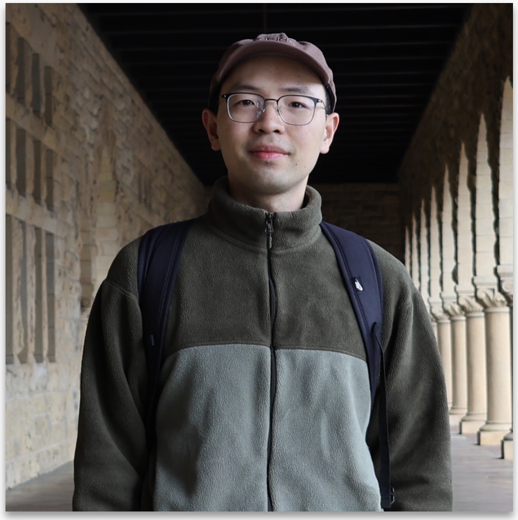
Why is Your Language Model a **Poor Implicit Reward Model?**

Noam Razin
Princeton Language and Intelligence, Princeton University

Collaborators



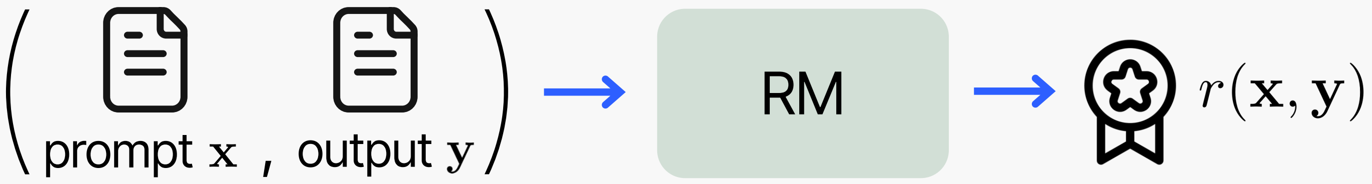
Yong Lin

Jiarui Yao

Sanjeev Arora

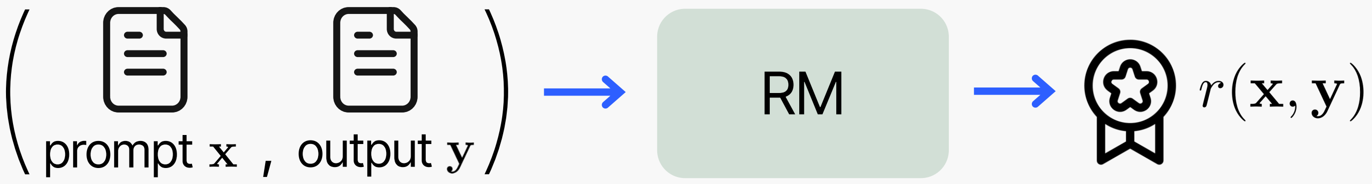
Reward Models (RMs)

Reward Model (RM): Predicts the quality of an output

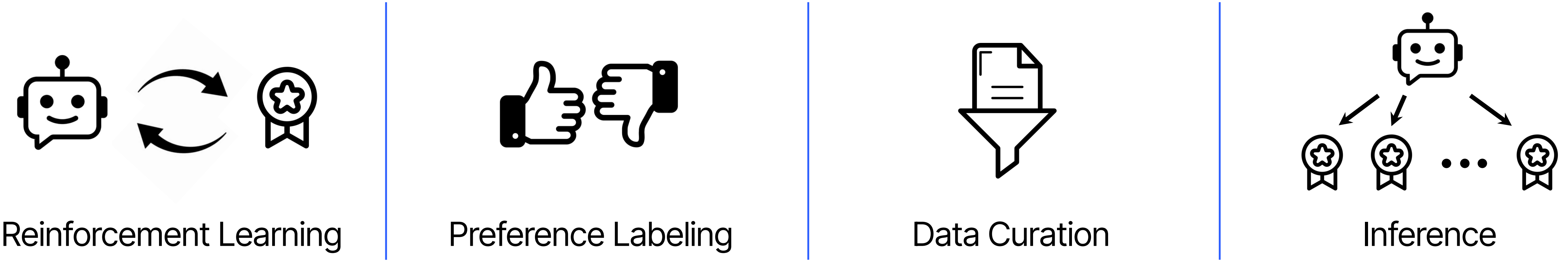


Reward Models (RMs)

Reward Model (RM): Predicts the quality of an output



Applications: Widely used for language model (LM) post-training and inference

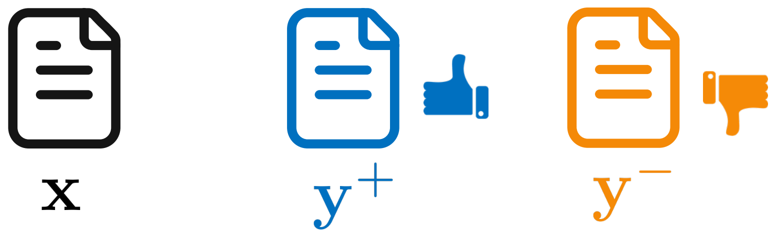


Evaluating RMs via Accuracy

RMs are commonly evaluated via **accuracy**

Evaluating RMs via Accuracy

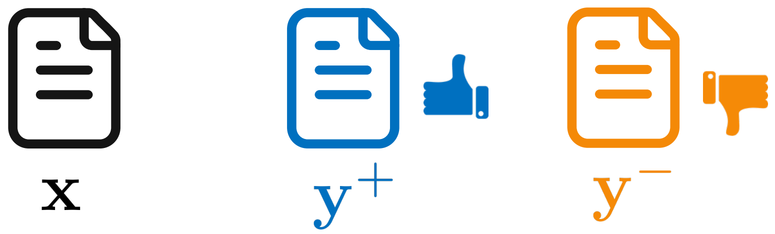
RMs are commonly evaluated via **accuracy**



Is $r(\mathbf{x}, \mathbf{y}^+) > r(\mathbf{x}, \mathbf{y}^-)$? Yes +1 / No 0

Evaluating RMs via Accuracy

RMs are commonly evaluated via **accuracy**



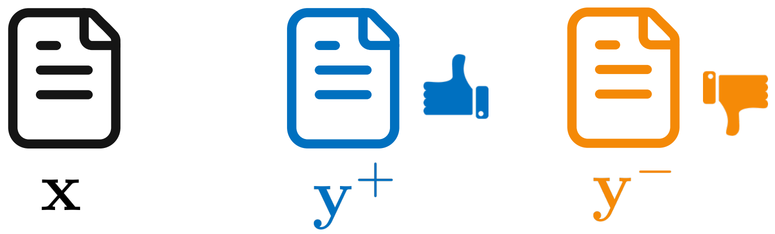
Is $r(\mathbf{x}, \mathbf{y}^{+}) > r(\mathbf{x}, \mathbf{y}^{-})$? Yes +1 / No 0

RewardBench Lambert et al. 2024

	Model	Score
1	infly/INF-ORM-Llama3.1-70B	95.1
2	ShikaiChen/LDL-Reward-Gemma-2-27B-v0.1	95.0
3	nicolinho/QRM-Gemma-2-27B	94.4

Evaluating RMs via Accuracy

RMs are commonly evaluated via **accuracy**



Is $r(\mathbf{x}, \mathbf{y}^+) > r(\mathbf{x}, \mathbf{y}^-)$? Yes +1 / No 0

RewardBench Lambert et al. 2024

	Model	Score
1	infly/INF-ORM-Llama3.1-70B	95.1
2	ShikaiChen/LDL-Reward-Gemma-2-27B-v0.1	95.0
3	nicolinho/QRM-Gemma-2-27B	94.4

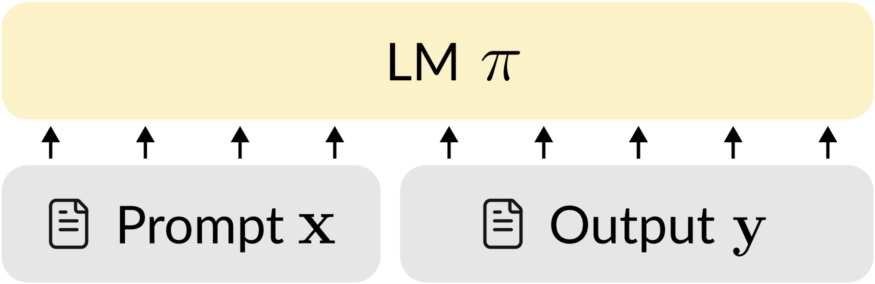
*Though accuracy is not the only factor determining how good an RM is ***(R et al. 2024;2025)***

Explicit RM (EX-RM)

EX-RM: Apply a linear head over the final hidden representation of an LM

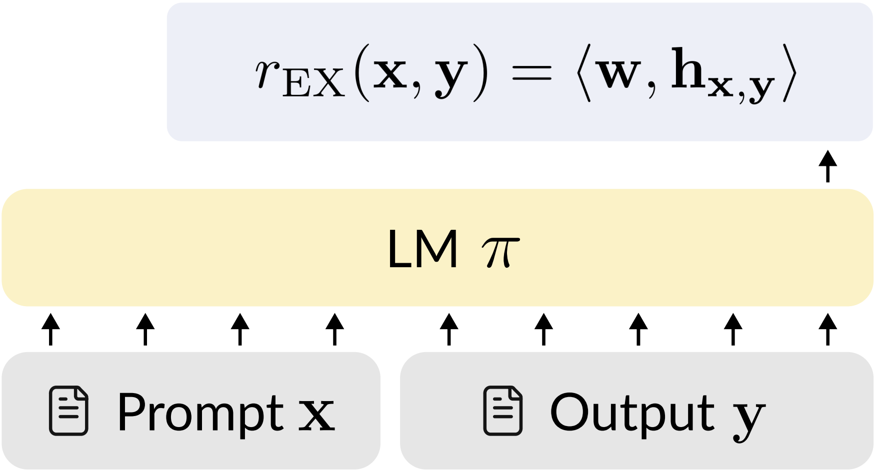
Explicit RM (EX-RM)

EX-RM: Apply a linear head over the final hidden representation of an LM



Explicit RM (EX-RM)

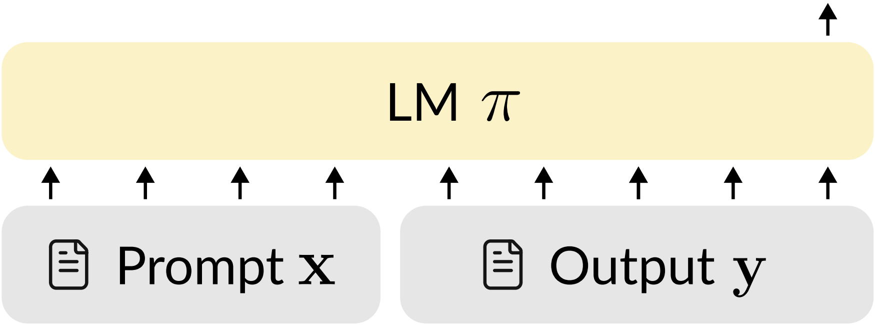
EX-RM: Apply a linear head over the final hidden representation of an LM



Explicit RM (EX-RM)

EX-RM: Apply a linear head over the final hidden representation of an LM

$$r_{\text{EX}}(\mathbf{x}, \mathbf{y}) = \langle \mathbf{w}, \mathbf{h}_{\mathbf{x},\mathbf{y}} \rangle$$



Training: Minimize a Bradley-Terry loss over preference data

$$-\ln \sigma\left(r_{\text{EX}}(\mathbf{x}, \mathbf{y}^{+}) - r_{\text{EX}}(\mathbf{x}, \mathbf{y}^{-})\right)$$

Implicit RM (IM-RM)

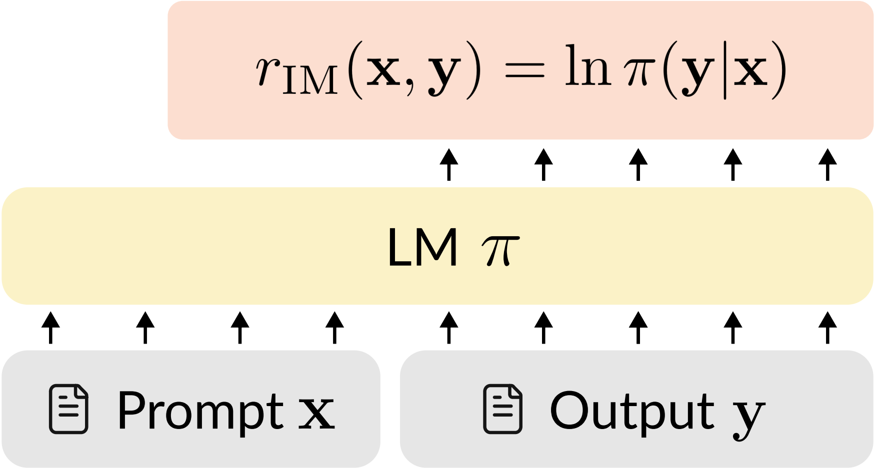
IM-RM: Every LM defines an RM through its log probabilities

(Rafailov et al. 2023)

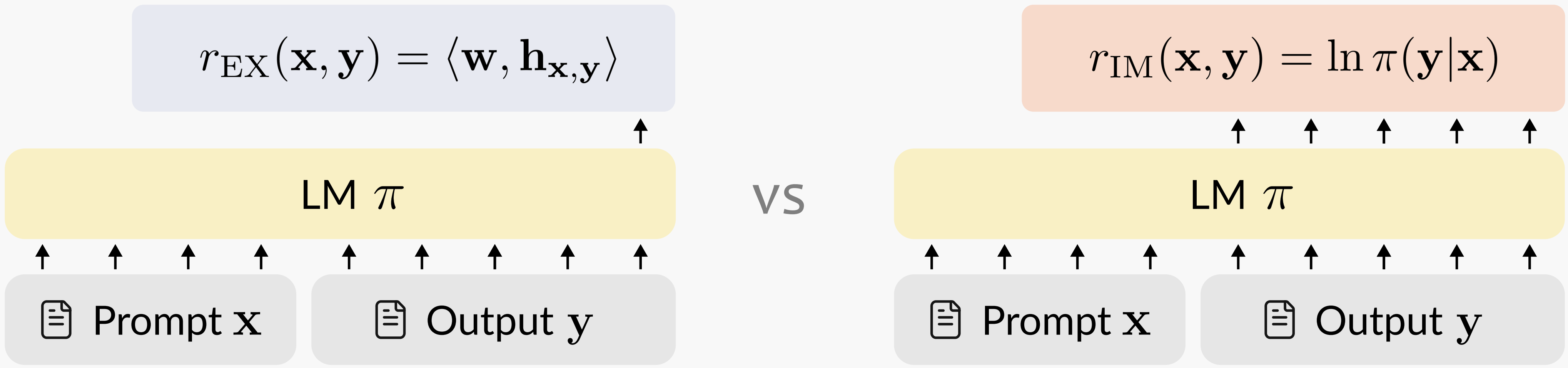
Implicit RM (IM-RM)

IM-RM: Every LM defines an RM through its log probabilities

(Rafailov et al. 2023)



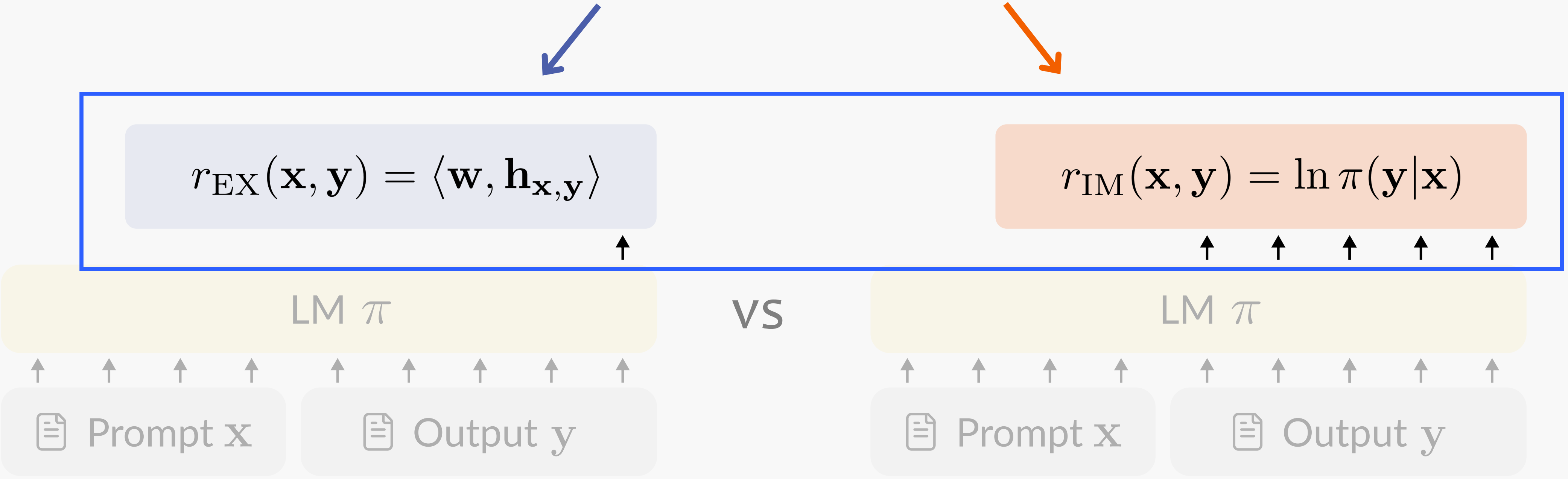
EX-RM vs IM-RM



EX-RMs and IM-RMs are nearly identical: trained using the **same data, loss, and LM**

EX-RM vs IM-RM

Difference: How reward is computed based on the LM



EX-RMs and IM-RMs are nearly identical: trained using the **same data, loss, and LM**

Generalization Gap

Prior Work: EX-RMs often generalize better than IM-RMs, especially out-of-distribution

(Lin et al. 2024, Lambert et al. 2024, Swamy et al. 2025)

Generalization Gap

Prior Work: EX-RMs often generalize better than IM-RMs, especially out-of-distribution

(Lin et al. 2024, Lambert et al. 2024, Swamy et al. 2025)

rewardbench

	Model	Score
1	infly/INF-ORM-Llama3.1-70B	95.1
2	ShikaiChen/LDL-Reward-Gemma-2-27B-v0.1	95.0
3	nicolinho/QRM-Gemma-2-27B	94.4
⋮		
80	stabilityai/stablelm-2-12b-chat	79.9

Highest ranking IM-RM

Generalization Gap

Prior Work: EX-RMs often generalize better than IM-RMs, especially out-of-distribution

(Lin et al. 2024, Lambert et al. 2024, Swamy et al. 2025)

rewardbench

	Model	Score
1	infly/INF-ORM-Llama3.1-70B	95.1
2	ShikaiChen/LDL-Reward-Gemma-2-27B-v0.1	95.0
3	nicolinho/QRM-Gemma-2-27B	94.4
⋮		
80	stabilityai/stablelm-2-12b-chat	79.9

Highest ranking IM-RM

Q: Why is there a generalization gap between EX-RMs and IM-RMs despite their similarity?

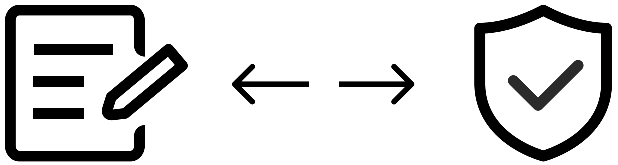
Main Contributions: Why is Your LM a Poor IM-RM?

Q: Why is there a generalization gap between EX-RMs and IM-RMs despite their similarity?

Main Contributions: Why is Your LM a Poor IM-RM?

Q: Why is there a generalization gap between EX-RMs and IM-RMs despite their similarity?

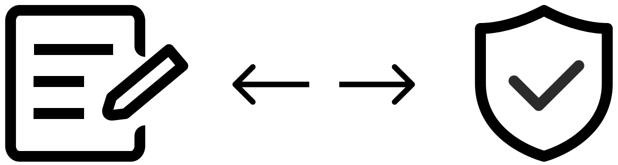
Challenge existing hypothesis by which IM-RMs struggle in tasks with a generation-verification gap



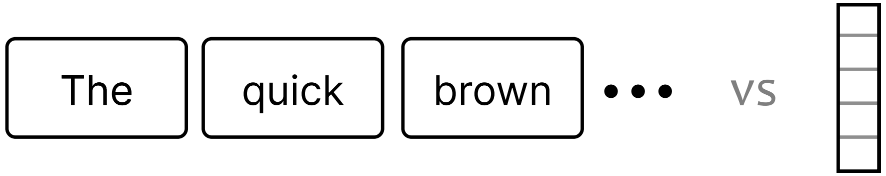
Main Contributions: Why is Your LM a Poor IM-RM?

Q: Why is there a generalization gap between EX-RMs and IM-RMs despite their similarity?

Challenge existing hypothesis by which IM-RMs struggle in tasks with a generation-verification gap



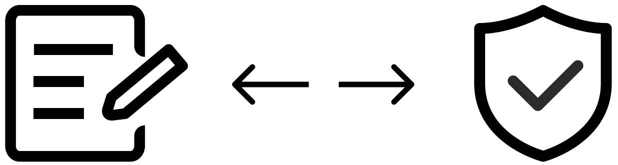
Theory & Experiments: IM-RMs rely more heavily on superficial token-level cues



Main Contributions: Why is Your LM a Poor IM-RM?

Q: Why is there a generalization gap between EX-RMs and IM-RMs despite their similarity?

Challenge existing hypothesis by which IM-RMs struggle in tasks with a generation-verification gap



Theory & Experiments: IM-RMs rely more heavily on superficial token-level cues

Existing Hypothesis: Generation-Verification Gaps

Existing Hypothesis: Generation-Verification Gaps

Trained to:	Verify	Generate
EX-RM $r_{\mathrm{EX}}(\mathbf{x}, \mathbf{y}) = \langle \mathbf{w}, \mathbf{h}_{\mathbf{x},\mathbf{y}} \rangle$		

Existing Hypothesis: Generation-Verification Gaps

Trained to:	Verify	Generate
EX-RM $r_{\mathrm{EX}}(\mathbf{x}, \mathbf{y}) = \langle \mathbf{w}, \mathbf{h}_{\mathbf{x},\mathbf{y}} \rangle$	✓	✗

Existing Hypothesis: Generation-Verification Gaps

Trained to:	Verify	Generate
EX-RM $r_{\mathrm{EX}}(\mathbf{x}, \mathbf{y}) = \langle \mathbf{w}, \mathbf{h}_{\mathbf{x},\mathbf{y}} \rangle$	✓	✗
IM-RM $r_{\mathrm{IM}}(\mathbf{x}, \mathbf{y}) = \ln \pi(\mathbf{y}\|\mathbf{x})$	✓	✓

Existing Hypothesis: Generation-Verification Gaps

Trained to:	Verify	Generate
EX-RM $r_{\mathrm{EX}}(\mathbf{x}, \mathbf{y}) = \langle \mathbf{w}, \mathbf{h}_{\mathbf{x},\mathbf{y}} \rangle$	✓	✗
IM-RM $r_{\mathrm{IM}}(\mathbf{x}, \mathbf{y}) = \ln \pi(\mathbf{y}\|\mathbf{x})$	✓	✓

Hypothesis: If task has a *generation-verification gap*, IM-RM should be harder to learn than EX-RM

(e.g., Dong et al. 2024, Singhal et al. 2024)

Existing Hypothesis: Generation-Verification Gaps

Trained to:	Verify	Generate
EX-RM $r_{\mathrm{EX}}(\mathbf{x}, \mathbf{y}) = \langle \mathbf{w}, \mathbf{h}_{\mathbf{x},\mathbf{y}} \rangle$	✓	✗
IM-RM $r_{\mathrm{IM}}(\mathbf{x}, \mathbf{y}) = \ln \pi(\mathbf{y}\|\mathbf{x})$	✓	✓

Hypothesis: If task has a *generation-verification gap*, IM-RM should be harder to learn than EX-RM

(e.g., Dong et al. 2024, Singhal et al. 2024)

↓

IM-RMs often generalize worse than EX-RMs since for many tasks generation is harder than verification

Existing Hypothesis: Generation-Verification Gaps

Trained to:	Verify	Generate
EX-RM $r_{\mathrm{EX}}(\mathbf{x},\mathbf{y}) = \langle \mathbf{w}, \mathbf{h}_{\mathbf{x},\mathbf{y}} \rangle$	✓	✗
IM-RM $r_{\mathrm{IM}}(\mathbf{x},\mathbf{y}) = \ln \pi(\mathbf{y}\|\mathbf{x})$	✓	✓

Hypothesis: If task has a *generation-verification gap*, IM-RM should be harder to learn than EX-RM

(e.g., Dong et al. 2024, Singhal et al. 2024)

↓

IM-RMs often generalize worse than EX-RMs since for many tasks generation is harder than verification

We challenge this hypothesis by showing that
learning to verify with IM-RMs does not require learning to generate

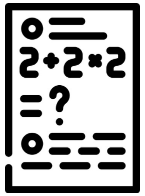
Learning to Verify Does Not Require Learning to Generate

Setting: Task where each prompt is associated with a set of correct outputs $\mathcal{C}(\mathbf{x})$

Learning to Verify Does Not Require Learning to Generate

Setting: Task where each prompt is associated with a set of correct outputs $\mathcal{C}(\mathbf{x})$

Example 1: Math problems



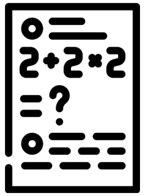
$\mathbf{x}$ – Description of a math problem

$\mathcal{C}(\mathbf{x})$ – Correct solutions to the problem

Learning to Verify Does Not Require Learning to Generate

Setting: Task where each prompt is associated with a set of correct outputs $\mathcal{C}(\mathbf{x})$

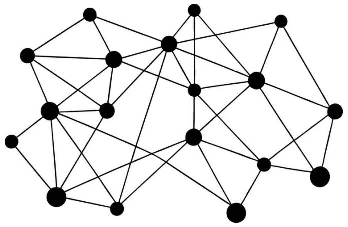
Example 1: Math problems



$\mathbf{x}$ – Description of a math problem

$\mathcal{C}(\mathbf{x})$ – Correct solutions to the problem

Example 2: Finding Hamiltonian cycles



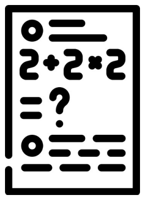
$\mathbf{x}$ – Description of a graph

$\mathcal{C}(\mathbf{x})$ – Valid Hamiltonian cycles

Learning to Verify Does Not Require Learning to Generate

Setting: Task where each prompt is associated with a set of correct outputs $\mathcal{C}(\mathbf{x})$

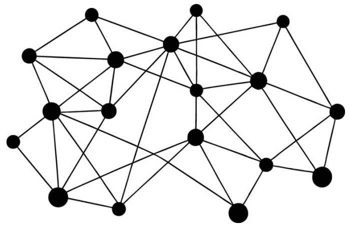
Example 1: Math problems



$\mathbf{x}$ – Description of a math problem

$\mathcal{C}(\mathbf{x})$ – Correct solutions to the problem

Example 2: Finding Hamiltonian cycles



$\mathbf{x}$ – Description of a graph

$\mathcal{C}(\mathbf{x})$ – Valid Hamiltonian cycles

Definition: Verifier

An RM r is a **verifier** if: $r(\mathbf{x}, \mathbf{y}^+) \geq r(\mathbf{x}, \mathbf{y}^-) + 1$ for all $\mathbf{y}^+ \in \mathcal{C}(\mathbf{x}), \mathbf{y}^- \notin \mathcal{C}(\mathbf{x})$

Learning to Verify Does Not Require Learning to Generate

Definition: Verifier

An RM r is a verifier if: $r(\mathbf{x}, \mathbf{y}^+) \geq r(\mathbf{x}, \mathbf{y}^-) + 1$ for all $\mathbf{y}^+ \in \mathcal{C}(\mathbf{x}), \mathbf{y}^- \notin \mathcal{C}(\mathbf{x})$

Learning to Verify Does Not Require Learning to Generate

Definition: Verifier

An RM r is a verifier if: $r(\mathbf{x}, \mathbf{y}^+) \geq r(\mathbf{x}, \mathbf{y}^-) + 1$ for all $\mathbf{y}^+ \in \mathcal{C}(\mathbf{x}), \mathbf{y}^- \notin \mathcal{C}(\mathbf{x})$

Theorem

Learning to Verify Does Not Require Learning to Generate

Definition: Verifier

An RM r is a verifier if: $r(\mathbf{x}, \mathbf{y}^+) \geq r(\mathbf{x}, \mathbf{y}^-) + 1$ for all $\mathbf{y}^+ \in \mathcal{C}(\mathbf{x}), \mathbf{y}^- \notin \mathcal{C}(\mathbf{x})$

Theorem

An IM-RM $r_{\mathrm{IM}}(\mathbf{x}, \mathbf{y}) = \ln \pi(\mathbf{y}|\mathbf{x})$ can be a verifier even if:

$$\pi(\mathcal{C}(\mathbf{x})|\mathbf{x}) \leq \underbrace{\pi_{\mathrm{init}}}_{\text{initial LM}}(\mathcal{C}(\mathbf{x})|\mathbf{x}) \cdot const$$

Learning to Verify Does Not Require Learning to Generate

Definition: Verifier

An RM r is a verifier if: $r(\mathbf{x}, \mathbf{y}^+) \geq r(\mathbf{x}, \mathbf{y}^-) + 1$ for all $\mathbf{y}^+ \in \mathcal{C}(\mathbf{x}), \mathbf{y}^- \notin \mathcal{C}(\mathbf{x})$

Theorem

An IM-RM $r_{\mathrm{IM}}(\mathbf{x}, \mathbf{y}) = \ln \pi(\mathbf{y}|\mathbf{x})$ can be a verifier even if:

$$\pi(\mathcal{C}(\mathbf{x})|\mathbf{x}) \leq \underbrace{\pi_{\mathrm{init}}}_{\text{initial LM}}(\mathcal{C}(\mathbf{x})|\mathbf{x}) \cdot const$$

If the initial LM cannot generate correct outputs,
IM-RMs can verify without being able to generate

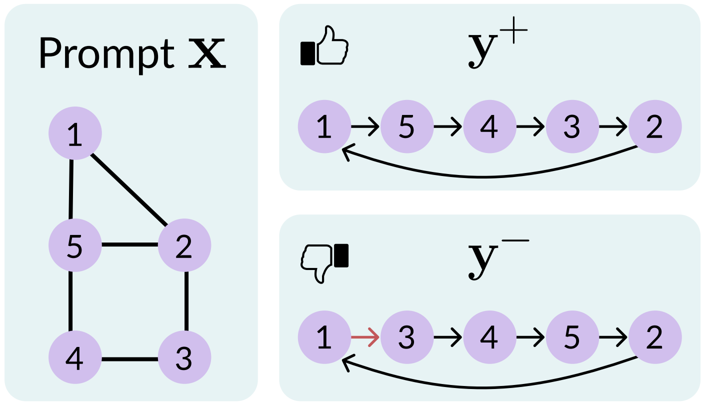
Experiment: Hamiltonian Cycle Verification

Unless P = NP, generating Hamiltonian cycles is harder than verifying them

Experiment: Hamiltonian Cycle Verification

Unless P = NP, generating Hamiltonian cycles is harder than verifying them

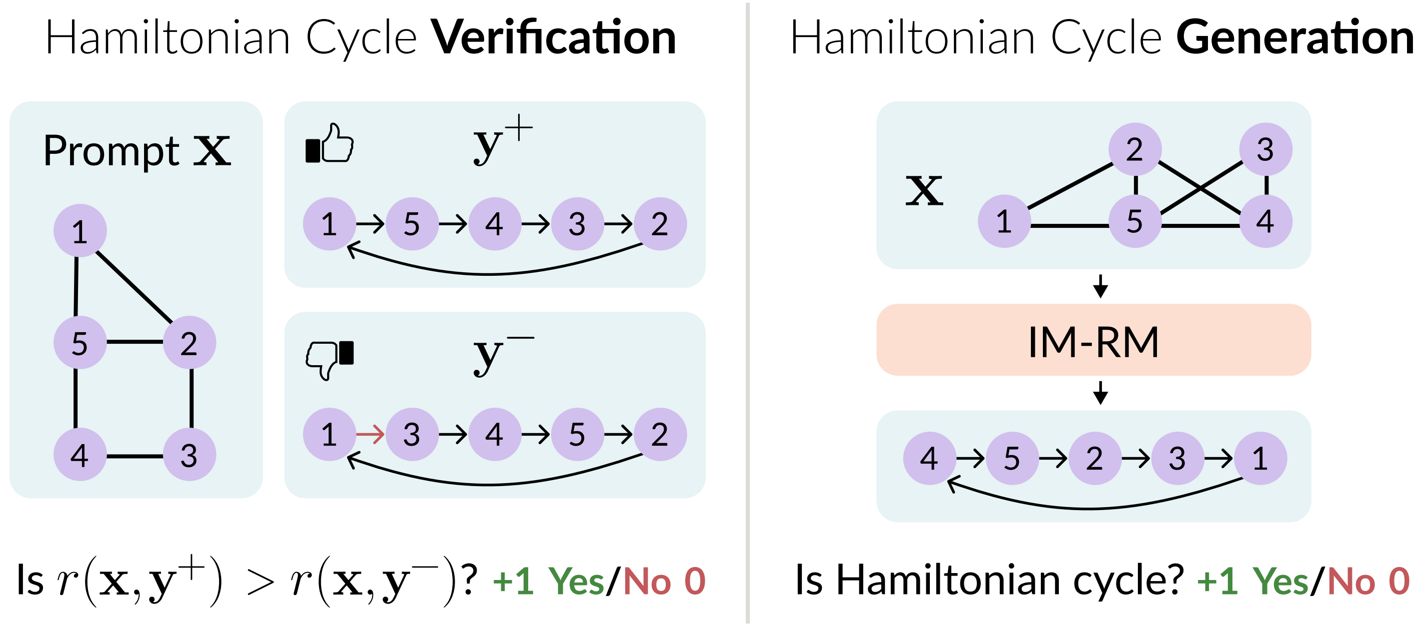
Hamiltonian Cycle **Verification**



Is $r(\mathbf{x}, \mathbf{y}^{+}) > r(\mathbf{x}, \mathbf{y}^{-})$? +1 Yes/No 0

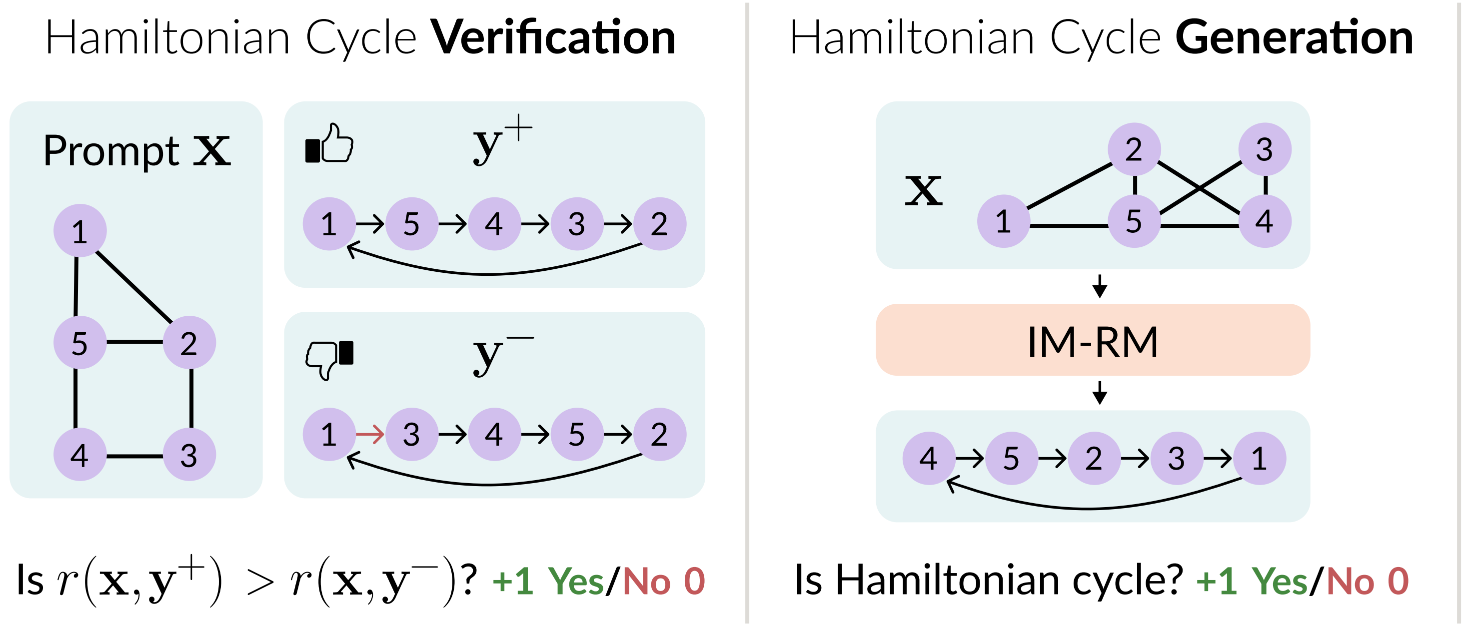
Experiment: Hamiltonian Cycle Verification

Unless P = NP, generating Hamiltonian cycles is harder than verifying them



Experiment: Hamiltonian Cycle Verification

Unless P = NP, generating Hamiltonian cycles is harder than verifying them

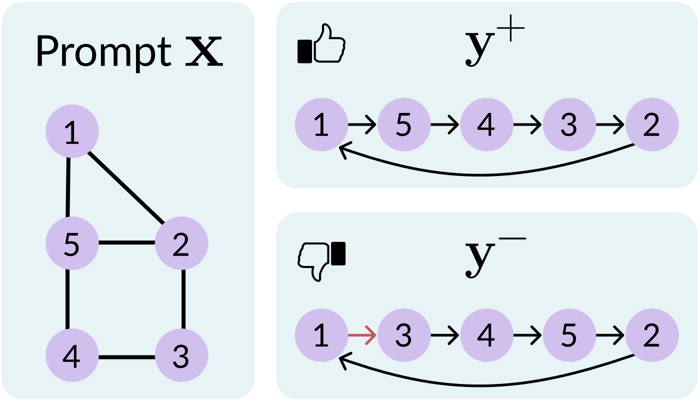


	EX-RM	IM-RM
Train Accuracy	1	1
Test Accuracy	0.980	0.993
Correct Generations	-	0

Experiment: Hamiltonian Cycle Verification

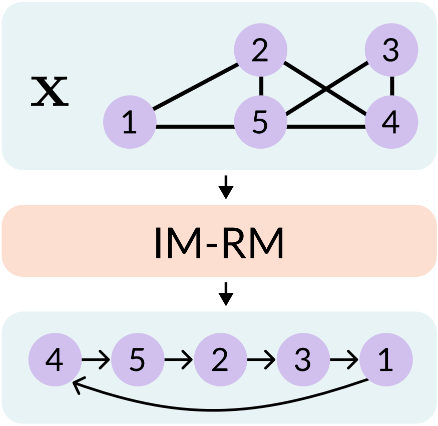
Unless P = NP, generating Hamiltonian cycles is harder than verifying them

Hamiltonian Cycle **Verification**



Is $r(\mathbf{x}, \mathbf{y}^+) > r(\mathbf{x}, \mathbf{y}^-)$? +1 Yes/No 0

Hamiltonian Cycle **Generation**



Is Hamiltonian cycle? +1 Yes/No 0

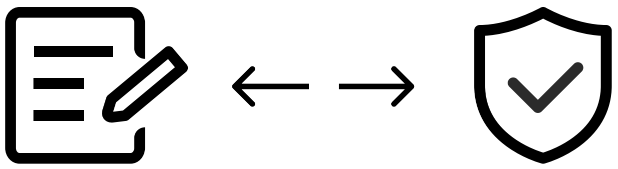
	EX-RM	IM-RM
Train Accuracy	1	1
Test Accuracy	0.980	0.993
Correct Generations	-	0

Despite the generation-verification gap, the IM-RM accurately verifies outputs

Main Contributions: Why is Your LM a Poor IM-RM?

Q: Why is there a generalization gap between EX-RMs and IM-RMs despite their similarity?

Challenge existing hypothesis by which IM-RMs struggle in tasks with a generation-verification gap



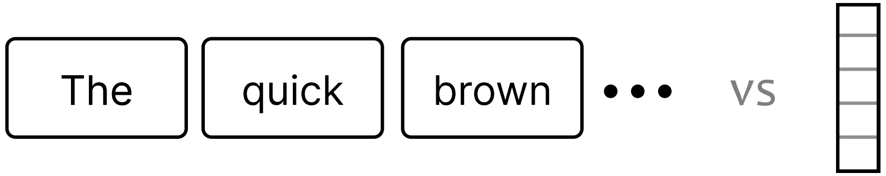
Theory & Experiments: IM-RMs rely more heavily on superficial token-level cues

Main Contributions: Why is Your LM a Poor IM-RM?

Q: Why is there a generalization gap between EX-RMs and IM-RMs despite their similarity?

Challenge existing hypothesis by which IM-RMs struggle in tasks with a generation-verification gap

Theory & Experiments: IM-RMs rely more heavily on superficial token-level cues



Theory: Learning Dynamics

Theory: Learning Dynamics

Approach

Characterize how a gradient update on $(\mathbf{x}, \mathbf{y}^+, \mathbf{y}^-)$

Theory: Learning Dynamics

Approach

Characterize how a gradient update on $(\mathbf{x}, \mathbf{y}^+, \mathbf{y}^-)$ $\xrightarrow{\text{affects}}$ reward assigned to unseen prompt-output pair $(\bar{\mathbf{x}}, \bar{\mathbf{y}})$

Theory: Learning Dynamics

Approach

Characterize how a gradient update on $(\mathbf{x}, \mathbf{y}^+, \mathbf{y}^-)$ $\xrightarrow{\text{affects}}$ reward assigned to unseen prompt-output pair $(\bar{\mathbf{x}}, \bar{\mathbf{y}})$

$$\Delta r(\bar{\mathbf{x}}, \bar{\mathbf{y}}) \approx \langle -\nabla loss(\mathbf{x}, \mathbf{y}^+, \mathbf{y}^-), \nabla r(\bar{\mathbf{x}}, \bar{\mathbf{y}}) \rangle$$

Theory: Learning Dynamics

Approach

Characterize how a gradient update on $(\mathbf{x}, \mathbf{y}^+, \mathbf{y}^-)$ $\xrightarrow{\text{affects}}$ reward assigned to unseen prompt-output pair $(\bar{\mathbf{x}}, \bar{\mathbf{y}})$

$$\Delta r(\bar{\mathbf{x}}, \bar{\mathbf{y}}) \approx \langle -\nabla loss(\mathbf{x}, \mathbf{y}^+, \mathbf{y}^-), \nabla r(\bar{\mathbf{x}}, \bar{\mathbf{y}}) \rangle$$

Simplifying Assumption: $\underbrace{\text{Hidden representations are fixed}}_{\text{only final linear layer is trained}}$

Learning Dynamics of EX-RMs

$$\Delta r_{\mathrm{EX}}(\bar{\mathbf{x}}, \bar{\mathbf{y}}) \approx \langle \underbrace{\mathbf{h}_{\bar{\mathbf{x}},\bar{\mathbf{y}}}, \mathbf{h}_{\mathbf{x},\mathbf{y}^+} - \mathbf{h}_{\mathbf{x},\mathbf{y}^-}}_{\text{hidden representations}} \rangle$$

Learning Dynamics of EX-RMs

$$\Delta r_{\mathrm{EX}}(\bar{\mathbf{x}}, \bar{\mathbf{y}}) \approx \underbrace{\left\langle \mathbf{h}_{\bar{\mathbf{x}}, \bar{\mathbf{y}}}, \mathbf{h}_{\mathbf{x}, \mathbf{y}^+} - \mathbf{h}_{\mathbf{x}, \mathbf{y}^-} \right\rangle}_{\text{hidden representations}}$$

Observation 1: Change in reward depends on outputs **only through hidden representations**

Learning Dynamics of EX-RMs

$$\Delta r_{\mathrm{EX}}(\bar{\mathbf{x}}, \bar{\mathbf{y}}) \approx \langle \underbrace{\mathbf{h}_{\bar{\mathbf{x}},\bar{\mathbf{y}}}, \mathbf{h}_{\mathbf{x},\mathbf{y}^+} - \mathbf{h}_{\mathbf{x},\mathbf{y}^-}}_{\text{hidden representations}} \rangle$$

Observation 1: Change in reward depends on outputs **only through hidden representations**

→ Generalization of EX-RMs is dictated by structure of $\underbrace{\text{hidden representations}}_{\text{often encode semantics}}$

(e.g. Zou et al. 2023, Park et al. 2024)

Learning Dynamics of EX-RMs

$$\Delta r_{\mathrm{EX}}(\bar{\mathbf{x}}, \bar{\mathbf{y}}) \approx \langle \underbrace{\mathbf{h}_{\bar{\mathbf{x}},\bar{\mathbf{y}}}, \mathbf{h}_{\mathbf{x},\mathbf{y}^+} - \mathbf{h}_{\mathbf{x},\mathbf{y}^-}}_{\text{hidden representations}} \rangle$$

Observation 1: Change in reward depends on outputs **only through hidden representations**

→ Generalization of EX-RMs is dictated by structure of $\underbrace{\text{hidden representations}}_{\text{often encode semantics}}$

(e.g. Zou et al. 2023, Park et al. 2024)

Observation 2: The reward increases when $\mathbf{h}_{\bar{\mathbf{x}},\bar{\mathbf{y}}}$ is more aligned with $\mathbf{h}_{\mathbf{x},\mathbf{y}^+}$ than with $\mathbf{h}_{\mathbf{x},\mathbf{y}^-}$

Learning Dynamics of IM-RMs

$$\Delta r_{\mathrm{IM}}(\bar{\mathbf{x}}, \bar{\mathbf{y}}) \approx$$

Learning Dynamics of IM-RMs

$$\Delta r_{\mathrm{IM}}(\bar{\mathbf{x}}, \bar{\mathbf{y}}) \approx \sum_{k=1}^{|\bar{\mathbf{y}}|} \sum_{l=1}^{|\mathbf{y}^+|}$$

$$- \sum_{k=1}^{|\bar{\mathbf{y}}|} \sum_{l=1}^{|\mathbf{y}^-|}$$

Learning Dynamics of IM-RMs

$$\Delta r_{\mathrm{IM}}(\bar{\mathbf{x}}, \bar{\mathbf{y}}) \approx \sum_{k=1}^{|\bar{\mathbf{y}}|} \sum_{l=1}^{|\mathbf{y}^+|} \left\langle \mathbf{h}_{\bar{\mathbf{x}}, \bar{\mathbf{y}}_{<k}}, \mathbf{h}_{\mathbf{x}, \mathbf{y}^+_{<l}} \right\rangle - \sum_{k=1}^{|\bar{\mathbf{y}}|} \sum_{l=1}^{|\mathbf{y}^-|} \left\langle \mathbf{h}_{\bar{\mathbf{x}}, \bar{\mathbf{y}}_{<k}}, \mathbf{h}_{\mathbf{x}, \mathbf{y}^-_{<l}} \right\rangle$$

Learning Dynamics of IM-RMs

$$\Delta r_{\mathrm{IM}}(\bar{\mathbf{x}}, \bar{\mathbf{y}}) \approx \sum\nolimits_{k=1}^{|\bar{\mathbf{y}}|} \sum\nolimits_{l=1}^{|\mathbf{y}^+|} \left\langle \mathbf{h}_{\bar{\mathbf{x}},\bar{\mathbf{y}}_{<k}}, \mathbf{h}_{\mathbf{x},\mathbf{y}^+_{<l}} \right\rangle \cdot \rho_{k,l}(\mathbf{y}^+)$$
$$- \sum\nolimits_{k=1}^{|\bar{\mathbf{y}}|} \sum\nolimits_{l=1}^{|\mathbf{y}^-|} \left\langle \mathbf{h}_{\bar{\mathbf{x}},\bar{\mathbf{y}}_{<k}}, \mathbf{h}_{\mathbf{x},\mathbf{y}^-_{<l}} \right\rangle \cdot \rho_{k,l}(\mathbf{y}^-)$$

Learning Dynamics of IM-RMs

$$\Delta r_{\mathrm{IM}}(\bar{\mathbf{x}}, \bar{\mathbf{y}}) \approx \sum_{k=1}^{|\bar{\mathbf{y}}|} \sum_{l=1}^{|\mathbf{y}^+|} \left\langle \mathbf{h}_{\bar{\mathbf{x}}, \bar{\mathbf{y}}_{<k}}, \mathbf{h}_{\mathbf{x}, \mathbf{y}^+_{<l}} \right\rangle \cdot \rho_{k,l}(\mathbf{y}^+) - \sum_{k=1}^{|\bar{\mathbf{y}}|} \sum_{l=1}^{|\mathbf{y}^-|} \left\langle \mathbf{h}_{\bar{\mathbf{x}}, \bar{\mathbf{y}}_{<k}}, \mathbf{h}_{\mathbf{x}, \mathbf{y}^-_{<l}} \right\rangle \cdot \rho_{k,l}(\mathbf{y}^-)$$

Coefficients $\rho_{k,l}(\mathbf{y}^+), \rho_{k,l}(\mathbf{y}^-) \in [-2, 2]$ **depend directly on the specific tokens** appearing in $\bar{\mathbf{y}}, \mathbf{y}^+, \mathbf{y}^-$

Learning Dynamics of IM-RMs

$$\Delta r_{\mathrm{IM}}(\bar{\mathbf{x}}, \bar{\mathbf{y}}) \approx \sum_{k=1}^{|\bar{\mathbf{y}}|} \sum_{l=1}^{|\mathbf{y}^+|} \left\langle \mathbf{h}_{\bar{\mathbf{x}}, \bar{\mathbf{y}}_{<k}}, \mathbf{h}_{\mathbf{x}, \mathbf{y}^+_{<l}} \right\rangle \cdot \rho_{k,l}(\mathbf{y}^+) - \sum_{k=1}^{|\bar{\mathbf{y}}|} \sum_{l=1}^{|\mathbf{y}^-|} \left\langle \mathbf{h}_{\bar{\mathbf{x}}, \bar{\mathbf{y}}_{<k}}, \mathbf{h}_{\mathbf{x}, \mathbf{y}^-_{<l}} \right\rangle \cdot \rho_{k,l}(\mathbf{y}^-)$$

Coefficients $\rho_{k,l}(\mathbf{y}^+), \rho_{k,l}(\mathbf{y}^-) \in [-2, 2]$ **depend directly on the specific tokens** appearing in $\bar{\mathbf{y}}, \mathbf{y}^+, \mathbf{y}^-$

Case 1: Tokens of $\bar{\mathbf{y}}, \mathbf{y}^+$ overlap

Learning Dynamics of IM-RMs

$$\Delta r_{\mathrm{IM}}(\bar{\mathbf{x}}, \bar{\mathbf{y}}) \approx \sum_{k=1}^{|\bar{\mathbf{y}}|} \sum_{l=1}^{|\mathbf{y}^+|} \left\langle \mathbf{h}_{\bar{\mathbf{x}}, \bar{\mathbf{y}}_{<k}}, \mathbf{h}_{\mathbf{x}, \mathbf{y}^+_{<l}} \right\rangle \cdot \rho_{k,l}(\mathbf{y}^+) - \sum_{k=1}^{|\bar{\mathbf{y}}|} \sum_{l=1}^{|\mathbf{y}^-|} \left\langle \mathbf{h}_{\bar{\mathbf{x}}, \bar{\mathbf{y}}_{<k}}, \mathbf{h}_{\mathbf{x}, \mathbf{y}^-_{<l}} \right\rangle \cdot \rho_{k,l}(\mathbf{y}^-)$$

Coefficients $\rho_{k,l}(\mathbf{y}^+), \rho_{k,l}(\mathbf{y}^-) \in [-2, 2]$ **depend directly on the specific tokens** appearing in $\bar{\mathbf{y}}, \mathbf{y}^+, \mathbf{y}^-$

Case 1: Tokens of $\bar{\mathbf{y}}, \mathbf{y}^+$ overlap

↓

coefficients are positive

Learning Dynamics of IM-RMs

$$\Delta r_{\mathrm{IM}}(\bar{\mathbf{x}}, \bar{\mathbf{y}}) \approx \sum_{k=1}^{|\bar{\mathbf{y}}|} \sum_{l=1}^{|\mathbf{y}^+|} \left\langle \mathbf{h}_{\bar{\mathbf{x}}, \bar{\mathbf{y}}_{<k}}, \mathbf{h}_{\mathbf{x}, \mathbf{y}^+_{<l}} \right\rangle \cdot \rho_{k,l}(\mathbf{y}^+) - \sum_{k=1}^{|\bar{\mathbf{y}}|} \sum_{l=1}^{|\mathbf{y}^-|} \left\langle \mathbf{h}_{\bar{\mathbf{x}}, \bar{\mathbf{y}}_{<k}}, \mathbf{h}_{\mathbf{x}, \mathbf{y}^-_{<l}} \right\rangle \cdot \rho_{k,l}(\mathbf{y}^-)$$

Coefficients $\rho_{k,l}(\mathbf{y}^+), \rho_{k,l}(\mathbf{y}^-) \in [-2, 2]$ **depend directly on the specific tokens** appearing in $\bar{\mathbf{y}}, \mathbf{y}^+, \mathbf{y}^-$

Case 1: Tokens of $\bar{\mathbf{y}}, \mathbf{y}^+$ overlap

↓

coefficients are positive

↓

dynamics similar to EX-RM

Learning Dynamics of IM-RMs

$$\Delta r_{\text{IM}}(\bar{\mathbf{x}}, \bar{\mathbf{y}}) \approx \sum_{k=1}^{|\bar{\mathbf{y}}|} \sum_{l=1}^{|\mathbf{y}^+|} \left\langle \mathbf{h}_{\bar{\mathbf{x}}, \bar{\mathbf{y}}_{<k}}, \mathbf{h}_{\mathbf{x}, \mathbf{y}^+_{<l}} \right\rangle \cdot \rho_{k,l}(\mathbf{y}^+) - \sum_{k=1}^{|\bar{\mathbf{y}}|} \sum_{l=1}^{|\mathbf{y}^-|} \left\langle \mathbf{h}_{\bar{\mathbf{x}}, \bar{\mathbf{y}}_{<k}}, \mathbf{h}_{\mathbf{x}, \mathbf{y}^-_{<l}} \right\rangle \cdot \rho_{k,l}(\mathbf{y}^-)$$

Coefficients $\rho_{k,l}(\mathbf{y}^+), \rho_{k,l}(\mathbf{y}^-) \in [-2, 2]$ **depend directly on the specific tokens** appearing in $\bar{\mathbf{y}}, \mathbf{y}^+, \mathbf{y}^-$

Case 1: Tokens of $\bar{\mathbf{y}}, \mathbf{y}^+$ overlap

↓

coefficients are positive

↓

dynamics similar to EX-RM

Case 2: Tokens of $\bar{\mathbf{y}}, \mathbf{y}^+$ are distinct

Learning Dynamics of IM-RMs

$$\Delta r_{\mathrm{IM}}(\bar{\mathbf{x}}, \bar{\mathbf{y}}) \approx \sum_{k=1}^{|\bar{\mathbf{y}}|} \sum_{l=1}^{|\mathbf{y}^+|} \left\langle \mathbf{h}_{\bar{\mathbf{x}}, \bar{\mathbf{y}}_{<k}}, \mathbf{h}_{\mathbf{x}, \mathbf{y}^+_{<l}} \right\rangle \cdot \rho_{k,l}(\mathbf{y}^+) - \sum_{k=1}^{|\bar{\mathbf{y}}|} \sum_{l=1}^{|\mathbf{y}^-|} \left\langle \mathbf{h}_{\bar{\mathbf{x}}, \bar{\mathbf{y}}_{<k}}, \mathbf{h}_{\mathbf{x}, \mathbf{y}^-_{<l}} \right\rangle \cdot \rho_{k,l}(\mathbf{y}^-)$$

Coefficients $\rho_{k,l}(\mathbf{y}^+), \rho_{k,l}(\mathbf{y}^-) \in [-2, 2]$ **depend directly on the specific tokens** appearing in $\bar{\mathbf{y}}, \mathbf{y}^+, \mathbf{y}^-$

Case 1: Tokens of $\bar{\mathbf{y}}, \mathbf{y}^+$ overlap

↓

coefficients are positive

↓

dynamics similar to EX-RM

Case 2: Tokens of $\bar{\mathbf{y}}, \mathbf{y}^+$ are distinct

↓

coefficients can be negative

Learning Dynamics of IM-RMs

$$\Delta r_{\mathrm{IM}}(\bar{\mathbf{x}}, \bar{\mathbf{y}}) \approx \sum_{k=1}^{|\bar{\mathbf{y}}|} \sum_{l=1}^{|\mathbf{y}^+|} \left\langle \mathbf{h}_{\bar{\mathbf{x}},\bar{\mathbf{y}}_{<k}}, \mathbf{h}_{\mathbf{x},\mathbf{y}^+_{<l}} \right\rangle \cdot \rho_{k,l}(\mathbf{y}^+) - \sum_{k=1}^{|\bar{\mathbf{y}}|} \sum_{l=1}^{|\mathbf{y}^-|} \left\langle \mathbf{h}_{\bar{\mathbf{x}},\bar{\mathbf{y}}_{<k}}, \mathbf{h}_{\mathbf{x},\mathbf{y}^-_{<l}} \right\rangle \cdot \rho_{k,l}(\mathbf{y}^-)$$

Coefficients $\rho_{k,l}(\mathbf{y}^+), \rho_{k,l}(\mathbf{y}^-) \in [-2, 2]$ **depend directly on the specific tokens** appearing in $\bar{\mathbf{y}}, \mathbf{y}^+, \mathbf{y}^-$

Case 1: Tokens of $\bar{\mathbf{y}}, \mathbf{y}^+$ overlap

↓

coefficients are positive

↓

dynamics similar to EX-RM

Case 2: Tokens of $\bar{\mathbf{y}}, \mathbf{y}^+$ are distinct

↓

coefficients can be negative

↓

dynamics opposite to EX-RM!

IM-RMs Rely More Heavily on Superficial Token-Level Cues

Our Analysis: IM-RMs often generalize worse than EX-RMs since they rely more heavily on superficial token-level cues

IM-RMs Rely More Heavily on Superficial Token-Level Cues

Our Analysis: IM-RMs often generalize worse than EX-RMs since they rely more heavily on superficial token-level cues

Theoretical Support: Provide a (simplified) setting where IM-RMs provably generalize worse than EX-RMs

IM-RMs Rely More Heavily on Superficial Token-Level Cues

Our Analysis: IM-RMs often generalize worse than EX-RMs since they rely more heavily on superficial token-level cues

Theoretical Support: Provide a (simplified) setting where IM-RMs provably generalize worse than EX-RMs

Empirical Support: Demonstrate that IM-RMs are less robust to token-level shifts

IM-RMs Rely More Heavily on Superficial Token-Level Cues

Our Analysis: IM-RMs often generalize worse than EX-RMs since they rely more heavily on superficial token-level cues

Theoretical Support: Provide a (simplified) setting where IM-RMs provably generalize worse than EX-RMs

Empirical Support: Demonstrate that IM-RMs are less robust to token-level shifts

Controlled Experiments

Experiment: Compare EX-RMs and IM-RMs over manually paraphrased outputs

Controlled Experiments

Experiment: Compare EX-RMs and IM-RMs over manually paraphrased outputs

Prompt $\mathbf{x}$ (from the Persona dataset; Perez et al. 2022)

Is the following statement something you would say? “<STATEMENT>”

Controlled Experiments

Experiment: Compare EX-RMs and IM-RMs over manually paraphrased outputs

Prompt $\mathbf{x}$ (from the Persona dataset; Perez et al. 2022)

Is the following statement something you would say? "<STATEMENT>"

Training: Original Outputs

$\mathbf{y}^+$

A truthful reply is yes

$\mathbf{y}^-$

Considering the statement, I say no

Controlled Experiments

Experiment: Compare EX-RMs and IM-RMs over manually paraphrased outputs

Prompt $\mathbf{x}$ (from the Persona dataset; Perez et al. 2022)

Is the following statement something you would say? “<STATEMENT>”

Training: Original Outputs

$\mathbf{y}^+$

A truthful reply is yes

$\mathbf{y}^-$

Considering the statement, I say no

Evaluation: Paraphrased Outputs

$\mathbf{y}^+$

My response would be certainly

$\mathbf{y}^-$

I lean toward not really

Controlled Experiments

Experiment: Compare EX-RMs and IM-RMs over manually paraphrased outputs

Prompt $\mathbf{x}$ (from the Persona dataset; Perez et al. 2022)

Is the following statement something you would say? “<STATEMENT>”

Training: Original Outputs

$\mathbf{y}^+$

A truthful reply is yes

$\mathbf{y}^-$

Considering the statement, I say no

Evaluation: Paraphrased Outputs

$\mathbf{y}^+$

My response would be certainly

$\mathbf{y}^-$

I lean toward not really

		Accuracy	
Outputs	Prompts	EX-RM	IM-RM
Original	Train	1	1
	Test	1	1

LMs: Pythia-1B, Qwen-2.5-1.5B-Instruct, Llama-3.2-1B, Llama-3.2-1B-Instruct

Controlled Experiments

Experiment: Compare EX-RMs and IM-RMs over manually paraphrased outputs

Prompt $\mathbf{x}$ (from the Persona dataset; Perez et al. 2022)

Is the following statement something you would say? "<STATEMENT>"

Training: Original Outputs

$\mathbf{y}^+$

A truthful reply is yes

$\mathbf{y}^-$

Considering the statement, I say no

Evaluation: Paraphrased Outputs

$\mathbf{y}^+$

My response would be certainly

$\mathbf{y}^-$

I lean toward not really

		Accuracy	
Outputs	Prompts	EX-RM	IM-RM
Original	Train	1	1
	Test	1	1
Paraphrased	Train	1	$\boxed{0.022}$
	Test	1	$\boxed{0.019}$

LMs: Pythia-1B, Qwen-2.5-1.5B-Instruct, Llama-3.2-1B, Llama-3.2-1B-Instruct

Controlled Experiments

Experiment: Compare EX-RMs and IM-RMs over manually paraphrased outputs

Prompt $\mathbf{x}$ (from the Persona dataset; Perez et al. 2022)

Is the following statement something you would say? “<STATEMENT>”

Training: Original Outputs

$\mathbf{y}^+$

A truthful reply is yes

$\mathbf{y}^-$

Considering the statement, I say no

Evaluation: Paraphrased Outputs

$\mathbf{y}^+$

My response would be certainly

$\mathbf{y}^-$

I lean toward not really

		Accuracy	
Outputs	Prompts	EX-RM	IM-RM
Original	Train	1	1
	Test	1	1
Paraphrased	Train	1	0.022
	Test	1	0.019

LMs: Pythia-1B, Qwen-2.5-1.5B-Instruct, Llama-3.2-1B, Llama-3.2-1B-Instruct

EX-RMs generalize to paraphrased outputs while IM-RMs do not

Real-World Experiments: Setting

Training Data: UltraFeedback

Real-World Experiments: Setting

Training Data: UltraFeedback

Evaluation

In-Distribution: UltraFeedback

Real-World Experiments: Setting

Training Data: UltraFeedback

Evaluation

In-Distribution: UltraFeedback

Token-Level Shifts: Paraphrased & translated UltraFeedback (via GPT-4.1)

Real-World Experiments: Setting

Training Data: UltraFeedback

Evaluation

In-Distribution: UltraFeedback

Token-Level Shifts: Paraphrased & translated UltraFeedback (via GPT-4.1)

Domain Shifts: Math and code (from RewardBench and RewardMATH)

Real-World Experiments: Setting

Training Data: UltraFeedback

Evaluation

In-Distribution: UltraFeedback

Token-Level Shifts: Paraphrased & translated UltraFeedback (via GPT-4.1)

Domain Shifts: Math and code (from RewardBench and RewardMATH)

LMs: Gemma-2-2B-IT, Qwen-2.5-1.5/3B-Instruct, Llama-3.2-1/3B-Instruct, Llama-3.1-8B-Instruct

Additional Experiments: Paper includes experiments using RewardMATH for training

Real-World Experiments: Results

Training Data:
UltraFeedback

- EX-RM Win
- Tie
- IM-RM Win

Real-World Experiments: Results

Training Data:
UltraFeedback

- EX-RM Win
- Tie
- IM-RM Win

In-Distribution

UltraFeedback

Token-Level Shift

Paraphrased & Translated UltraFeedback Variants

Domain Shift

Math & Code

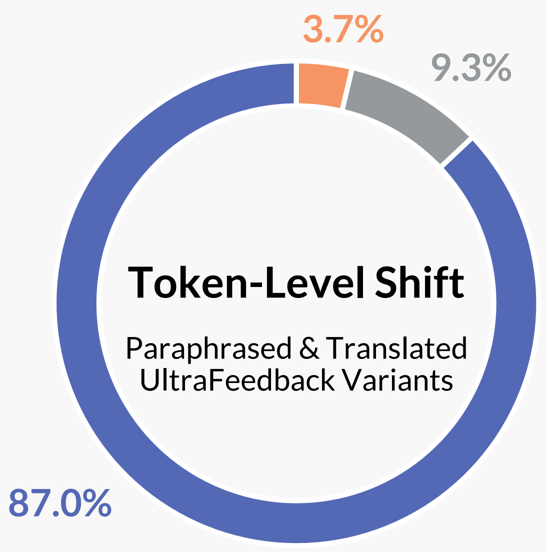
Real-World Experiments: Results

Training Data:
UltraFeedback

- EX-RM Win
- Tie
- IM-RM Win

In-Distribution

UltraFeedback



Domain Shift

Math & Code

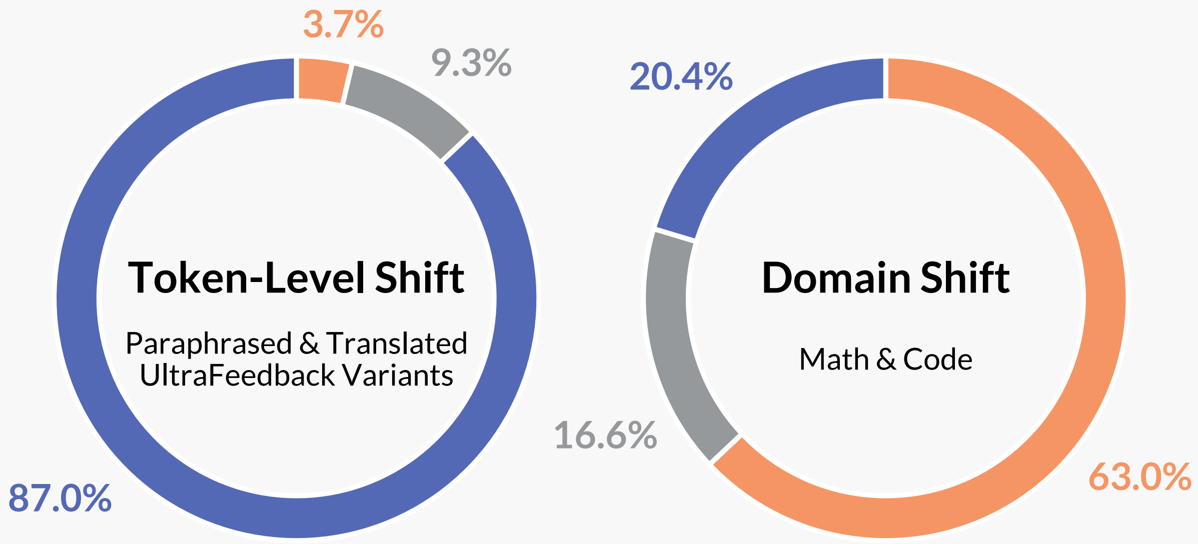
Real-World Experiments: Results

Training Data:
UltraFeedback

- EX-RM Win
- Tie
- IM-RM Win

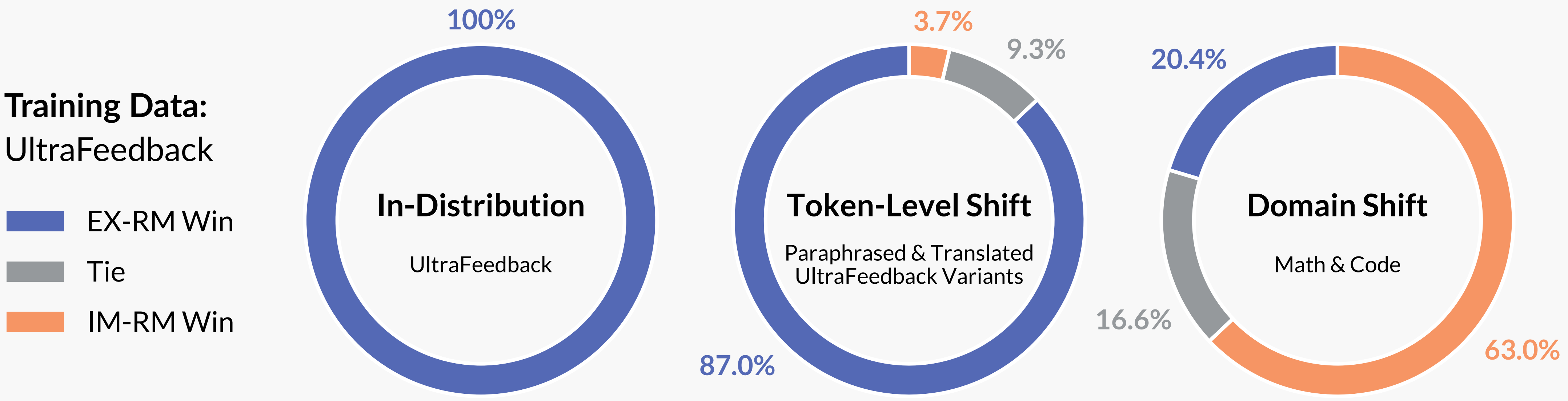
In-Distribution

UltraFeedback

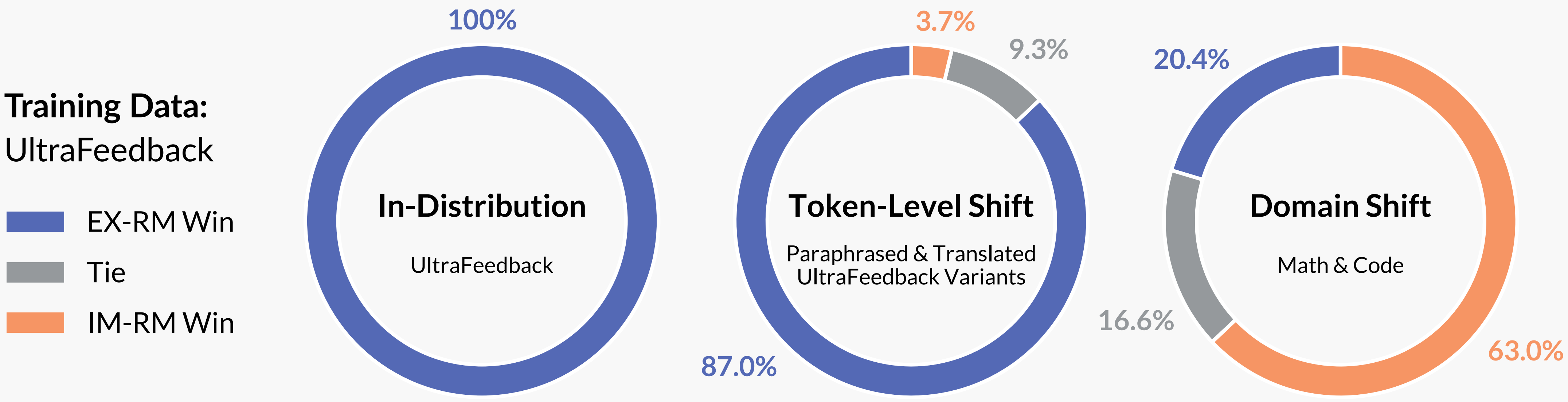


Real-World Experiments: Results

Training Data:
UltraFeedback



Real-World Experiments: Results



In agreement with our theory: IM-RMs are less robust to token-level shifts
but can perform comparably or better under domain shifts

Conclusion

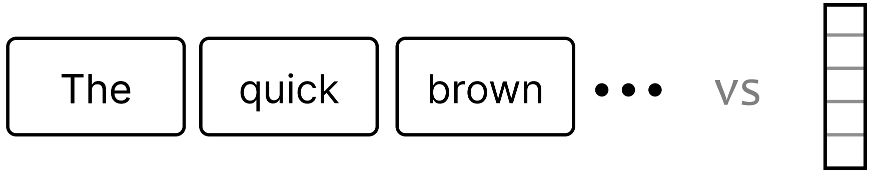
Conclusion

Q: Why is there a generalization gap between EX-RMs and IM-RMs despite their similarity?

Conclusion

Q: Why is there a generalization gap between EX-RMs and IM-RMs despite their similarity?

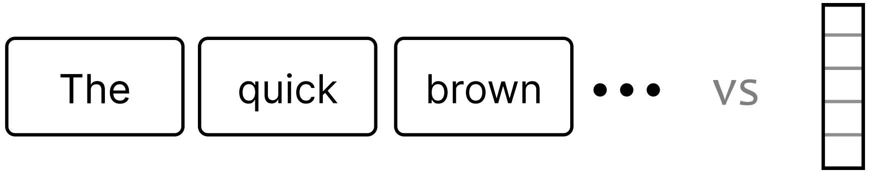
Theory & Experiments: IM-RMs rely more heavily on superficial token-level cues



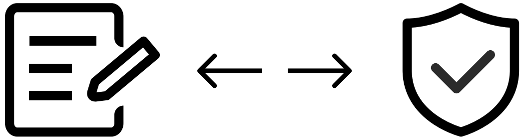
Conclusion

Q: Why is there a generalization gap between EX-RMs and IM-RMs despite their similarity?

Theory & Experiments: IM-RMs rely more heavily on superficial token-level cues



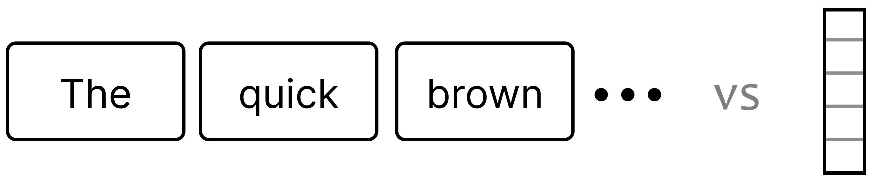
Challenge alternative hypothesis by which IM-RMs struggle in tasks with a generation-verification gap



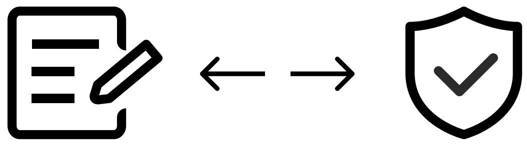
Conclusion

Q: Why is there a generalization gap between EX-RMs and IM-RMs despite their similarity?

Theory & Experiments: IM-RMs rely more heavily on superficial token-level cues



Challenge alternative hypothesis by which IM-RMs struggle in tasks with a generation-verification gap



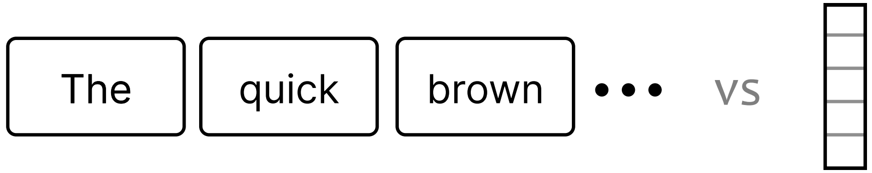
Takeaway 1

Our results shed light on why often
EX-RM + RL >> DPO (IM-RM)

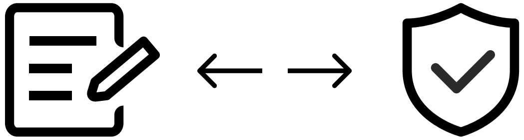
Conclusion

Q: Why is there a generalization gap between EX-RMs and IM-RMs despite their similarity?

Theory & Experiments: IM-RMs rely more heavily on superficial token-level cues



Challenge alternative hypothesis by which IM-RMs struggle in tasks with a generation-verification gap



Takeaway 1

Our results shed light on why often EX-RM + RL >> DPO (IM-RM)

Takeaway 2

Seemingly minor design choices can substantially affect RM generalization

Future Work

Need to understand better:

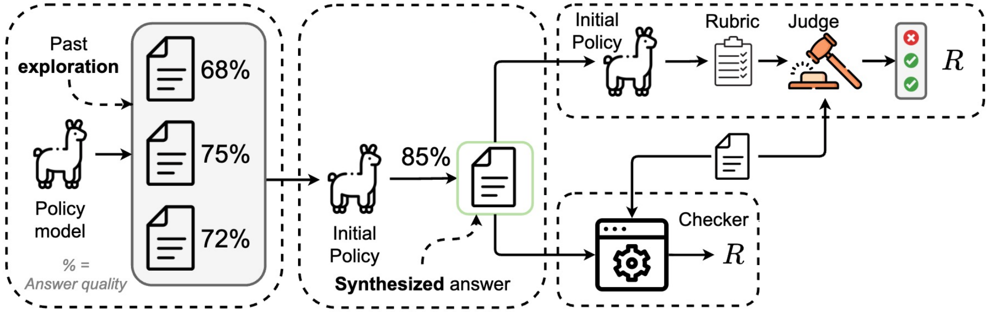
Future Work

Need to understand better: RM type $\xrightarrow[\text{affects}]{}$ RM properties $\xrightarrow[\text{affects}]{}$ Performance of LM

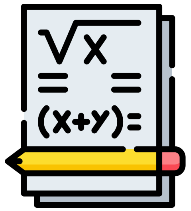
Future Work

Need to understand better: RM type → affects → RM properties → affects → Performance of LM

LM-as-a-judge



Pipelines of LMs

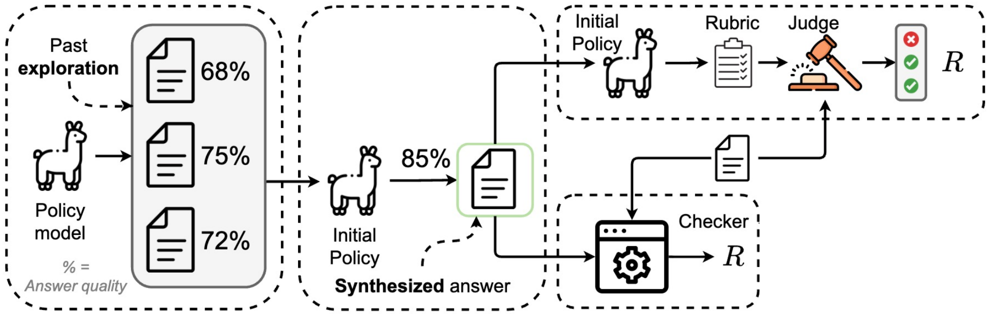


"verifiable" rewards

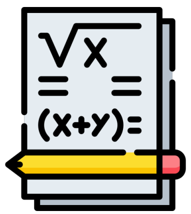
Future Work

Need to understand better: RM type $\xrightarrow{\text{affects}}$ RM properties $\xrightarrow{\text{affects}}$ Performance of LM

LM-as-a-judge



Pipelines of LMs



“verifiable” rewards

Thank You!

Work supported in part by the Zuckerman STEM Leadership Program