

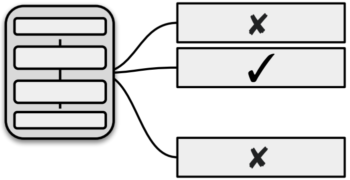
Large reasoning models show strong alignment with human reasoning behavior. One possible explanation for this convergence is that reasoning models follow a *learning trajectory* that resembles human acquisition of new reasoning skills

Pre-training **RL**

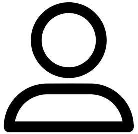


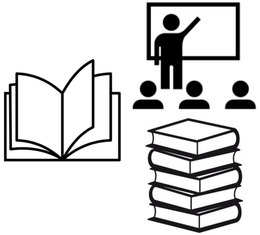


observe a wide range of correct reasoning examples



learn from feedback





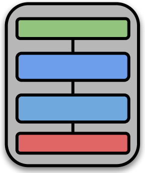
learn from textbooks or instruction

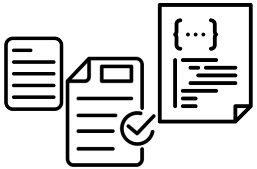
Why?

Large reasoning models show strong alignment with human reasoning behavior. One possible explanation for this convergence is that reasoning models follow a *learning trajectory* that resembles human acquisition of new reasoning skills

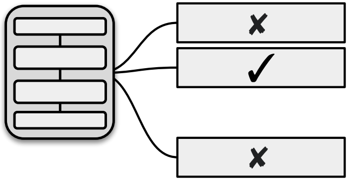
Pre-training

RL

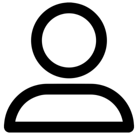


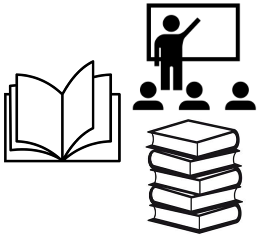


observe a wide range of correct reasoning examples

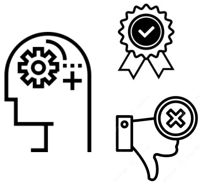


learn from feedback





learn from textbooks or instruction



practicing problems and learning from success or failure

Model internals

Reasoning-optimized LLMs allocate resources to cognitive tasks in a way that is similar to humans.

→ Does the **internal organization** of the models’ reasoning systems mirror the human brain?

Reasoning systems in the human brain

Intelligent behavior in humans is supported by a set of distributed brain networks that are functionally specialized for certain cognitive domains.

Kanwisher et al., 1997; Saxe & Kanwisher, 2003; Fedorenko et al., 2011

Reasoning systems in the human brain

Intelligent behavior in humans is supported by a set of distributed brain networks that are functionally specialized for certain cognitive domains.

Kanwisher et al., 1997; Saxe & Kanwisher, 2003; Fedorenko et al., 2011

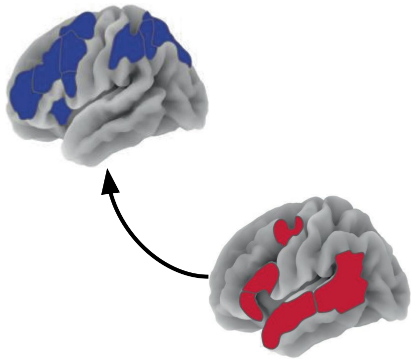
Clear segregation between the cortical regions that support:

- Linguistic processing (understanding and producing language)

Reasoning systems in the human brain

Intelligent behavior in humans is supported by a set of distributed brain networks that are functionally specialized for certain cognitive domains.

Kanwisher et al., 1997; Saxe & Kanwisher, 2003; Fedorenko et al., 2011



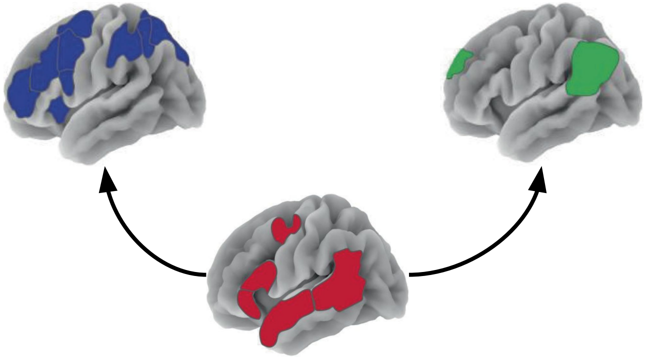
Clear segregation between the cortical regions that support:

- Linguistic processing (understanding and producing language)
- Knowledge and reasoning
  Fedorenko et al., 2024; Mahowald, Ivanova et al., 2024
  - Domain-general reasoning

Reasoning systems in the human brain

Intelligent behavior in humans is supported by a set of distributed brain networks that are functionally specialized for certain cognitive domains.

Kanwisher et al., 1997; Saxe & Kanwisher, 2003; Fedorenko et al., 2011



Clear segregation between the cortical regions that support:

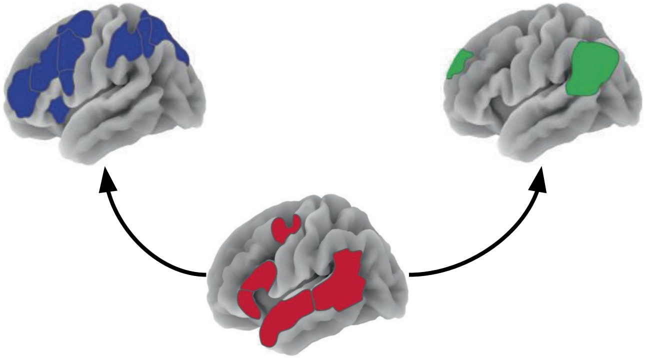
- Linguistic processing (understanding and producing language)
- Knowledge and reasoning

  Fedorenko et al., 2024; Mahowald, Ivanova et al., 2024
  - Domain-general reasoning
  - Domain-specific reasoning (theory of mind, intuitive physics, …)

Reasoning systems in the human brain

Intelligent behavior in humans is supported by a set of distributed brain networks that are functionally specialized for certain cognitive domains.

Kanwisher et al., 1997; Saxe & Kanwisher, 2003; Fedorenko et al., 2011



Clear segregation between the cortical regions that support:

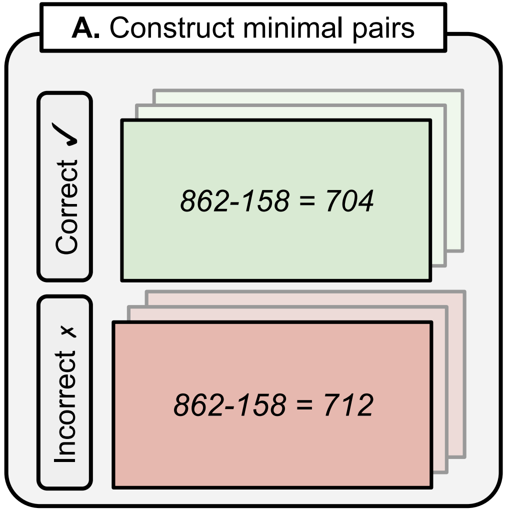
- Linguistic processing (understanding and producing language)
- Knowledge and reasoning

  Fedorenko et al., 2024; Mahowald, Ivanova et al., 2024
  - Domain-general reasoning
  - Domain-specific reasoning (theory of mind, intuitive physics, …)

→ Do we see similar segregation in LLMs?

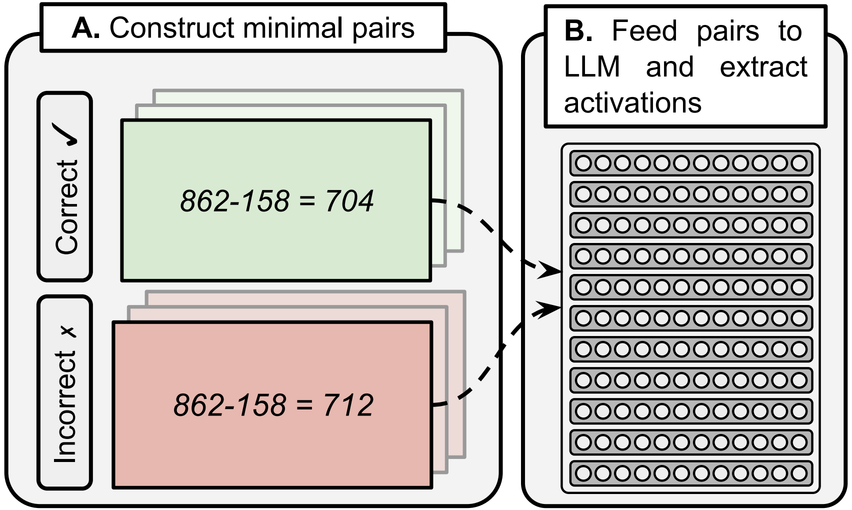
Approach

Starting from **minimal pairs** of problems followed by the correct vs. incorrect solutions, we found the units that maximally discriminate between them, and tested whether they overlap or segregate across tasks



Approach

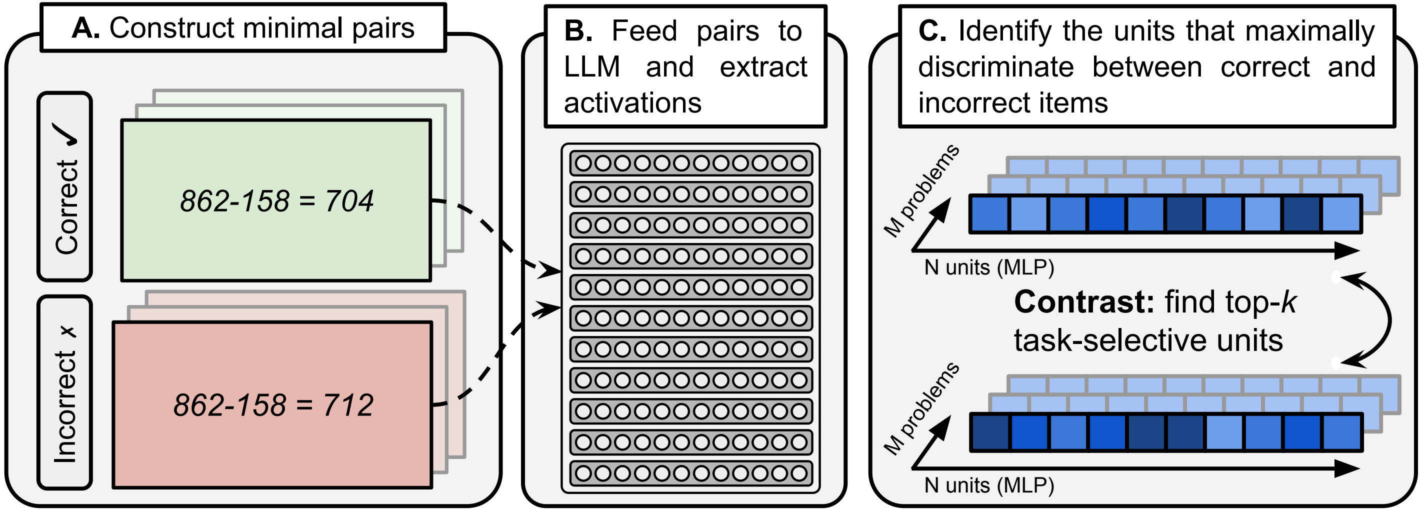
Starting from **minimal pairs** of problems followed by the correct vs. incorrect solutions, we found the units that maximally discriminate between them, and tested whether they overlap or segregate across tasks



Pengrui Han

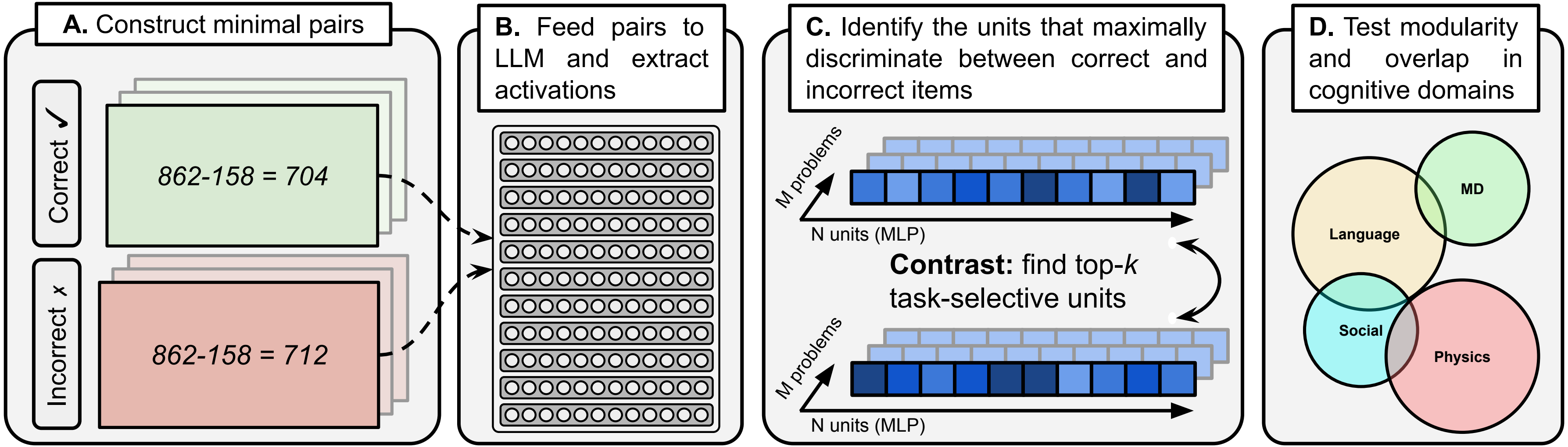
Approach

Starting from **minimal pairs** of problems followed by the correct vs. incorrect solutions, we found the units that maximally discriminate between them, and tested whether they overlap or segregate across tasks



Approach

Starting from **minimal pairs** of problems followed by the correct vs. incorrect solutions, we found the units that maximally discriminate between them, and tested whether they overlap or segregate across tasks



Modularity of reasoning systems in LLMs

We localized components in LLMs supporting **linguistic processing** and different *kinds* of **reasoning** across a total of N = 42 tasks:

Language

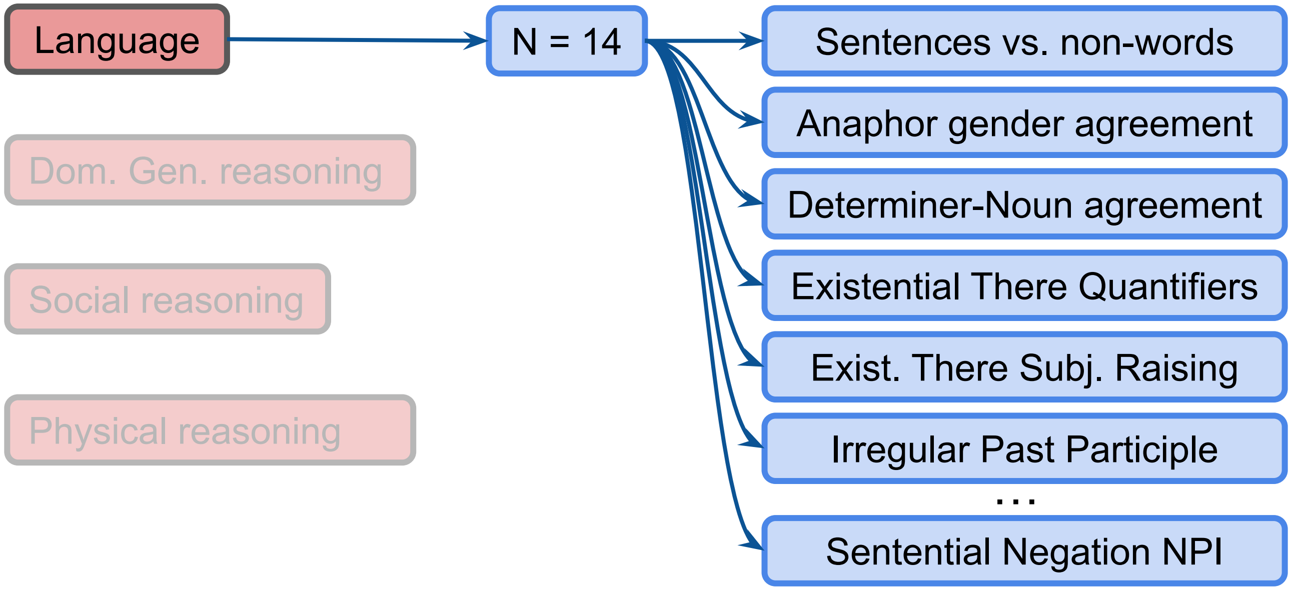
Dom. Gen. reasoning

Social reasoning

Physical reasoning

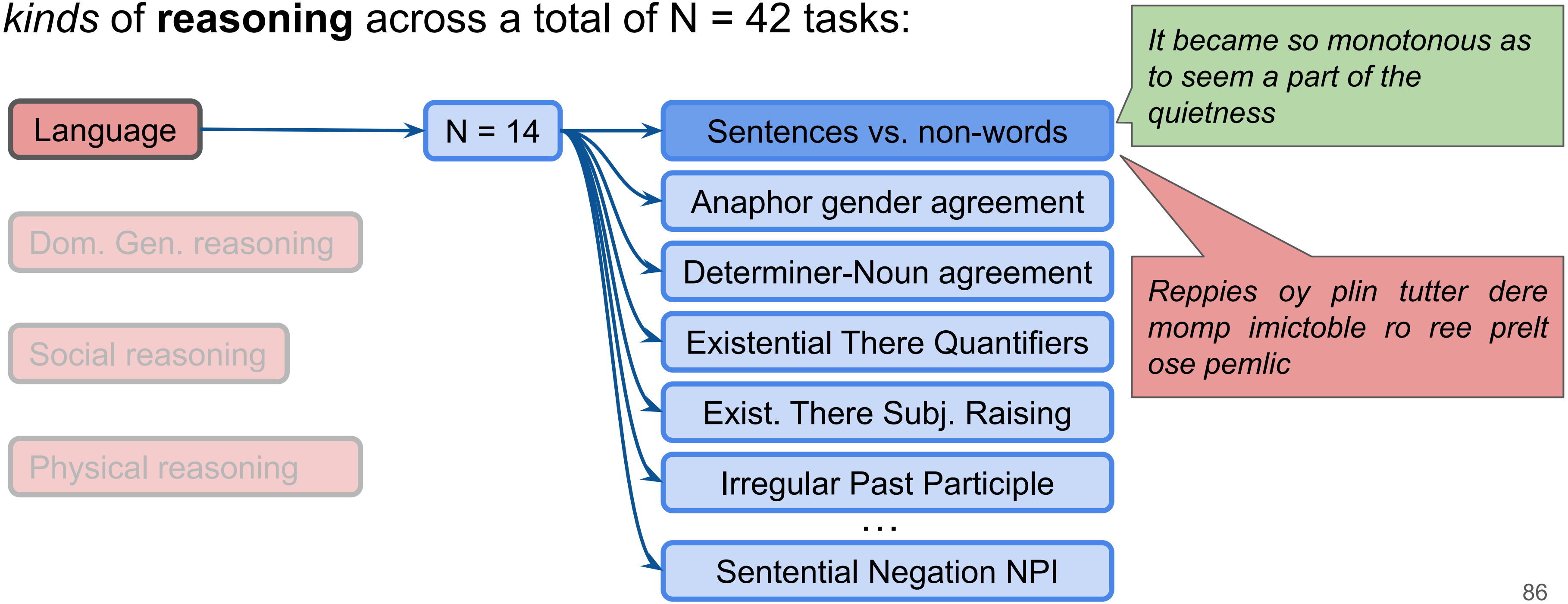
Modularity of reasoning systems in LLMs

We localized components in LLMs supporting **linguistic processing** and different *kinds* of **reasoning** across a total of N = 42 tasks:



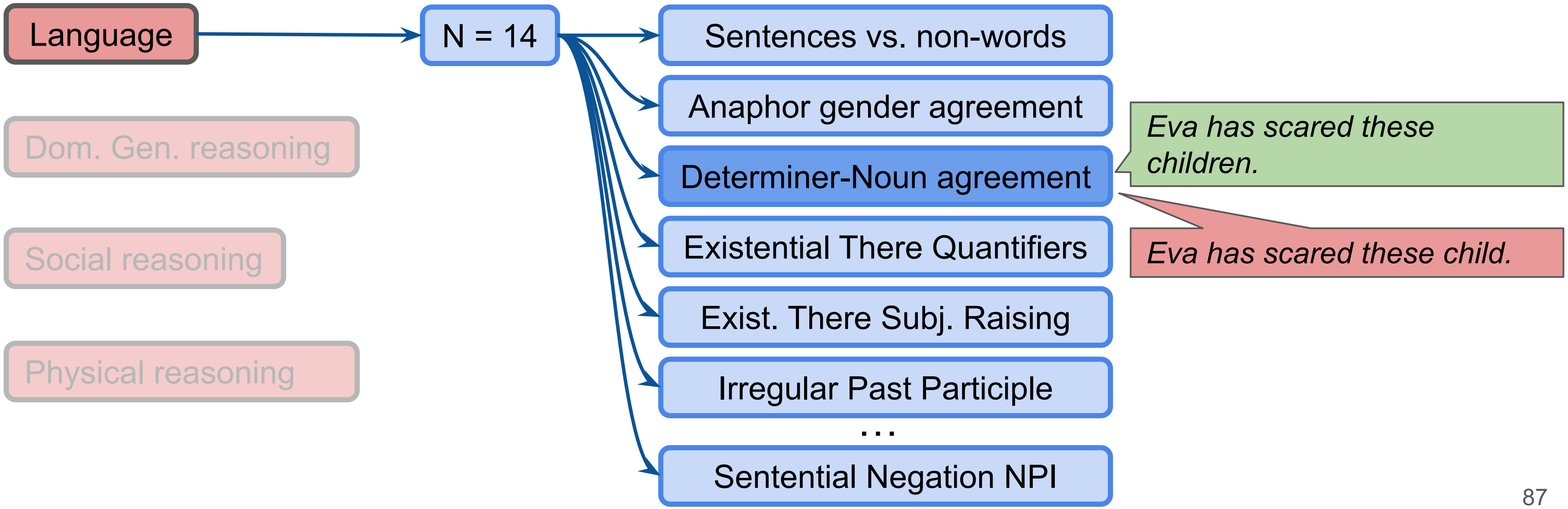
Modularity of reasoning systems in LLMs

We localized components in LLMs supporting **linguistic processing** and different *kinds* of **reasoning** across a total of N = 42 tasks:



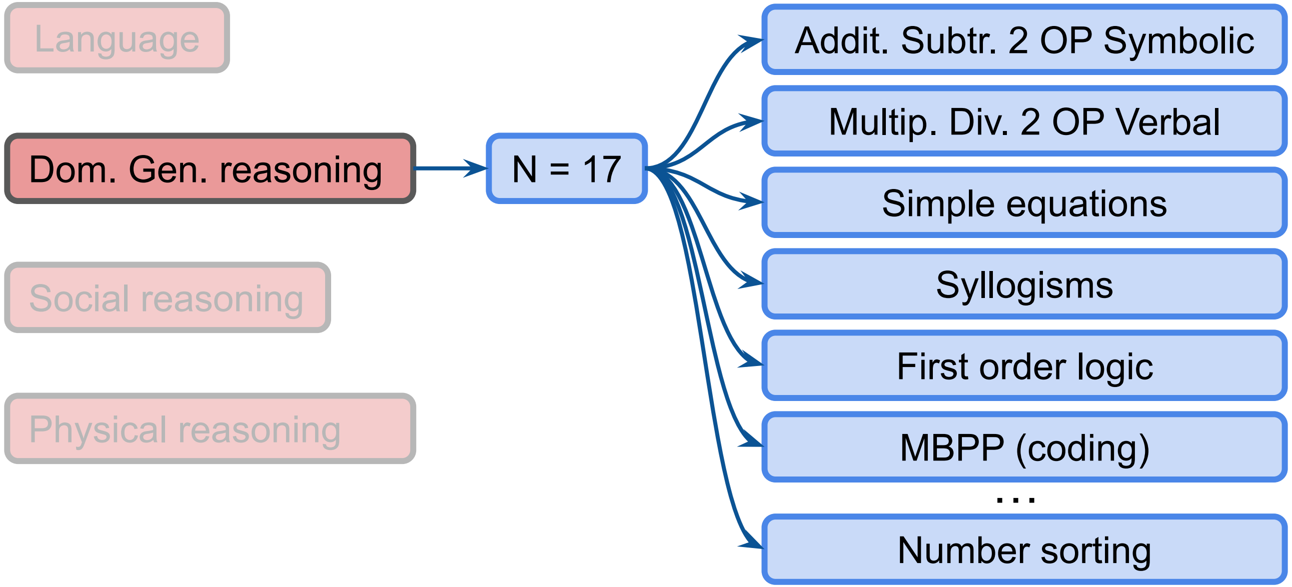
Modularity of reasoning systems in LLMs

We localized components in LLMs supporting **linguistic processing** and different *kinds* of **reasoning** across a total of N = 42 tasks:



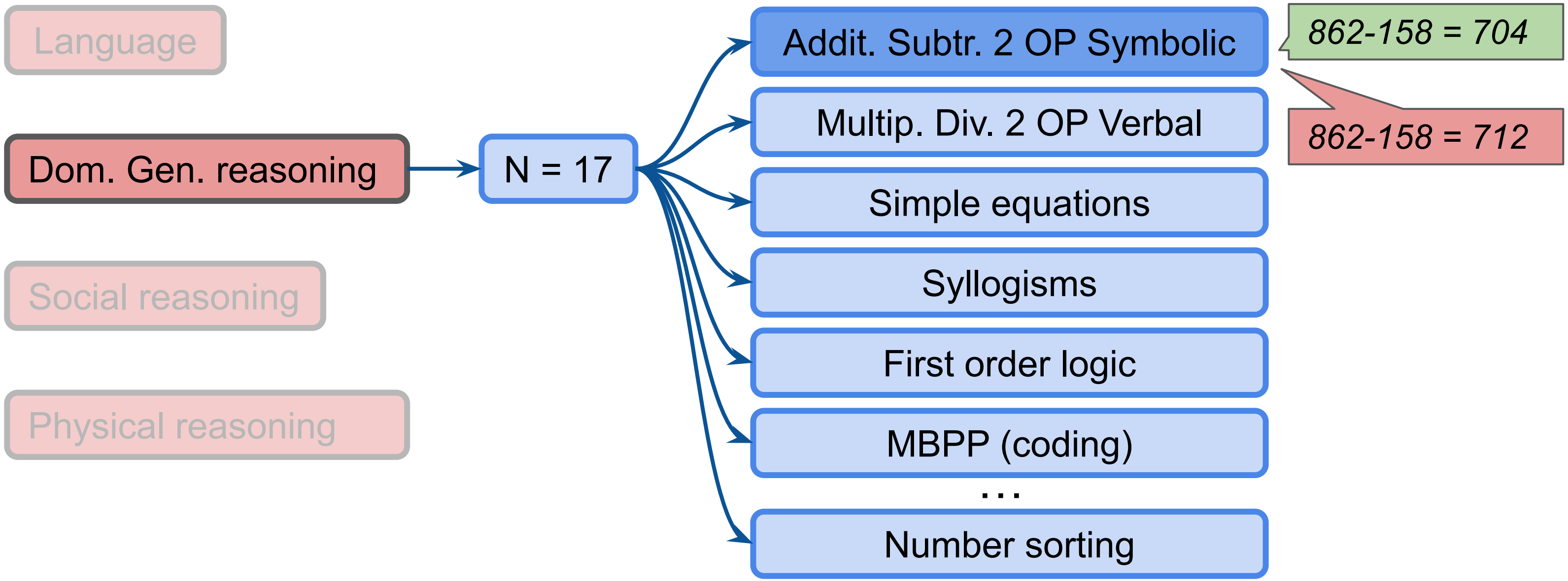
Modularity of reasoning systems in LLMs

We localized components in LLMs supporting **linguistic processing** and different *kinds* of **reasoning** across a total of N = 42 tasks:



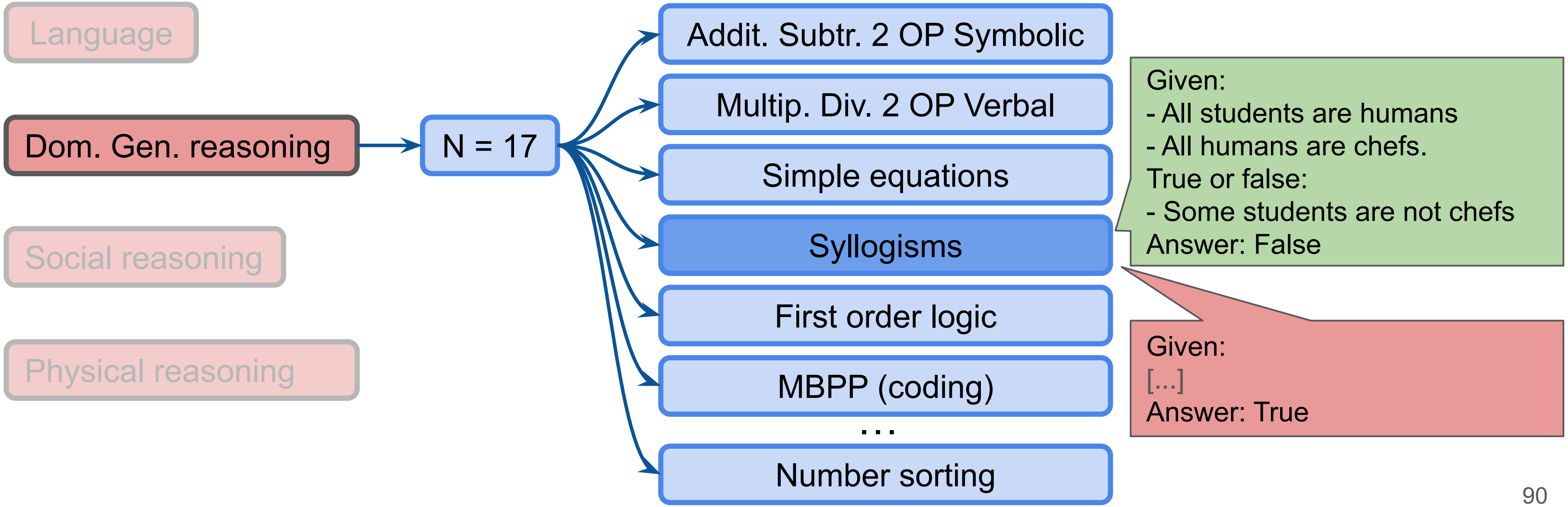
Modularity of reasoning systems in LLMs

We localized components in LLMs supporting **linguistic processing** and different *kinds* of **reasoning** across a total of N = 42 tasks:



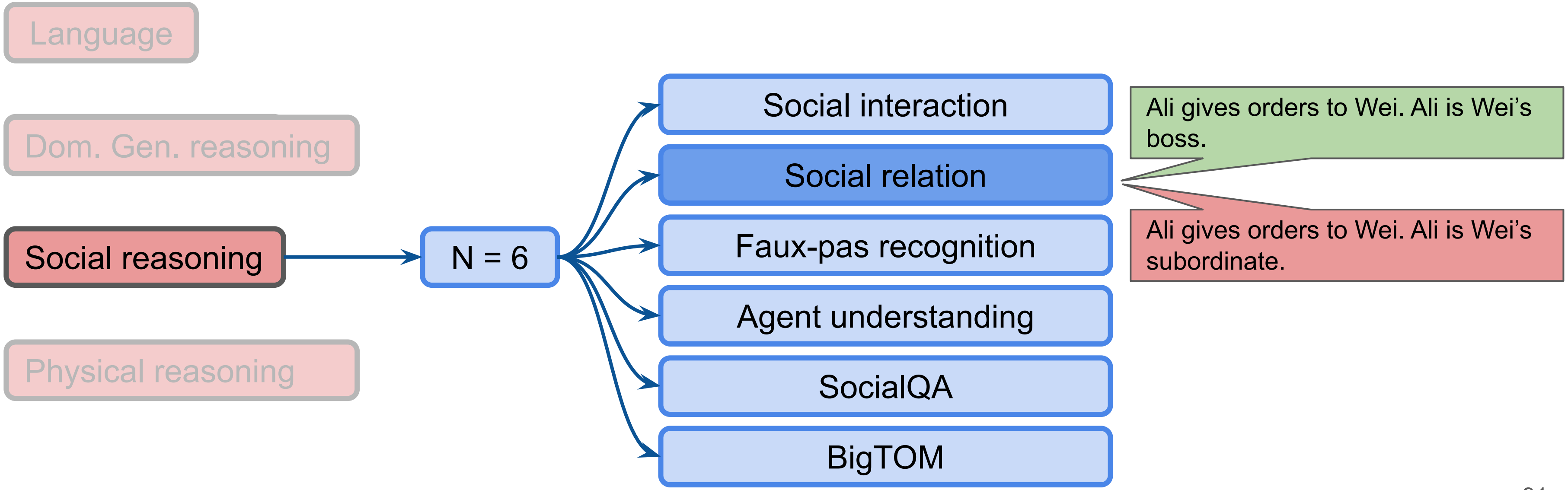
Modularity of reasoning systems in LLMs

We localized components in LLMs supporting **linguistic processing** and different *kinds* of **reasoning** across a total of N = 42 tasks:



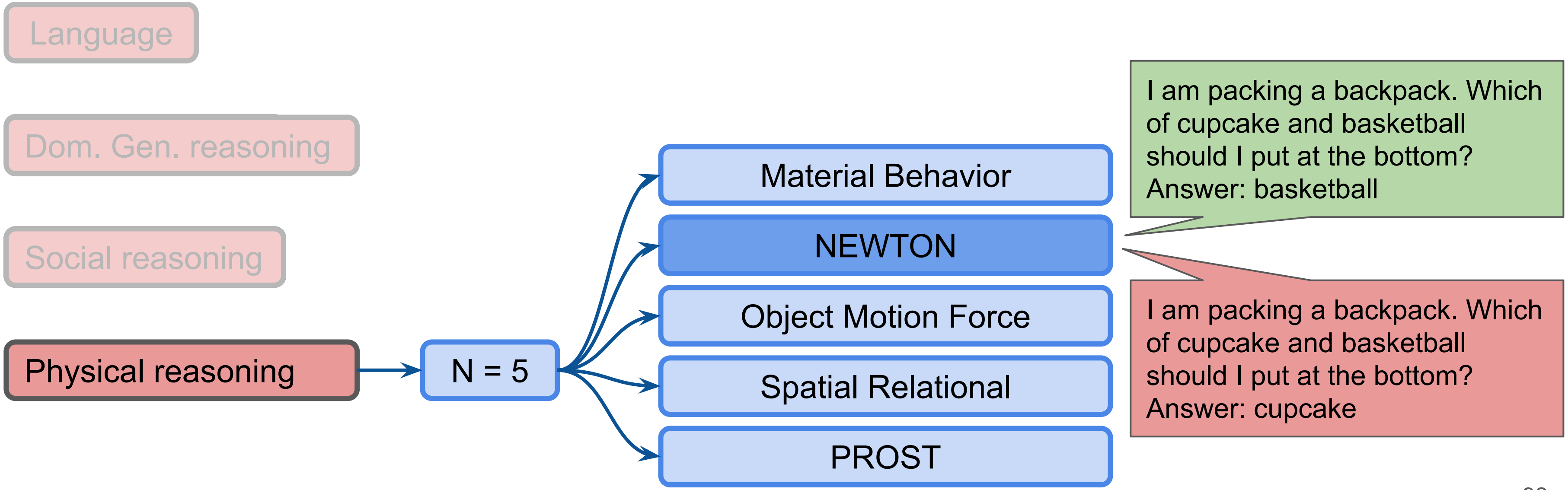
Modularity of reasoning systems in LLMs

We localized components in LLMs supporting **linguistic processing** and different *kinds* of **reasoning** across a total of N = 42 tasks:



Modularity of reasoning systems in LLMs

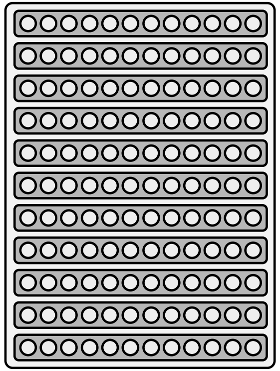
We localized components in LLMs supporting **linguistic processing** and different *kinds* of **reasoning** across a total of N = 42 tasks:



Models

We tested 6 LLMs of intermediate-to-large size (24–123B):

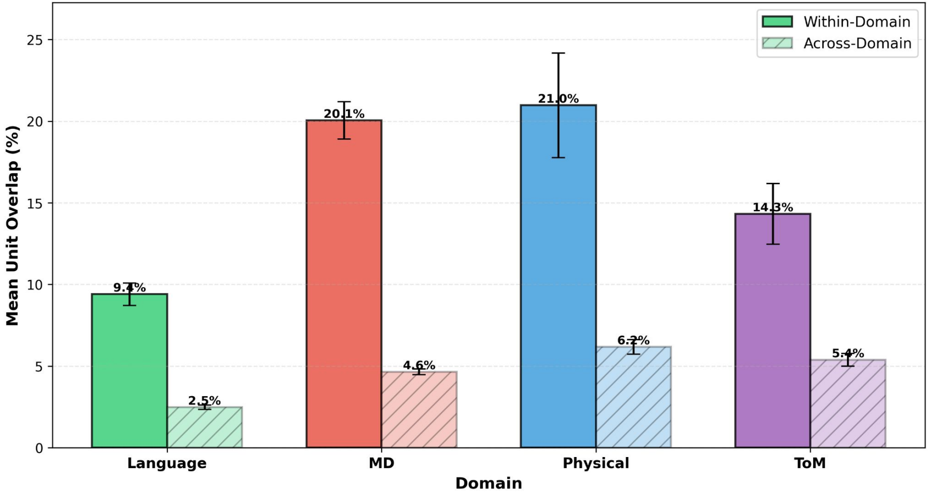
- ❏ Qwen 2.5 32B Instruct
- ❏ Qwen 2.5 72B Instruct
- ❏ Llama 3.1 70B Instruct
- ❏ Mistral 24B Instruct
- ❏ Mistral 123B Instruct
- ❏ Olmo2 32B Instruct



We only kept models that could accurately solve the problems in our meta-dataset (accuracy > 0.8 in 95% of the tasks)

Results

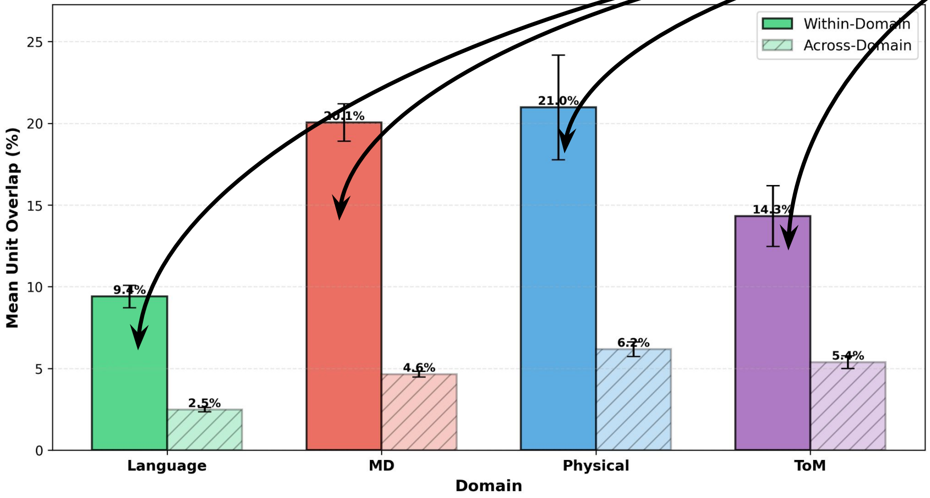
Across the various cognitive domains, more task-selective units (top 1%) are shared *within* a given domain than *across* domains.



Results

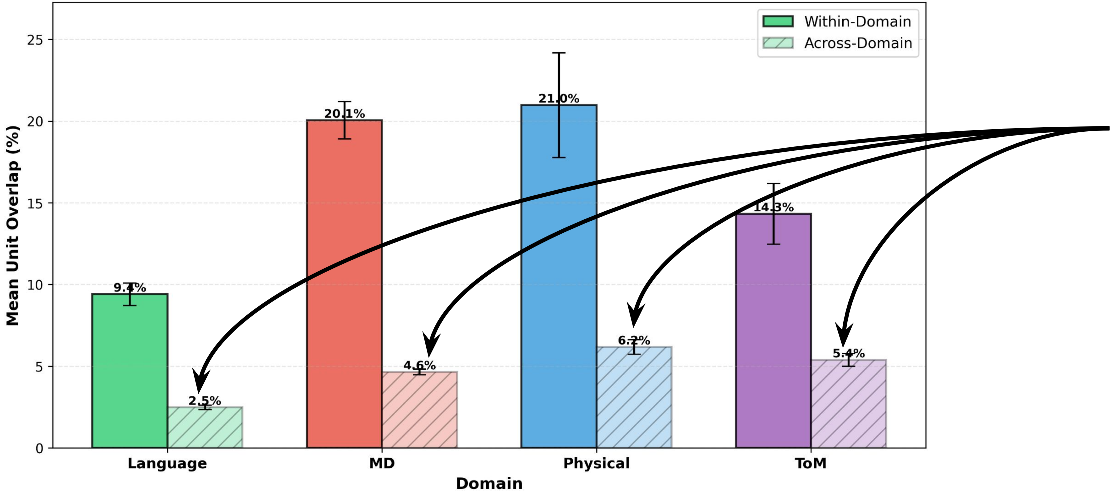
Across the various cognitive domains, more task-selective units (top 1%) are shared *within* a given domain than *across* domains.

High overlap of resources within domains



Results

Across the various cognitive domains, more task-selective units (top 1%) are shared *within* a given domain than *across* domains.



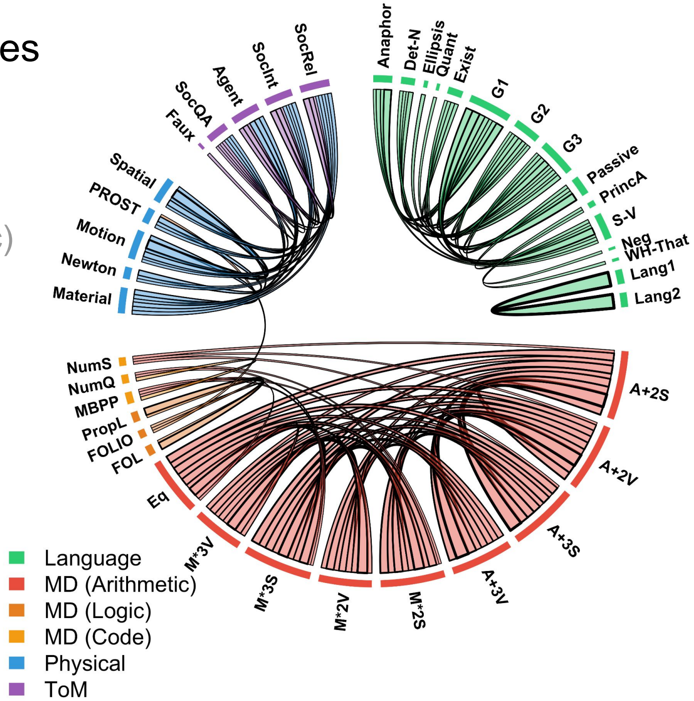
High segregation across domains

Results

Across the various cognitive domains, more task-selective units (top 1%) are shared *within* a given domain than *across* domains.

High overlap of resources within domains

Even for diverse tasks that load onto the same brain networks (math, code, logic)



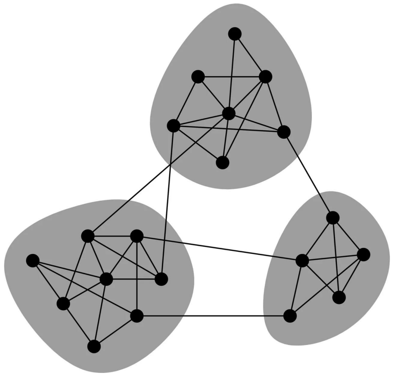
High segregation across domains

Discussion

The internal organization of reasoning systems in LLMs mirrors the **modular** organization of the human mind.

LLMs are not subject to the same **constraints** as the brain (e.g., cost for long connections).

→ Segregation of information may come from general principles of efficient computation



LLMs and reasoning models offer a unified account of the behavioral correlates of reasoning and the internal organization of reasoning systems in humans.

Thank you!





YANG TAN COLLECTIVE

McGOVERN INSTITUTE

brain+cognitive sciences

Thank you!

CoT ≠ language

Reasoning models can be trained with CoT directly in latent space Hao et al., 2024

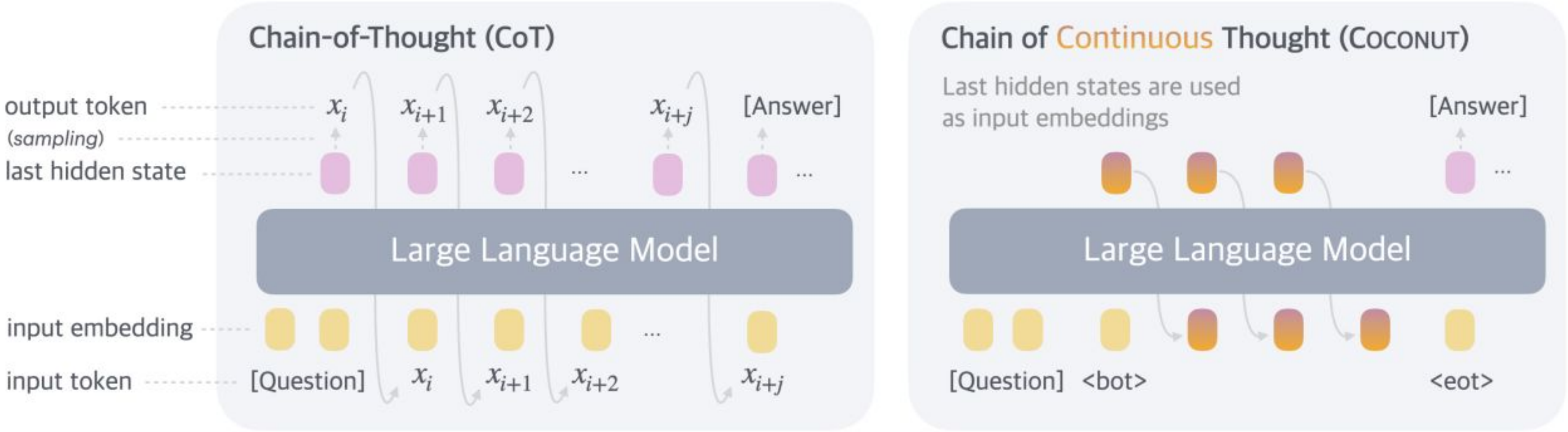


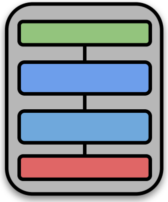
Figure from Hao et al., 2024

CoT ≠ language

Some portions of the actual CoT text are not language:

1. Brian is above and to the right of Henry.
2. Brian is above and to the left of Lucas.
3. Derek is above and to the left of Henry.

LUCAS IS ABOVE AND TO THE RIGHT OF DEREK



...
Now, D_x < H_x and H_x < B_x? From B_x > H_x, so H_x < B_x.
D_x < H_x < B_x < L_x? Let's see.
D_x < H_x (from prem 3)
H_x < B_x (from prem 1, since B_x > H_x)
B_x < L_x (from prem 2, since L_x > B_x)
So D_x < H_x < B_x < L_x
Therefore, D_x < L_x, so L_x > D_x, meaning Lucas is to the right of Derek.

CoT ≠ language

Even though LRMs use language to perform chain-of-thought reasoning, the underlying computations are likely non-linguistic, as in humans Fedorenko et al., 2024

CoT text often misrepresents the true internal processes Barez et al., 2025

- Prompt injection (e.g., "the answer is C") Anthropic Team, 2025
  LRMs never admitted the hint's influence, even though they would often pick a different answer without it
- Post-hoc rationalization of order effects Turpin et al., 2023
- Performance can improve through *filler tokens* (" ", "...") Pfau et al., 2024
- Models trained on random or corrupted traces performed comparatively to those trained on correct reasoning paths Stechly et al., 2025
- Correct solution despite errors in CoT Lanham et al., 2023; Arcuschin et al., 2025; Stechly et al., 2025

The verbal content of the CoT is at best a "lossy projection" of a model's internal computation Dutta et al., 2024

What drives the alignment

Problem length

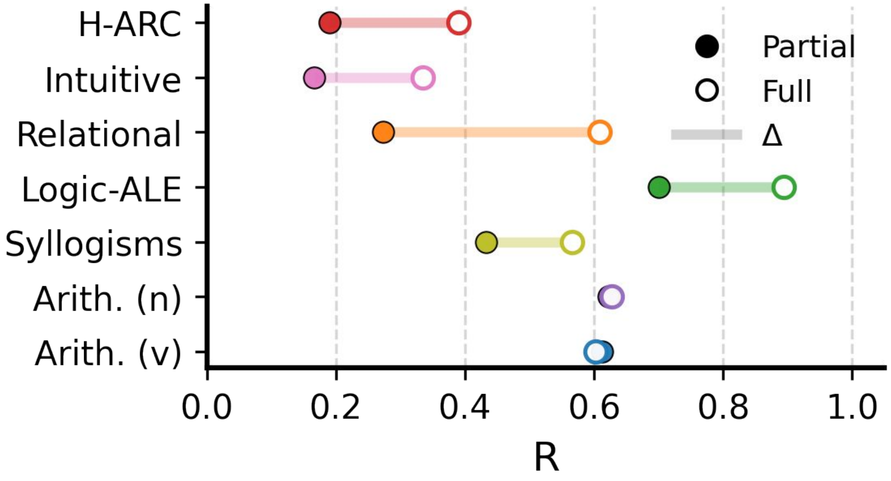
Problem length (number of tokens in the prompt) was also correlated with human RTs.

What drives the alignment

Problem length

Problem length (number of tokens in the prompt) was also correlated with human RTs.

Controlling for it via partial correlation (residualizing both RTs and reasoning length) still yielded significant effects (partial $\bar{r} = 0.43$, all $p < 0.05$).

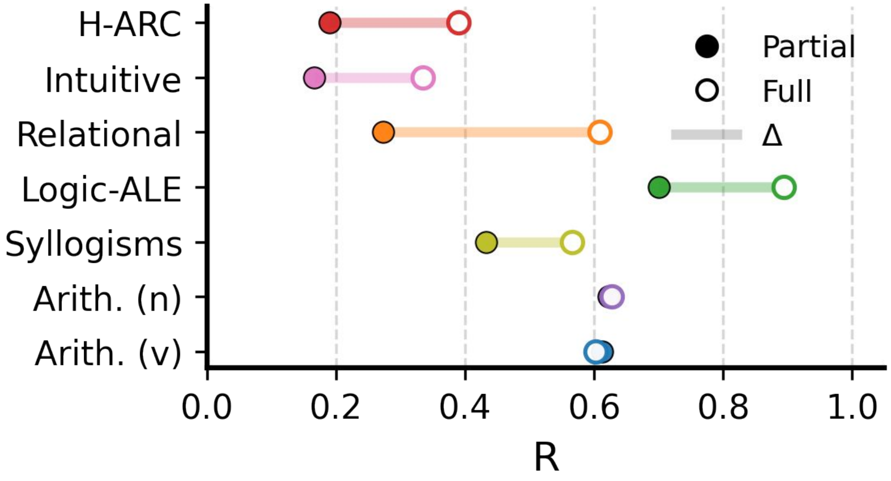


What drives the alignment

Problem length

Problem length (number of tokens in the prompt) was also correlated with human RTs.

Controlling for it via partial correlation (residualizing both RTs and reasoning length) still yielded significant effects (partial $\bar{r} = 0.43$, all $p < 0.05$).



GRPO

$$\mathcal{J}_{GRPO}(\theta) = \mathbb{E}[q \sim P(Q), \{o_i\}_{i=1}^{G} \sim \pi_{\theta_{old}}(O|q)]$$

$$\frac{1}{G}\sum_{i=1}^{G}\left(\min\left(\frac{\pi_\theta(o_i|q)}{\pi_{\theta_{old}}(o_i|q)}A_i, \text{clip}\left(\frac{\pi_\theta(o_i|q)}{\pi_{\theta_{old}}(o_i|q)}, 1-\varepsilon, 1+\varepsilon\right)A_i\right) - \beta\mathbb{D}_{KL}\left(\pi_\theta||\pi_{ref}\right)\right)$$

$$A_i = \frac{r_i - \text{mean}(\{r_1, r_2, \cdots, r_G\})}{std(\{r_1, r_2, \cdots, r_G\})}$$