SPICE 🌶️: Self-Play In Corpus Environments Improves Reasoning

Bo Liu[1,2,*], **Chuanyang Jin**[1,‡], **Seungone Kim**[1,‡], **Weizhe Yuan**[1], **Wenting Zhao**[1], **Ilia Kulikov**[1], **Xian Li**[1], **Sainbayar Sukhbaatar**[1], **Jack Lanchantin**[1,∇], **Jason Weston**[1,∇]

[1]FAIR at Meta, [2]National University of Singapore
*Work done at Meta, ‡Joint second author, ∇Joint last author

Presenters: Aryan Shah and Ruining Yang

What is Self-Play Reinforcement Learning?

Description of Self-Play RL

- A model challenges itself: challenger & reasoner
  - Role of challenger: Creates more challenging questions
  - Role of reasoner: Provides more accurate answers

Why is Self-Play Important?

- Eliminates the need for human-annotated training datasets
- Models can provide more diverse answers, potentially outperforming human experts

Related Work

Absolute Zero

- Absolute Zero set the groundwork by creating a framework of a challenger and reasoner
- Created a curriculum of reasoning tasks and trajectories
- Self-play loop consisted of asking and solving coding tasks and results were verified on math and coding benchmarks

R-Zero

- Used a very similar structure to Absolute Zero but also verified on general reasoning benchmarks
- Self-play loop used math questions and answers rather than code
- Adapted the reward of the challenger to target maximum uncertainty rather than maximum difficulty

Research Gap and Contributions

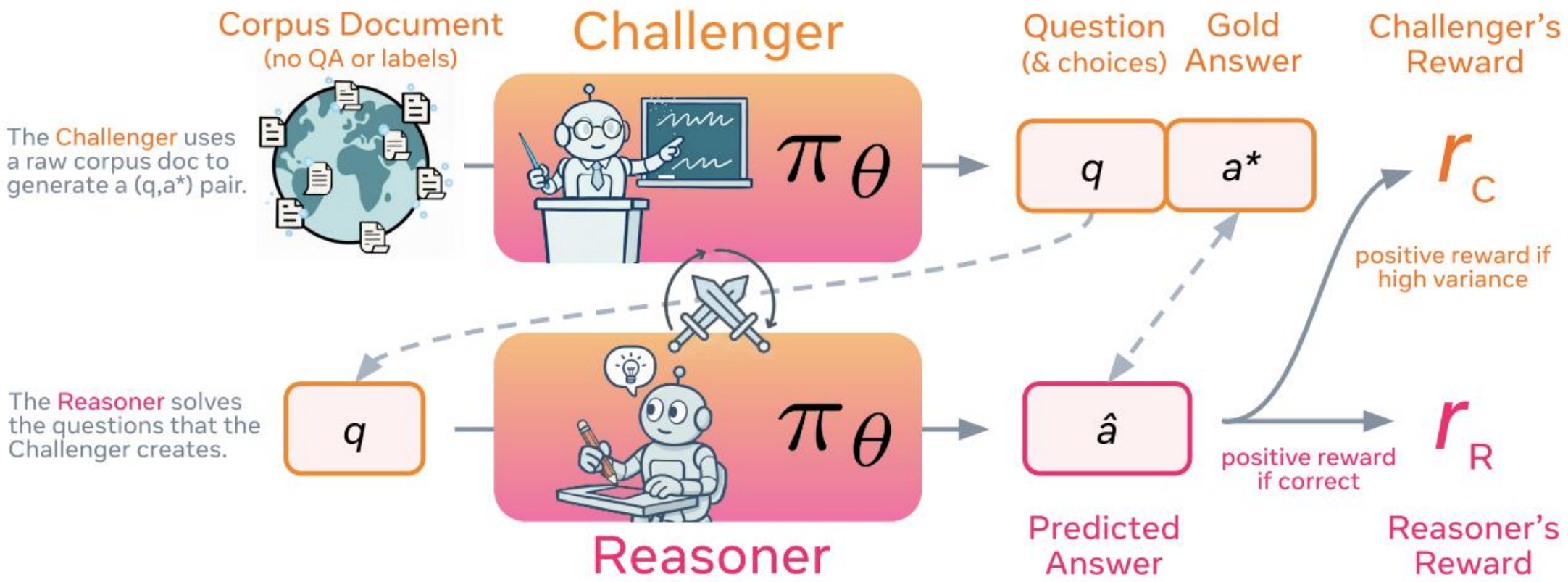
Limitations

- Two previous papers in self-play reinforcement learning (Absolute Zero, R-Zero) suffered from three major issues
  - Hallucination Amplification: Model creates factual errors
  - Information Symmetry: Model isn't challenged enough and has simple patterns
  - Restricted to Verifiable Domains: Only works on math and code domains because ground truth can be verified

Contributions

- A **large document corpus** as an external knowledge source avoids factual errors
- Corpus-grounded self-play create information asymmetries forcing the model to reason
- The performance gains are verified across math and general reasoning domains, generalizable to broad domains

High-Level Overview of the Self-Play Loop



Workflow of the Challenger

- Sample passages from the document corpus
- Create question-answer pairs with answers from the passage
- Improve itself according to a **variance-based curriculum reward signal**

Variance-Based Curriculum Reward. For each valid question (q, a^*), we sample K responses $\hat{a}_i$ from the Reasoner and compute the Challenger's reward using a scaled variance of the responses:

$$r_C(q, a^*) = \begin{cases} \exp\left(-\frac{(\mathrm{Var}(\{l_1,\ldots,l_K\})-0.25)^2}{2\cdot 0.01}\right) & \text{if } q \text{ is valid} \\ \rho & \text{otherwise (penalty)} \end{cases} \quad (1)$$

where $l_i = \mathbb{1}[\hat{a}_i = a^*]$. This Gaussian-shaped reward function is scaled to [0,1], maximizing at 1.0 when variance equals 0.25 (50% pass rate), indicating optimal task difficulty. Tasks that are too easy or too hard receive exponentially lower rewards; as the Reasoner improves, the Challenger is rewarded for increasing task difficulty, creating an automatic curriculum.

Reasoner Workflow and Key Improvements

Reasoner Workflow

- Answer the challenger's question without access to the document corpus
- Improve itself with a binary correctness reward objective

Improvements of the SPICE Algorithm

- Question-answer pairs are corpus grounded so the model cannot hallucinate
- Challenger has access to document corpus while reasoner doesn't, creating challenging asymmetries
- Since ground truth is based on the document corpus, this method works on unverifiable domains

Full SPICE Algorithm

Algorithm 1 SPICE: Self-Play In Corpus Environments

Require: Pretrained LLM π_θ; corpus $\mathcal{D}$; batch size B; group size G; iterations T; penalty ρ

for $t \leftarrow 1$ to T **do**
  Challenger Role: Generate challenging problems with document access
  for $b \leftarrow 1$ to B **do**
    Sample document $d \sim \mathcal{D}$
    Generate multiple attempts: $\{(q_i, a_i^*)\}_{i=1}^{N} \leftarrow \pi_\theta(d, \text{role} = C)$ ▷ MCQ or free-form
    $\mathcal{T}_C \leftarrow$ Subsample G trajectories preserving valid:invalid ratio
    if $q_i \in \mathcal{T}_C$ is valid **then**
      $\{\hat{a}_k\}_{k=1}^{G} \leftarrow \pi_\theta(q_i, \text{role} = R)$ ▷ No document access
      $r_C(q_i, a_i^*) \leftarrow$ Compute reward using Eq. (1) ▷ Gaussian variance reward
    else
      $r_C(q_i, a_i^*) \leftarrow \rho$ ▷ Invalid task penalty
    end if
  end for
  Reasoner Role: Solve generated problems without document access
  Select a random valid task (q, a^*) from Challenger phase
  $\mathcal{T}_R \leftarrow \{\hat{a}_i\}_{i=1}^{G} \sim \pi_\theta(q, \text{role} = R)$ ▷ G responses for training
  for $i \leftarrow 1$ to G **do**
    $r_R(\hat{a}_i, a_i^*) \leftarrow \mathbb{1}[\hat{a}_i = a^*]$ ▷ Binary correctness reward
  end for
  Update Phase: Optimize π_θ with role-specific advantages
  for trajectory i in $\mathcal{T}_C$ **do**
    $\hat{A}_C^i \leftarrow r_C^i - \text{mean}(\{r_C^j : j \in \mathcal{T}_C\})$ ▷ DrGRPO
  end for
  for trajectory i in $\mathcal{T}_R$ **do**
    $\hat{A}_R^i \leftarrow r_R^i - \text{mean}(\{r_R^j : j \in \mathcal{T}_R\})$ ▷ DrGRPO
  end for
  Update π_θ via policy gradient with advantages $\{\hat{A}_C^i\}$ and $\{\hat{A}_R^i\}$
end for
return Trained model π_θ

Experiment Settings

Baseline

- Base model - Pretrained; No post Training
- Strong Challenger - Fixed stronger model as challenger → Challenger Learning Impacts
- R-Zero - Self-play; No corpus grounding → Corpus grounding
- Absolute-Zero - Self-play; Verified programing tasks → Reward design
- SPICE - Self-play; Corpus grounding

Base model

Model Types	Small Scale	Large Scale
General Purpose LLMs	Qwen3-4B-Base	Qwen3-8B-Base
Reasoning-oriented Models	OctoThinker-3B-Hybrid-Base	OctoThinker-8B-Hybrid-Base

Evaluation for Mathematical Reasoning & General Reasoning

- Mathematical reasoning: MATH-500, AIME25, OlympiadBench, GSM8K, etc.
- General reasoning: GPQA-Diamond, MMLU-Pro, SuperGPQA, and BBEH.

Training Setup

- 20,000 high-quality documents used as the training corpus
- mathematical reasoning (Nemotron-CC-Math) + general reasoning (NaturalReasoning)

Configuration	SPICE	Strong Challenger	R-Zero	Absolute Zero
Data Source				
Corpus documents	20,000	20,000	–	–
Question source	Document-grounded	Document-grounded	Self-generated	Self-generated
External grounding	✓	✓	×	Python executor
Training Details				
Challenger training	✓	× (fixed)	✓	✓
Challenger sampling	8	–	4	1
Reasoner training	✓	✓	✓	✓
Reasoner sampling	8	8	5	1
Temperature	1.0	1.0	1.0	1.0
Optimizer	DrGRPO	DrGRPO	GRPO	REINFORCE++
Reward Design				
Challenger reward	Gaussian Variance (max 1.0)	N/A	$1 - 2\|p - 0.5\|$	$1 - p$
Reasoner reward	Binary correctness	Binary correctness	Binary correctness	Binary correctness
Invalid penalty	-0.1	N/A	-1 (Challenger)	-0.5/-1*
Performance				
Training iterations	640	640	5†	640

Insights from Training

Adversarial Training Dynamics

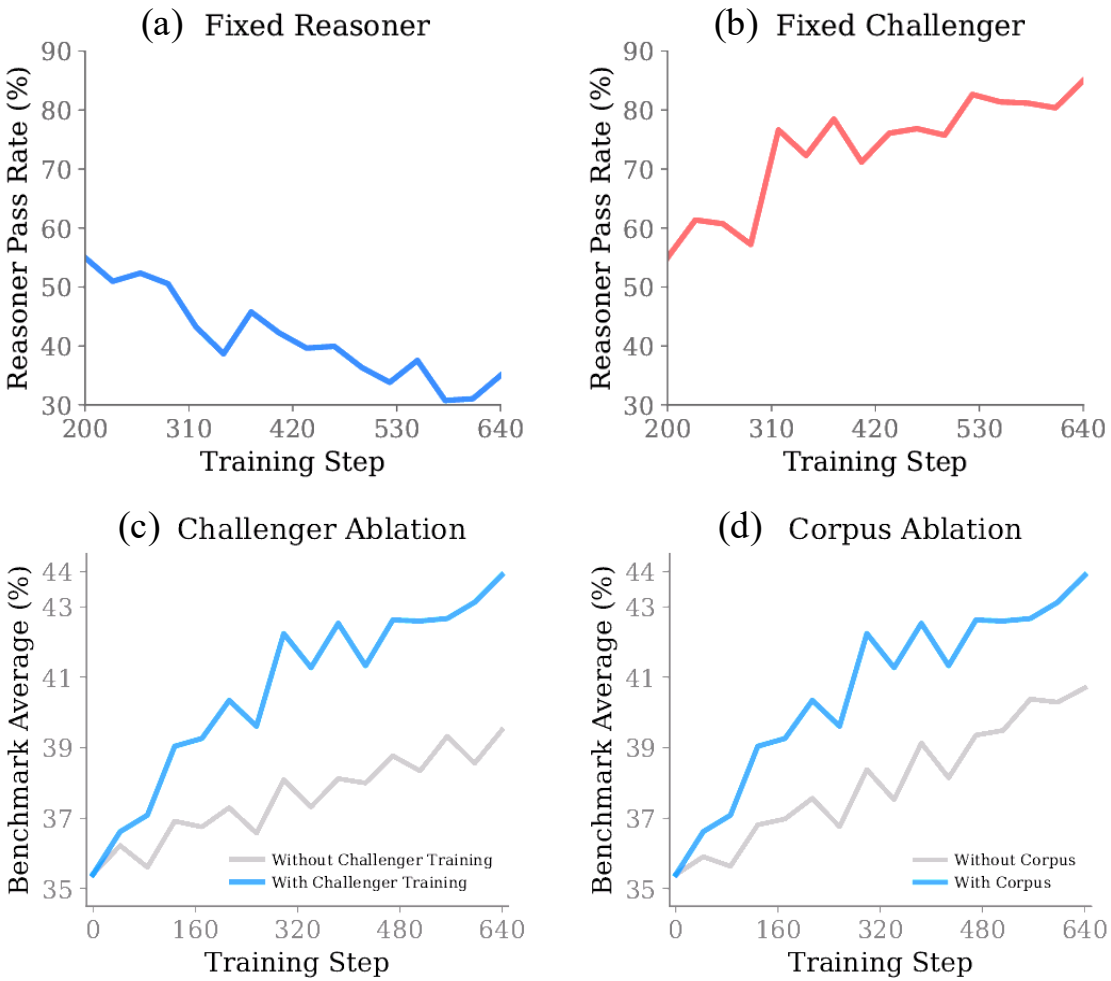
- Reduced pass rate → Challenger improves
- Increased pass rates → Reasoner improves

Challenger Learning Impact

- With challenger training → Higher benchmark average

Corpus Grounding Impact

- With corpus → Higher benchmark average



Insights from Training

Surface-level Question → Complex Multi-step Questions

Early Training (Step 50): Surface-level Question

Document: [Passage about solar system scale] "The diameter of the Sun is 1,391,000 km. The diameter of the Moon is 3,475 km. The distance from Earth to the Sun is 149,600,000 km. For a solar eclipse to occur, the Moon must appear the same angular size as the Sun from Earth. In a scale model, a circle with diameter 2 cm represents Earth. When Earth is scaled to 0.3 mm (like a period), the entire solar system model changes dramatically..."
Generated MCQ: What is the diameter of the Moon?
A) 1,391 km
B) 3,475 km
C) 34,750 km
D) 347,500 km
Answer: B

Late Training (Step 480): Multi-step Reasoning

Document: [Passage about solar system scale] "The diameter of the Sun is 1,391,000 km. The diameter of the Moon is 3,475 km. The distance from Earth to the Sun is 149,600,000 km. For a solar eclipse to occur, the Moon must appear the same angular size as the Sun from Earth. In a scale model, a circle with diameter 2 cm represents Earth. When Earth is scaled to 0.3 mm (like a period), the entire solar system model changes dramatically..."
Generated MCQ: An alien civilization has a moon with diameter 3,475 km that creates perfect solar eclipses. Their star has the same diameter as our Sun. If their moon orbits at 374,000 km (maintaining the same angular size ratio as Earth's eclipse geometry), approximately what would be the distance from their planet to their star in kilometers?
A) 374,000,000 km
B) 149,600,000 km
C) 1,391,000,000 km
D) 74,800,000 km
Answer: B

Insights from Training

Unstructured Reasoning

↓

Structured Multi-step Reasoning

Early Training: Unstructured Reasoning

Question: An alien civilization has a moon with diameter 3,475 km that creates perfect solar eclipses. Their star has the same diameter as our Sun. If their moon orbits at 374,000 km (maintaining the same angular size ratio as Earth's eclipse geometry), approximately what would be the distance from their planet to their star in kilometers?
A) 374,000,000 km B) 149,600,000 km C) 1,391,000,000 km D) 74,800,000 km
Reasoner Output: "Perfect eclipses mean moon and star look same size. Moon is at 374,000 km. Stars are way farther than moons, so maybe 1000 times farther? That gives 374,000,000 km. The answer is $\boxed{A)374,000,000km}$."

Late Training: Structured Multi-step Reasoning

Question: An alien civilization has a moon with diameter 3,475 km that creates perfect solar eclipses. Their star has the same diameter as our Sun. If their moon orbits at 374,000 km (maintaining the same angular size ratio as Earth's eclipse geometry), approximately what would be the distance from their planet to their star in kilometers?
A) 374,000,000 km B) 149,600,000 km C) 1,391,000,000 km D) 74,800,000 km
Reasoner Output: "Let me analyze this step-by-step.

Step 1: Identify given information
- Alien moon diameter = 3,475 km
- Alien star diameter = same as Sun = 1,391,000 km
- Moon orbital distance = 374,000 km
- Need: star distance for perfect eclipses

Step 2: Understand perfect eclipse condition
For perfect eclipses, angular sizes must be equal:
$\frac{\text{Moon diameter}}{\text{Moon distance}} = \frac{\text{Star diameter}}{\text{Star distance}}$

Step 3: Set up the equation
$\frac{3,475}{374,000} = \frac{1,391,000}{d}$

Step 4: Solve for star distance d
Cross multiply:
$3,475 \times d = 1,391,000 \times 374,000$
$d = \frac{1,391,000 \times 374,000}{3,475}$

Step 5: Calculate
$d = \frac{520,234,000,000}{3,475} = 149,708,489$ km

Step 6: Match to closest option
149,708,489 km $\approx$ 149,600,000 km

Step 7: Verify
Moon angular size: $\frac{3,475}{374,000} = 0.00929$
Star angular size: $\frac{1,391,000}{149,600,000} = 0.00930$
These match (small rounding difference), confirming our solution.

The answer is $\boxed{B)149,600,000km}$"

Evaluation Results

		Mathematical Reasoning							General Reasoning				
Method	Overall	AMC	Minerva	MATH 500	GSM8K	Olymp.	AIME 24	AIME 25	Super-GPQA	GPQA-Diamond	MMLU-Pro	BBEH	Δ vs Base
Qwen3-4B-Base													
Base Model	35.8	47.5	42.3	68.2	72.6	34.8	10.3	6.7	25.4	26.3	51.6	8.1	–
+ Strong Challenger	43.0	57.5	44.9	78.0	91.6	41.5	12.7	12.9	27.4	37.9	56.1	12.3	+7.2
+ R-Zero	39.5	45.0	52.2	72.6	90.5	39.7	10.1	5.2	27.8	27.8	53.7	10.4	+3.7
+ Absolute Zero	40.7	50.0	41.9	76.2	89.3	41.5	12.2	13.4	27.1	35.3	52.6	8.3	+4.9
+ **SPICE (ours)**	**44.9**	**57.5**	**51.9**	**78.0**	**92.7**	**42.7**	**12.2**	**19.1**	**30.2**	**39.4**	**58.1**	**12.3**	**+9.1**
Qwen3-8B-Base													
Base Model	43.0	57.5	49.3	74.4	91.2	40.4	15.3	12.1	31.0	33.3	58.1	10.5	–
+ Strong Challenger	45.6	60.0	52.5	78.2	92.4	41.6	15.3	12.1	35.0	35.8	64.8	14.4	+2.6
+ R-Zero	46.3	70.0	59.2	78.0	91.8	41.3	11.7	14.2	32.8	36.4	61.7	12.0	+3.3
+ Absolute Zero	46.5	62.5	52.9	76.6	92.0	47.8	18.4	18.2	33.5	36.8	62.5	10.8	+3.5
+ **SPICE (ours)**	**48.7**	**70.0**	**59.2**	**79.4**	**92.7**	**42.5**	**18.4**	**18.2**	**35.7**	**39.4**	**65.0**	**14.9**	**+5.7**
OctoThinker-3B-Hybrid-Base													
Base Model	14.7	15.0	14.3	38.2	44.3	10.8	3.4	3.5	12.8	3.5	14.9	1.0	–
+ Strong Challenger	21.0	15.0	15.1	39.7	73.7	14.4	0.3	0.1	15.7	25.3	24.2	7.0	+6.3
+ R-Zero	20.3	20.4	22.1	48.0	74.5	15.4	0.2	0.4	12.6	6.6	18.7	4.0	+5.6
+ Absolute Zero	21.7	17.5	18.7	43.2	75.8	13.2	2.7	0.5	17.7	19.2	24.3	6.1	+7.0
+ **SPICE (ours)**	**25.2**	**22.5**	**22.1**	**48.0**	**78.8**	**15.4**	**3.4**	**3.5**	**18.2**	**26.8**	**28.9**	**9.3**	**+10.5**
OctoThinker-8B-Hybrid-Base													
Base Model	20.5	27.5	22.1	44.2	68.6	16.7	2.7	0.8	11.4	15.7	14.7	0.6	–
+ Strong Challenger	28.2	25.0	27.2	54.4	86.4	20.0	6.6	3.3	17.6	27.8	32.9	9.5	+7.7
+ R-Zero	29.9	32.5	33.1	58.4	85.2	22.6	9.9	1.9	17.9	21.7	37.4	7.8	+9.4
+ Absolute Zero	29.4	32.5	34.9	56.8	87.0	25.6	0.1	3.3	18.8	27.8	31.4	5.0	+8.9
+ **SPICE (ours)**	**32.4**	**35.0**	**34.9**	**58.4**	**87.3**	**25.6**	**9.9**	**7.4**	**18.8**	**31.8**	**37.4**	**9.5**	**+11.9**

Ablation Study: Corpus Distribution

How corpus composition affects reasoning?

Algorithms

- With NaturalReasoning only
- With Nemotron-CC-Math only
- NaturalReasoning + Nemotron-CC-Math (SPICE)

Insights

- Domain-specific corpora improve performance in respective tasks
- Math + general reasoning corpora yields the best overall results

Corpus Type	Math Avg	General Avg	Overall Avg
NaturalReasoning	44.4	**37.0**	41.7
Nemotron-CC-Math	**53.4**	29.8	43.2
NaturalReasoning + Nemotron-CC-Math	50.6	35.0	**44.9**

Ablation Study: Task Type

How task formats affect reasoning?

Algorithms

- Multi-choice questions (MCQ) only
- Free-form questions only
- MCQ + free-form (SPICE)

Insights

- MCQ provides clear answer space and stable evaluation
- Free-form questions enable more flexible reasoning
- MCQ + Free-form achieves the best overall performance

Task Type	Math Avg	General Avg	Overall Avg
MCQ only	46.9	**35.7**	42.0
Free-form only	**52.5**	31.8	43.7
MCQ + Free-form	50.6	35.0	**44.9**

Ablation Study: Challenger's Reward

How challenger reward design affects training?

Algorithm

- Binary correctness reward
- Difficulty-based reward
- Variance-based reward (SPICE)

Insights

- Binary reward provides weak learning signals
- Difficulty-aware reward encourages harder questions
- Variance-based reward generates diverse and informative challenges

Challenger Reward	Math	General	Overall
Absolute Zero	48.2	30.8	40.7
Threshold	48.6	31.6	41.4
R-Zero	<u>50.0</u>	<u>33.9</u>	<u>43.6</u>
Variance (SPICE)	**50.6**	**35.0**	**44.9**

Conclusions

Key Findings from Experiments

- Corpus-grounded self-play significantly improves reasoning performance
- SPICE consistently outperforms prior self-play methods across multiple models
- Adversarial challenger-reasoner dynamics create an effective automatic curriculum

What These Results Suggest

- External grounding is critical for sustained self-improvement
- Pure self-play tends to stagnate due to hallucination and information symmetry
- Large corpora can act as an open-ended learning environment

Limitation

- Requires access to large external corpora
- No direct comparison with GRPO trained on curated datasets

Personal Takeaway

Personal Takeaway

- This work connects to the trend of self-improving reasoning systems for LLMs
- External knowledge grounding can act as a learning environment
- Future LLM training may rely on self-generated tasks rather than curated datasets
- Corpus grounding introduces a new paradigm of self-play RL algorithms

Future Works

- Moving beyond math and general reasoning benchmarks to specialized tasks
- Hybrid training with curated reasoning datasets
- Exploring richer learning environments