STaR: Self-Taught Reasoner

Bootstrapping Reasoning With Reasoning

Aniket Tiwari(MDS202308), Kironmoy Roy(MDS202332)

Instructor: Pranabendu Misra(CMI), Dinesh Kirthivasan(Kantar)

Chennai Mathematical Institute

19 May 2025

Outline

The Reasoning Challenge

Current Approaches:

- Manual rationales
  (Expensive, unscalable)

The Reasoning Challenge

Current Approaches:

- Manual rationales
  (Expensive, unscalable)
- Few-shot CoT
  (Limited accuracy)

The Reasoning Challenge

Current Approaches:

- Manual rationales
  (Expensive, unscalable)
- Few-shot CoT
  (Limited accuracy)
- Direct fine-tuning
  (No reasoning)

The Reasoning Challenge

Current Approaches:

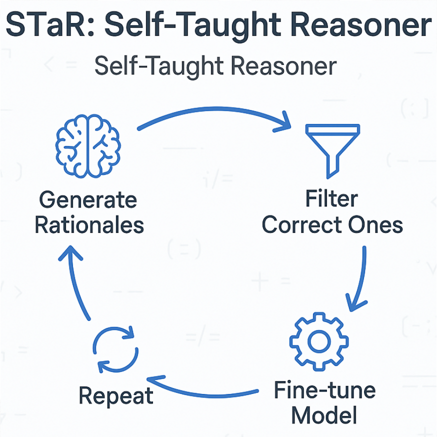
- Manual rationales (Expensive, unscalable)
- Few-shot CoT (Limited accuracy)
- Direct fine-tuning (No reasoning)

Research Question

Can models teach themselves to reason better?

Understanding the STaR Algorithm

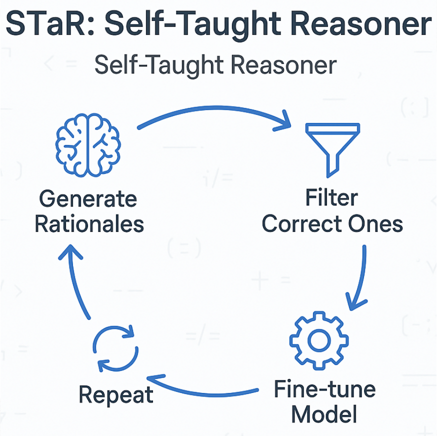
The STaR algorithm introduces a loop-based mechanism that allows a language model to generate and refine its reasoning capabilities over time through rationale generation and iterative learning.



Understanding the STaR Algorithm

The STaR algorithm introduces a loop-based mechanism that allows a language model to generate and refine its reasoning capabilities over time through rationale generation and iterative learning.

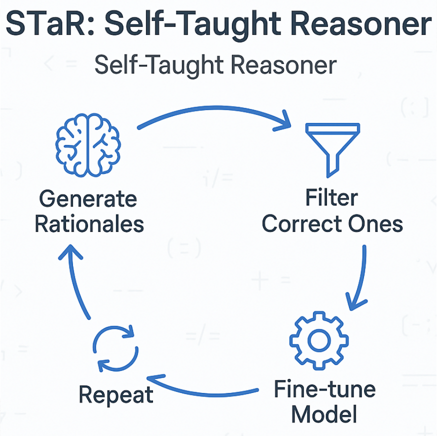
- Creates step-by-step explanations (rationales)



Understanding the STaR Algorithm

The STaR algorithm introduces a loop-based mechanism that allows a language model to generate and refine its reasoning capabilities over time through rationale generation and iterative learning.

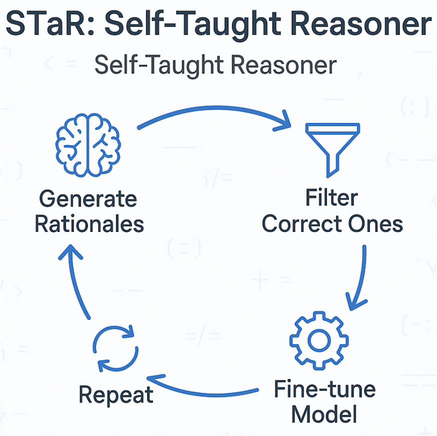
- Creates step-by-step explanations (rationales)
- Continuously improves by learning from generated data



Understanding the STaR Algorithm

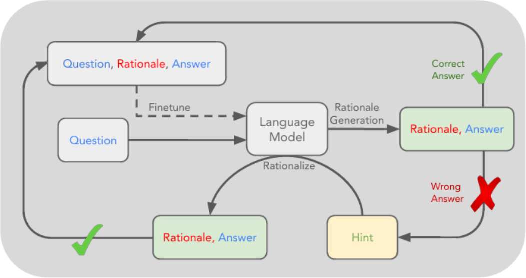
The STaR algorithm introduces a loop-based mechanism that allows a language model to generate and refine its reasoning capabilities over time through rationale generation and iterative learning.

- Creates step-by-step explanations (rationales)
- Continuously improves by learning from generated data
- Combines rationale generation with rationalization



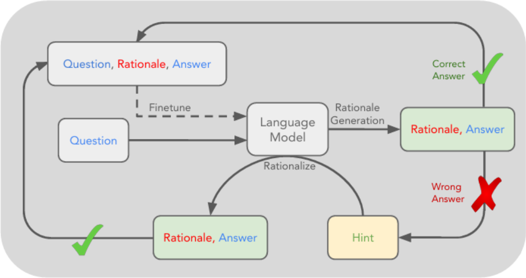
The STaR Process

1. Starting Point: Begins with a Question



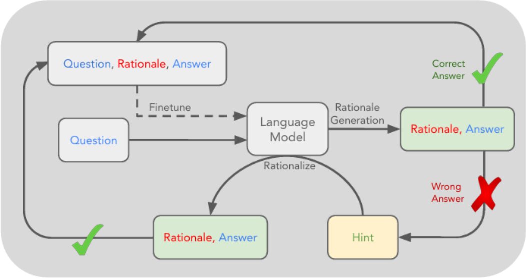
The STaR Process

1. Starting Point: Begins with a Question
2. Rationale Generation: Model generates reasoning + answer
   - Correct: Added to training data
   - Incorrect: Given correct answer as hint



The STaR Process

1. Starting Point: Begins with a Question
2. Rationale Generation: Model generates reasoning + answer
   - Correct: Added to training data
   - Incorrect: Given correct answer as hint
3. Feedback Loop: Rationales used to fine-tune model



Example: Correct Answer

Question: "What is the best way to carry a small dog?"
Choices: (a) Swimming pool (b) Basket (c) Dog show (d) Backyard (e) Own home

Example: Correct Answer

Question: "What is the best way to carry a small dog?"
Choices: (a) Swimming pool (b) Basket (c) Dog show (d) Backyard (e) Own home

Model Output (Correct)

Rationale: "Baskets are portable and designed to hold items. Other options are not for carrying."
Answer: (b) Basket

Example: Correct Answer

Question: "What is the best way to carry a small dog?"
Choices: (a) Swimming pool (b) Basket (c) Dog show (d) Backyard (e) Own home

Model Output (Correct)

Rationale: "Baskets are portable and designed to hold items. Other options are not for carrying."
Answer: (b) Basket

- Correct answer/rationale added to training data

Example: Correct Answer

Question: "What is the best way to carry a small dog?"
Choices: (a) Swimming pool (b) Basket (c) Dog show (d) Backyard (e) Own home

Model Output (Correct)

Rationale: "Baskets are portable and designed to hold items. Other options are not for carrying."
Answer: (b) Basket

- Correct answer/rationale added to training data
- Reinforces valid reasoning paths

Example: Incorrect Answer Handling

Initial Incorrect Output

Rationale: "Dog show has space for movement."
Answer: (c) Dog show

Example: Incorrect Answer Handling

Initial Incorrect Output

Rationale: "Dog show has space for movement."
Answer: (c) Dog show

Correction Process

- Provide hint: Correct answer is (b) Basket

Example: Incorrect Answer Handling

Initial Incorrect Output

Rationale: "Dog show has space for movement."
Answer: (c) Dog show

Correction Process

- Provide hint: Correct answer is (b) Basket
- Model generates new rationale supporting basket

Example: Incorrect Answer Handling

Initial Incorrect Output

Rationale: “Dog show has space for movement.”
Answer: (c) Dog show

Correction Process

- Provide hint: Correct answer is (b) Basket
- Model generates new rationale supporting basket
- New rationale added to training data

Example: Incorrect Answer Handling

Initial Incorrect Output

Rationale: “Dog show has space for movement.”
Answer: (c) Dog show

Correction Process

- Provide hint: Correct answer is (b) Basket
- Model generates new rationale supporting basket
- New rationale added to training data
- Model fine-tuned to avoid similar mistakes

STaR without Rationalization

Algorithm 1 Rationale Generation Bootstrapping (STaR without rationalization)

Input $\boldsymbol{M}$: a pretrained LLM; dataset $\boldsymbol{\mathcal{D}} = \{(x_i, y_i)\}_{i=1}^{D}$ (w/ few-shot prompts)

1: $\boldsymbol{M_0} \leftarrow \boldsymbol{M}$ # Copy the original model
2: **for** n in $1 \ldots N$ **do** # Outer loop
3: $\quad (\hat{r}_i, \hat{y}_i) \leftarrow M_{n-1}(x_i) \quad \forall i \in [1, D]$ # Perform rationale generation
4: $\quad \mathcal{D}_n \leftarrow \{(x_i, \hat{r}_i, y_i) \mid i \in [1, D] \wedge \hat{y}_i = y_i\}$ # Filter rationales using ground truth answers
5: $\quad M_n \leftarrow \text{train}(M, \mathcal{D}_n)$ # Finetune the original model on the correct solutions - inner loop

STaR Algorithm

Algorithm 2 STaR

Input M: a pretrained LLM; dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{D}$ (w/ few-shot prompts)

1: $M_0 \leftarrow M$ # Copy the original model
2: **for** n in $1 \ldots N$ **do** # Outer loop
3: $(\hat{r}_i, \hat{y}_i) \leftarrow M_{n-1}(x_i) \quad \forall i \in [1, D]$ # Perform rationale generation
4: $(\hat{r}_i^{\text{rat}}, \hat{y}_i^{\text{rat}}) \leftarrow M_{n-1}(\texttt{add\_hint}(x_i, y_i)) \quad \forall i \in [1, D]$ # Perform rationalization
5: $\mathcal{D}_n \leftarrow \{(x_i, \hat{r}_i, y_i) \mid i \in [1, D] \wedge \hat{y}_i = y_i\}$ # Filter rationales using ground truth
6: $\mathcal{D}_n^{\text{rat}} \leftarrow \{(x_i, \hat{r}_i^{\text{rat}}, y_i) \mid i \in [1, D] \wedge \hat{y}_i \neq y_i \wedge \hat{y}_i^{\text{rat}} = y_i\}$ # Filter rationalized rationales
7: $M_n \leftarrow \text{train}(M, \mathcal{D}_n \cup \mathcal{D}_n^{\text{rat}})$ # Finetune on correct solutions

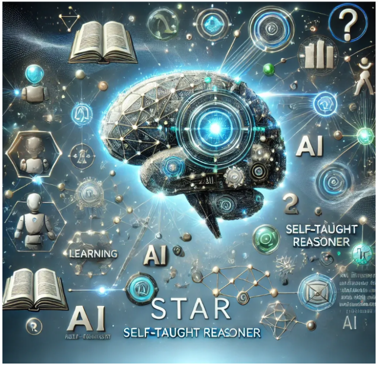
Technical Implementation

Mathematical Formulation

Treat as latent variable model:

$$p_M(y|x) = \sum_r p(r|x)p(y|x, r)$$

$$\nabla J = \sum_i \mathbb{E}_{r,y}[\mathbb{I}(y_i = \hat{y}_i) \cdot \nabla \log p_M(\hat{y}_i, \hat{r}_i|x_i)]$$



Technical Implementation

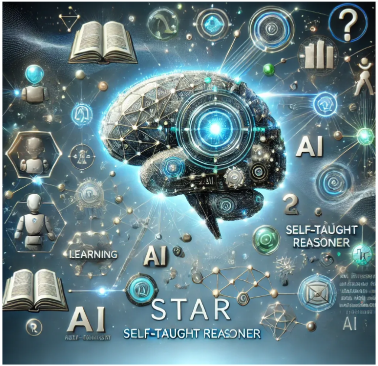
Mathematical Formulation

Treat as latent variable model:

$$p_M(y|x) = \sum_r p(r|x)p(y|x,r)$$

$$\nabla J = \sum_i \mathbb{E}_{r,y}[\mathbb{I}(y_i = \hat{y}_i) \cdot \nabla \log p_M(\hat{y}_i, \hat{r}_i|x_i)]$$

- Base model: GPT-J (6B parameters)
- Batch size: 8 sequences $\times$ 1024 tokens
- Learning rate: 1e-6 (Adam optimizer)
- TPU-v3 hardware



CommonsenseQA Results

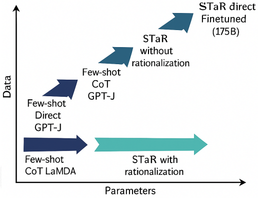
Method	CQA Dev Set Accuracy (%)	Train Data Used (%)
GPT-3 Direct Finetuned(175B)	73.0	100
Few-shot Direct GPT-J	20.9	$\sim$0
Few-shot CoT GPT-J	36.6	$\sim$0
Few-shot CoT LaMDA 137B	55.6	$\sim$0
GPT-J Direct Finetuned	60.0	100
STaR without rationalization	68.8	69.7
STaR with rationalization	**72.5**	86.7

CommonsenseQA Results

Method	CQA Dev Set Accuracy (%)	Train Data Used (%)
GPT-3 Direct Finetuned(175B)	73.0	100
Few-shot Direct GPT-J	20.9	$\sim$0
Few-shot CoT GPT-J	36.6	$\sim$0
Few-shot CoT LaMDA 137B	55.6	$\sim$0
GPT-J Direct Finetuned	60.0	100
STaR without rationalization	68.8	69.7
STaR with rationalization	**72.5**	86.7

Key Observations

- STaR achieves 72.5% accuracy vs. GPT-3's 73.0%



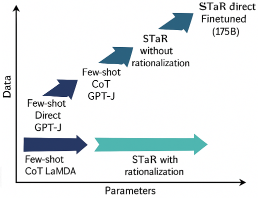
STaR matches GPT-3 accuracy with less data and fewer parameters, outperforming larger models like LaMDA.

CommonsenseQA Results

Method	CQA Dev Set Accuracy (%)	Train Data Used (%)
GPT-3 Direct Finetuned(175B)	73.0	100
Few-shot Direct GPT-J	20.9	$\sim$0
Few-shot CoT GPT-J	36.6	$\sim$0
Few-shot CoT LaMDA 137B	55.6	$\sim$0
GPT-J Direct Finetuned	60.0	100
STaR without rationalization	68.8	69.7
STaR with rationalization	**72.5**	86.7

Key Observations

- STaR achieves 72.5% accuracy vs. GPT-3's 73.0%
- Uses only 86.7% training data (78.2% generation + 8.5% rationalization)



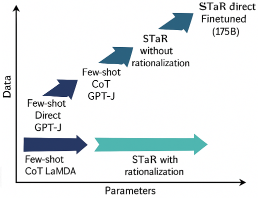
STaR matches GPT-3 accuracy with less data and fewer parameters, outperforming larger models like LaMDA.

CommonsenseQA Results

Method	CQA Dev Set Accuracy (%)	Train Data Used (%)
GPT-3 Direct Finetuned(175B)	73.0	100
Few-shot Direct GPT-J	20.9	∼0
Few-shot CoT GPT-J	36.6	∼0
Few-shot CoT LaMDA 137B	55.6	∼0
GPT-J Direct Finetuned	60.0	100
STaR without rationalization	68.8	69.7
STaR with rationalization	**72.5**	86.7

Key Observations

- STaR achieves 72.5% accuracy vs. GPT-3's 73.0%
- Uses only 86.7% training data (78.2% generation + 8.5% rationalization)
- Outperforms LaMDA (137B params) with just 6B params



STaR matches GPT-3 accuracy with less data and fewer parameters, outperforming larger models like LaMDA.

GSM8K Results (Math Word Problems)

Method	Test Accuracy (%)	Train Data Used (%)
Few-shot Direct GPT-J	3.0	$\sim$0
Few-shot CoT GPT-J	3.1	$\sim$0
GPT-J Direct Finetuned	5.8	100
STaR without rationalization	10.1	25.0
STaR with rationalization	**10.7**	28.7

GSM8K Results (Math Word Problems)

Method	**Test Accuracy (%)**	**Train Data Used (%)**
Few-shot Direct GPT-J	3.0	$\sim$0
Few-shot CoT GPT-J	3.1	$\sim$0
GPT-J Direct Finetuned	5.8	100
STaR without rationalization	10.1	25.0
STaR with rationalization	**10.7**	28.7

- STaR achieves 2-3$\times$ improvement over baselines

GSM8K Results (Math Word Problems)

Method	Test Accuracy (%)	Train Data Used (%)
Few-shot Direct GPT-J	3.0	$\sim$0
Few-shot CoT GPT-J	3.1	$\sim$0
GPT-J Direct Finetuned	5.8	100
STaR without rationalization	10.1	25.0
STaR with rationalization	**10.7**	28.7

- STaR achieves 2-3$\times$ improvement over baselines
- Uses only 25–28% of training data

GSM8K Results (Math Word Problems)

Method	Test Accuracy (%)	Train Data Used (%)
Few-shot Direct GPT-J	3.0	$\sim$0
Few-shot CoT GPT-J	3.1	$\sim$0
GPT-J Direct Finetuned	5.8	100
STaR without rationalization	10.1	25.0
STaR with rationalization	**10.7**	28.7

- STaR achieves 2-3$\times$ improvement over baselines
- Uses only 25–28% of training data
- Rationalization provides modest gains in this domain

Simpler Solutions by the Model: GSM8K Solutions

Q: A van is delivering 180 bottles of drinks to a neighborhood, and each bottle contains either cider or beer or a mixture of two. Out of the 180 bottles, 40 contain only cider, 80 contain only beer, and the rest are a mixture of the two drinks. If the delivery man gives half the number of each bottle of drink to the first house, how many bottles does the first house get?

Human (Ground truth from GSM8K dataset):

The total number of bottles containing cider or beer is 40+80 = <<40+80=120>>120
If the number of bottles in the van was 180, then 180-120 =<<180-120=60>>60 contains a mixture of the two drinks.
At the first house, the delivery man dropped 1/2*40 = <<1/2*40=20>>20 bottles.
There were also 1/2*80 = <<1/2*80=40>>40 bottles of beer that the delivery man gave to the first house.
Together, the first house received 40+20 = <<40+20=60>>60 bottles of cider and beer
The total number of bottles that had a mixture of the two drinks that were dropped at the first house is 1/2*60 = <<1/2*60=30>>30
The first house got 60+30 = <<60+30=90>>90 bottles of all types of drinks.

90

STaR (on GPT-J):

The delivery man gives half the number of each bottle of drink to the first house.
So he gives 180/2 = <<180/2=90>>90 bottles of drink to the first house.

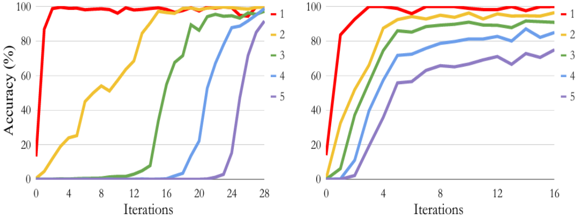
90

Fig: Example problem in the training set where STaR derives significantly simpler solution than the ground truth.

Why STaR is a Game-Changer?

Core Innovations

- **Human-like learning**:
  Generate → Verify → Improve cycle



(a) Without rationalization

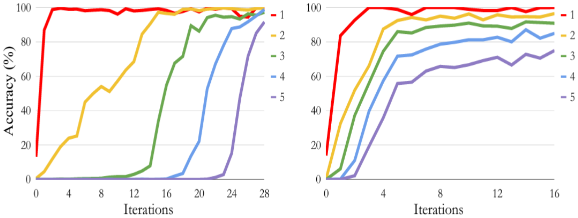
(b) With rationalization

Why STaR is a Game-Changer?

Core Innovations

- Human-like learning:

  Generate → Verify → Improve cycle
  Learns from both correct/incorrect outputs



(a) Without rationalization (b) With rationalization

Why STaR is a Game-Changer?

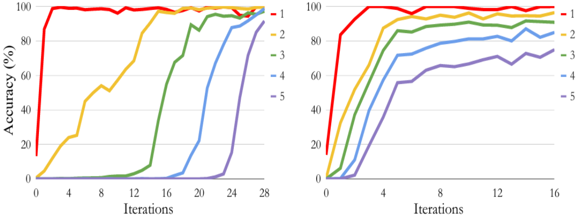
Core Innovations

- Human-like learning:

  Generate $\rightarrow$ Verify $\rightarrow$ Improve cycle
  Learns from both correct/incorrect outputs

- Mathematical edge:

  Explores $p(r|x, y)$ space not $p(r|x)$



(a) Without rationalization (b) With rationalization

Why STaR is a Game-Changer?

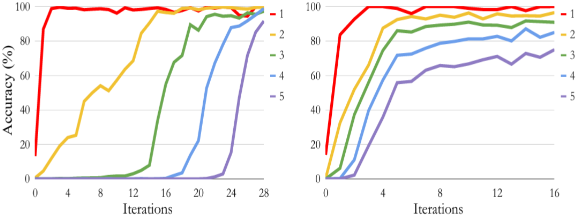
Core Innovations

- Human-like learning:

  Generate → Verify → Improve cycle
  Learns from both correct/incorrect outputs

- Mathematical edge:

  Explores $p(r|x, y)$ space not $p(r|x)$
  Stronger training signals



(a) Without rationalization (b) With rationalization

Why STaR is a Game-Changer?

Core Innovations

- **Human-like learning**:

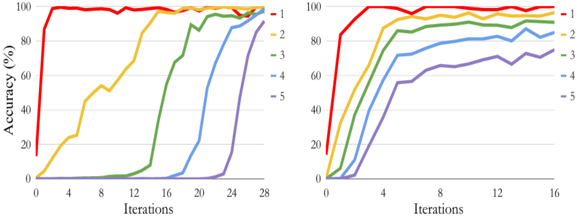
  Generate $\rightarrow$ Verify $\rightarrow$ Improve cycle
  Learns from both correct/incorrect outputs

- **Mathematical edge**:

  Explores $p(r|x, y)$ space not $p(r|x)$
  Stronger training signals

- **AI Development**

  Faster training of models

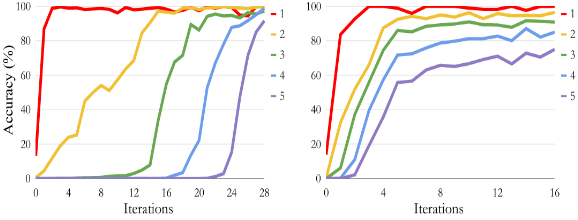


(a) Without rationalization (b) With rationalization

Why STaR is a Game-Changer?

Core Innovations

- **Human-like learning**:
  - Generate → Verify → Improve cycle
  - Learns from both correct/incorrect outputs
- **Mathematical edge**:
  - Explores $p(r|x, y)$ space not $p(r|x)$
  - Stronger training signals
- **AI Development**
  - Faster training of models
  - Less human-labeled data needed

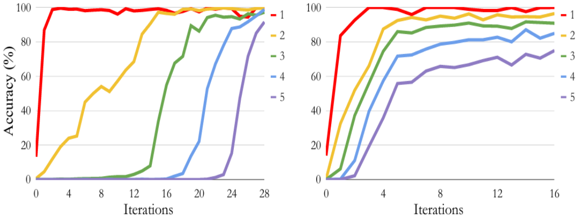


(a) Without rationalization
(b) With rationalization

Why STaR is a Game-Changer?

Core Innovations

- Human-like learning:
  - Generate → Verify → Improve cycle
  - Learns from both correct/incorrect outputs
- Mathematical edge:
  - Explores $p(r|x, y)$ space not $p(r|x)$
  - Stronger training signals
- AI Development
  - Faster training of models
  - Less human-labeled data needed



(a) Without rationalization (b) With rationalization

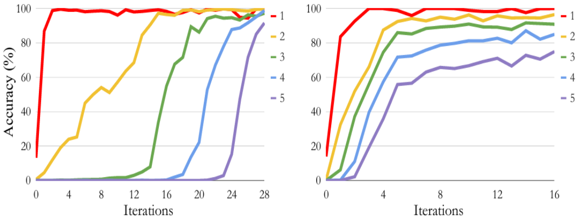
Results

- +35% accuracy over few-shot

Why STaR is a Game-Changer?

Core Innovations

- Human-like learning:
  - Generate → Verify → Improve cycle
  - Learns from both correct/incorrect outputs
- Mathematical edge:
  - Explores $p(r|x, y)$ space not $p(r|x)$
  - Stronger training signals
- AI Development
  - Faster training of models
  - Less human-labeled data needed



(a) Without rationalization (b) With rationalization

Results

- +35% accuracy over few-shot
- Matches models 30x larger
- Reverse-engineers solutions
- Prevents learning plateaus

Human-Evaluated Test Prompts

Comparing Few-Shot vs. STaR vs. Human Rationales

Evaluation Methodology

- 50 questions correctly answered by both models
- 20 crowdworkers ranked rationales (1=best, 3=worst)
- Examples shown with sources shuffled

Human-Evaluated Test Prompts

Comparing Few-Shot vs. STaR vs. Human Rationales

Evaluation Methodology

- 50 questions correctly answered by both models
- 20 crowdworkers ranked rationales (1=best, 3=worst)
- Examples shown with sources shuffled

Sample Question

Q: What should you do to become a good writer?
Choices: (a) word sentence (b) own animal (c) read newspaper (d) catch cold (e) study literature

Human-Evaluated Test Prompts

Comparing Few-Shot vs. STaR vs. Human Rationales

Evaluation Methodology

- 50 questions correctly answered by both models
- 20 crowdworkers ranked rationales (1=best, 3=worst)
- Examples shown with sources shuffled

Sample Question

Q: What should you do to become a good writer?
Choices: (a) word sentence (b) own animal (c) read newspaper (d) catch cold (e) study literature

1. (STaR) "Literature develops writing skills..."
2. (Few-Shot) "The answer must help writing..."
3. (Human) "Studying literature gives writing skills..."

Human-Evaluated Test Prompts

Comparing Few-Shot vs. STaR vs. Human Rationales

Evaluation Methodology

- 50 questions correctly answered by both models
- 20 crowdworkers ranked rationales (1=best, 3=worst)
- Examples shown with sources shuffled

Sample Question

Q: What should you do to become a good writer?
Choices: (a) word sentence (b) own animal (c) read newspaper (d) catch cold (e) study literature

1. (STaR) "Literature develops writing skills..."
2. (Few-Shot) "The answer must help writing..."
3. (Human) "studying literature gives writing skills..."

STaR rationales were preferred 30% over few-shot and 74% over human with $p < 0.01$

Key Limitations and Challenges

Major Limitations	Ethical Challenges
■ Requires human examples to start ■ Struggles with math reasoning ■ Slow processing speed ■ Needs careful tuning	■ Accountability for errors unclear ■ Potential for hidden biases ■ Autonomous learning risks

Conclusion & Broader Impact

Key Contributions

- First framework for **self-improving reasoning** via rationalization

Conclusion & Broader Impact

Key Contributions

- First framework for **self-improving reasoning** via rationalization
- Achieves performance of models **30× larger** with less human data

Conclusion & Broader Impact

Key Contributions

- First framework for **self-improving reasoning** via rationalization
- Achieves performance of models **30× larger** with less human data
- Creates more **transparent** & **explainable** AI decisions

Conclusion & Broader Impact

Key Contributions

- First framework for **self-improving reasoning** via rationalization
- Achieves performance of models **30× larger** with less human data
- Creates more **transparent** & **explainable** AI decisions

Transformative Potential

- **Applications:** More explainable healthcare/legal/education decisions

Conclusion & Broader Impact

Key Contributions

- First framework for **self-improving reasoning** via rationalization
- Achieves performance of models **30× larger** with less human data
- Creates more **transparent** & **explainable** AI decisions

Transformative Potential

- **Applications:** More explainable healthcare/legal/education decisions
- **Ethics:** Built-in rationale generation enables bias detection

Conclusion & Broader Impact

Key Contributions

- First framework for **self-improving reasoning** via rationalization
- Achieves performance of models **30× larger** with less human data
- Creates more **transparent** & **explainable** AI decisions

Transformative Potential

- **Applications:** More explainable healthcare/legal/education decisions
- **Ethics:** Built-in rationale generation enables bias detection
- **Future Research:** Inspires new RL methods (STaR-RL), ReSTaR, Q-STaR etc.

Conclusion & Broader Impact

Key Contributions

- First framework for **self-improving reasoning** via rationalization
- Achieves performance of models **30× larger** with less human data
- Creates more **transparent** & **explainable** AI decisions

Transformative Potential

- **Applications:** More explainable healthcare/legal/education decisions
- **Ethics:** Built-in rationale generation enables bias detection
- **Future Research:** Inspires new RL methods (STaR-RL), ReSTaR, Q-STaR etc.

Conclusion & Broader Impact

Key Contributions

- First framework for **self-improving reasoning** via rationalization
- Achieves performance of models **30× larger** with less human data
- Creates more **transparent** & **explainable** AI decisions

Transformative Potential

- **Applications:** More explainable healthcare/legal/education decisions
- **Ethics:** Built-in rationale generation enables bias detection
- **Future Research:** Inspires new RL methods (STaR-RL), ReSTaR, Q-STaR etc.

Future Outlook

General framework for creating more capable, efficient, and trustworthy AI systems

End

Thank You!